



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

TRINITY COLLEGE, DUBLIN

DEPARTMENT OF ELECTRONIC AND ELECTRICAL
ENGINEERING

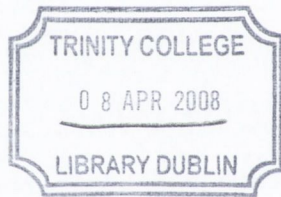
**Microphone Array Processing
Techniques for Classroom-Based
Videoconferencing**

Denis L. McCarthy

Submitted for the degree of Ph.D

January 2008

Supervisor: Frank Boland



THESIS

8344.

Acknowledgements

Firstly, I wish to thank my supervisor Professor Frank Boland both for the opportunity to pursue a PhD and for his advice, wisdom and support throughout the course of my research. Frank, I've truly enjoyed my years as a postgraduate and greatly appreciate all that you have done for me.

I would also like to thank Trinity College, Dublin and the Science Foundation of Ireland for their generous funding of this work.

Special mention should go to the staff and students in Trinity. In particular, I would like to thank Ami, Angela, Bernadette, Col, Conor, Damien, Darren, Deepti, Gavin, Robbie, Sean, and Shane for their help and friendship over the years.

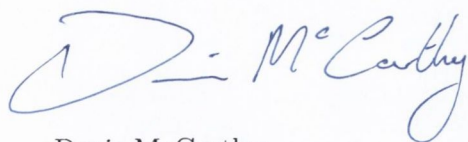
To Zoe, thank you for your love, kindness, patience and understanding. You kept me sane and happy and were the reason that, whatever else was going on, I always had something to look forward to.

I would like to thank my family for their love and support. I would also like to apologize for those multiple occasions on which all of you had to put up with my foul humour when things were not going well! To Órla, Aoibheann and Criona thank you.

Finally, to my Mum and Dad – none of this would have been possible without you. I owe you more than I can express and I love you.

Declaration

I, the undersigned, hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work. Furthermore, I agree that the Library may lend or copy this thesis upon request.

A handwritten signature in blue ink that reads "Denis McCarthy". The signature is written in a cursive style with a large initial "D" and a stylized "M".

Denis McCarthy
January 11, 2008

Abstract

This thesis is concerned with the design and development of microphone-array-processing techniques for videoconferencing applications in classroom environments. We argue that, in such environments, it is advantageous in terms of both performance and practical implementation to use widely distributed arrays. A likely consequence of the use of such arrays is that the relative locations of the microphones will be unknown, however, most previously published microphone-array-processing techniques assume and require knowledge of the microphone locations. We propose two novel algorithms which are designed for use with widely distributed arrays of unknown geometry.

The first of these is a Leaky-LMS-based method for precise beamformer steering. Beamforming is a classical array-signal-processing technique for the suppression of noise and reverberation. Traditionally, a necessary preliminary step - known as steering - has been the application of delays for the time-alignment of the target-signal components in each microphone output. When using time-sampled data, estimates of the appropriate delays are typically limited to being integer multiples of the sampling period. Consequently, the source locations to which we may precisely steer are also limited to a number of discrete locations, leading to missteering and reduced beamformer performance.

In this thesis we show that a leaky, multichannel, least-mean-squares algorithm may be used to achieve steering that is both more accurate and less computationally complex than traditional methods. We provide the results of experiments using both real and simulated data which demonstrate the efficacy of our approach in reverberant environments.

The second algorithm is a method for determining the distance between a sound source and the microphones in an array in reverberant environments. Following from the results of a series of experiments in real-room environments, we show that, in a spatially averaged sense, reverberation levels observed in a room are constant. This provides the theoretical justification for our technique, which we dub the "Range-Finder" algorithm.

We derive the Range-Finder algorithm and analyse the distribution of the source-microphone range-estimates it returns. We also provide the results of experiments in which we compare the performance of the Range-Finder algorithm to those of a naïve range estimation technique and a modified-steered-response-power technique, in both real and simulated reverberant environments.

Summary

This thesis opens with an introduction in chapter 1. In chapter 2 we discuss sound propagation in rooms, outlining relevant theory and presenting the results of experiments which investigate the variation of the Direct-to-Reverberant ratio (DRR) with distance from a sound source. These experiments were performed in real rooms. Our results show that, given an omni-directional source and receiver, the reverberation levels at any point in a room are independent of the source-microphone range and are instances of a normally-distributed random variable with a constant mean. Following from this result, we propose a new metric for describing the reverberant characteristics of a room – the “DRR-at-1m”. We define this metric and demonstrate instances in which the DRR-at-1m gives a more accurate impression of relative reverberation levels than the more commonly used measure, the “reverberation time”.

In chapter 3, we introduce the fundamentals of array-signal-processing and define the notation and much of the terminology to be used throughout the thesis. We develop the concept of spatial sampling and explore its implications with reference to spatial aliasing, “sidelobe” levels and wavenumber-vector resolution. The effects of frequency, direction-of-arrival and array geometry are explored. The issues arising from discrete-time sampling are also addressed.

In chapter 4, we review the literature concerning techniques for enhancing the perceptual characteristics of recorded speech. Starting with a review of conventional beamforming techniques, we explore the topic of “adaptive” beamformers using the “constrained-optimization” paradigm. We discuss the relationships between the popular “generalized-sidelobe-canceller” algorithm, its robust variants and the multichannel-Wiener-Filter. We also review the literature concerning dereverberation techniques.

In chapter 5, we review the literature relating to methods for determining the location of a sound source. Starting with a review of approaches to time-delay estimation, we discuss time-delay-estimate-based source-localization. Parametric and subspace-based methods are also reviewed.

In chapter 6, the inadequacies of currently-available microphone-array techniques are outlined with respect to classroom-based videoconferencing. We address the advantages to be obtained via a widely-distributed, ad-hoc and unknown microphone deployment, while also outlining the additional problems such an array-configuration

would pose. We thus define the scenario and applications for which the novel techniques presented in this thesis are designed.

In chapter 7, we present a technique for steering-vector estimation based upon a leaky-LMS filter. This method is suitable for practical implementation in situations in which the array geometry is unknown. The steering-vector thus obtained may be used to time-align the target-signal components of the outputs of multiple microphones – a necessary preliminary step for beamforming. While LMS-based techniques have been previously applied to the problem of obtaining time-delay estimates (from which a steering-vector may be determined), our approach calculates the steering-vector directly, without a requirement for intermediate time-delay estimation. Furthermore, we analyse our method under reverberant conditions. This is in contrast with previously published techniques which assume an anechoic environment. In discussion, we highlight how, in reverberant environments, the effective performance of our approach is dependent upon the selection of an appropriate step size and leakage coefficient.

We present the results of experiments using simulated and real recorded data that demonstrate that the steering-vector obtained by our method achieves accurate beamformer steering under reverberant conditions. Furthermore, we show that our approach to beamformer steering is, under certain conditions, both more accurate and less computationally complex than traditional methods.

In chapter 8, we present a method for determining the distance between a sound source and the microphones in an array - the "Range-Finder" technique. The theoretical underpinning of the Range-Finder follows from our results in chapter 2 regarding reverberation levels in rooms. As with our technique for delay-vector estimation, the Range-Finder method is suitable for application to arrays of unknown geometry. We show that the distribution of the range estimates returned by the Range-Finder is related to a Cauchy distribution which is itself a function of the relative positioning of the source and microphones. Based on our analysis, we discuss the conditions necessary for accurate range estimation and, following from this, outline possible applications for the Range-Finder technique.

Using simulated and real-room recordings, we compare the performance of our method with that of a modification of the classical Steered-Response-Power approach and a "naïve" range estimator, which assumes an anechoic environment. To the author's knowledge, the modified-Steered-Response-Power technique has not been presented elsewhere previously.

We conclude in chapter 9 with a discussion of potential future work.

Contents

1	Introduction	1
2	Room Acoustic Environments	5
2.1	Introduction	5
2.2	Sound Propagation in Rooms	5
2.2.1	Sound Energy Density	6
2.2.2	Wave Phenomena	6
2.3	Reverberation	9
2.3.1	A Linear Time-Invariant Room Model	10
2.3.2	The Room Response	10
2.3.3	Reverberation time	13
2.3.4	Direct-to-Reverberant Ratio	15
2.4	Noise	17
2.4.1	Ambient Noise	17
2.4.2	Feedback	18
2.4.3	Sensor/Quantization Noise	19
2.5	Perceptual Effects of Noise and Reverberation.	20
2.5.1	Intelligibility	20
2.5.2	Quality	22
2.6	Discussion	23
3	Array-Processing Theory	25
3.1	Introduction	25
3.2	Spatiotemporal Signals	25
3.2.1	Multidimensional Fourier Transform	26
3.3	Array Signal Processing	28
3.3.1	Spatial Sampling	28
3.3.2	Spatial Aliasing	30
3.3.3	Wavenumber Smoothing	30
3.3.4	Weighting	32
3.3.5	Sparse Arrays	33
3.4	Filter-and-Sum Array Signal Processing	33
3.4.1	Steering	36
3.4.2	Discrete-Time Sampling	36
3.4.3	Angular Resolution	37

3.4.4	Interpolation	39
3.4.5	Steering in the Frequency Domain	40
3.4.6	Non-Ideal Sensors	42
3.5	Signal Model	43
3.5.1	Anechoic Signal Model	44
3.5.2	Reverberant Signal Model	45
3.5.3	Filter-and-Sum Processing	45
3.5.4	The Spatospectral Correlation Matrix	46
3.6	Discussion	46
4	Speech-Enhancement Techniques	49
4.1	Introduction	49
4.2	Delay-and-Sum Beamforming	49
4.3	Data-Dependant Signal Enhancement	51
4.3.1	Constrained Optimization	51
4.3.2	Generalized Sidelobe Canceller	56
4.3.3	The Multichannel Wiener Filter	62
4.3.4	Blind Source Separation	63
4.4	Dereverberation	64
4.5	Discussion	65
5	Time-Delay Estimation and Source-Localization	67
5.1	Introduction	67
5.2	Time-Delay Estimation Techniques	68
5.2.1	Cross-Correlation	68
5.2.2	Generalized Cross-Correlation	68
5.2.3	Least-Mean-Squares Methods	70
5.2.4	Adaptive Eigenvalue Decomposition	71
5.3	TDE-based Source Localization	73
5.3.1	Viete's Solution	73
5.3.2	Direction-of-Arrival Estimation	75
5.3.3	Least-Square-Error Fitting	77
5.4	Parametric Methods	78
5.4.1	Maximum-Likelihood Method	79
5.4.2	Steered Response Power Techniques	80
5.5	Subspace-Based Techniques	82
5.5.1	MUSIC	82
5.6	Discussion	83
6	Classroom-Based Videoconferencing: A Problem Overview	85
6.1	Introduction	85
6.2	A Typical Videoconferencing Setup	86
6.2.1	The Far-Field Assumption	86
6.2.2	Known Array Geometry	87
6.2.3	Disadvantages	87
6.3	The Proposed Videoconferencing Setup	89

6.3.1	Steering	90
6.3.2	Source-Microphone Range Estimation	93
7	A Leaky-LMS-Based Method for Precise Beamformer Steering	95
7.1	Introduction	95
7.2	Leaky-LMS-Based Beamformer Steering	95
7.2.1	Updating the Weightvector	96
7.2.2	Decomposing the Wiener Solution	98
7.2.3	Flattening the Filter Responses	98
7.2.4	Implementation	99
7.2.5	Parameter Selection	100
7.3	Simulations and Experiments	101
7.3.1	Simulations	101
7.3.2	Real-Room Experiments	107
7.4	Computational Complexity	109
7.5	Discussion	113
8	Range Estimation	115
8.1	Introduction	115
8.2	Sound Propagation in Reverberant Environments	115
8.3	Range Estimation	117
8.3.1	A Naïve Range Estimator	117
8.3.2	The Steered-Response-Power Range Estimator	118
8.3.3	The Range-Finder Algorithm	118
8.4	Estimate Distribution and Accuracy	120
8.4.1	An Alternative Formulation of the Range-Finder	121
8.4.2	Cauchy Distribution	121
8.4.3	The Effect of Array Geometry	123
8.5	Simulations and Experiments	125
8.5.1	Simulations	125
8.5.2	Experiments	129
8.6	Discussion	130
9	Conclusion	135
9.1	Summary	135
9.2	Future work	137
9.2.1	Leaky-LMS-Based Steering	137
9.2.2	Range Estimation.	137
9.3	Conclusion	138
A	The Multichannel Wiener Filter	139
B	Constrained Optimization	154

List of Figures

2.1	Diffraction: Waves “spread” after passing through a gap or around an obstacle. (image sourced from http://www.gcscience.com/pwav37.htm)	7
2.2	Wave reflection from smooth and rough surfaces.	8
2.3	Sound absorption with respect to frequency for typical surface materials in rooms. Source: EASE 4.0 materials database, [110].	8
2.4	Interference: soundwaves generated by the monochromatic sources S1 and S2 interfere constructively (red lines) and destructively (blue lines). Note: for visual clarity, not all points at which constructive and destructive interference occur are shown.	9
2.5	An impulse response obtained in a classroom at a distance of $3.5m$ from the source.	12
2.6	Implementing the transient decay method for determining RT_{60} . The recorded acoustic signal used to generate this figure is bandlimited in the range $1 - 2kHz$.	14
2.7	A photograph of the experimental setup used to obtain $DRRs$ at varying distances from a sound source.	15
2.8	Direct-to-reverberant ratios versus $\log_2(r)$, where r is the source-microphone range. Results shown are for an office, classroom and reception hall.	16
2.9	A full-duplex communication system and the resulting “feedback loop”.	18
2.10	Acoustic echo cancellation: An adaptive filter estimates the loudspeaker-microphone response and subtracts an estimate of the acoustic echo from the microphone output.	19
3.1	A planar, non-attenuating wave with direction of propagation \vec{v} . $f(\vec{x}, t)$, the wavefield at \vec{x} , may be expressed in terms of $f(\vec{0}, t)$, where $\vec{0}$ is the origin.	27
3.2	Linear arrays; A shows a 7-element LEA; B-D are “sparse” arrays. The minimum intersensor spacing is d .	30
3.3	$ \widehat{W}(k_x) $ for a 7-element LEA. The “lobe” at $\frac{dk_x}{\pi} = 0$ is referred to as the mainlobe. The other periodically repeating lobes are known as “grating” lobes - a name with its origins in experiments where light is passed through a grating to produce an interference pattern with multiple periodically repeating bright lines.	31
3.4	$ W(k_x) $ for a 7-element LEA corresponding to a “square” window function, (a), and a Hamming window function, (b).	32

3.5	$ \widehat{W}(k_x) $ corresponding to the sparse array geometries in figure (3.2). Compared to a 4-element LEA they achieve a narrower mainlobe at the expense of larger sidelobes.	34
3.6	Filter and Sum: Sensor outputs undergo temporal filtering before being combined.	35
3.7	A planar wave propagating across a LEA. The angle of incidence, θ , is commonly referred to as the direction of arrival of the wave.	37
3.8	Steerable angles for a LEA where $d = 0.2m$, and $c = 340ms^{-1}$. Increasing the sampling rate (reducing T) increases the density of the steerable angles.	38
4.1	$ W(\omega, \theta) $ for a seven-element D&S beamformer: $W_m(\omega) = \frac{1}{\sqrt{7}} \forall m, \omega$. $d = 0.034m$. The array gain has been normalized to give unity gain in the look direction.	51
4.2	$ W(\omega, \theta) $ for a seven-element D&S beamformer: $W_m(\omega) = \frac{1}{\sqrt{7}} \forall m, \omega$. $d = 0.1m$. The array gain has been normalized to give unity gain in the look direction.	52
4.3	The Array Pattern for a MVDR beamformer.	54
4.4	$ W(\omega, \theta) $ for a seven-element superdirective beamformer: $d = 0.034m$. The array gain has been normalized to give unity gain in the look direction.	55
4.5	The Generalized Sidelobe Canceller.	57
4.6	Leaky adaptive filters (LAFs) in the blocking matrix of a GSC.	61
5.1	Time Delay Estimation using the LMS algorithm.	71
5.2	Source localization as an instance of Apollonius' problem of tangent circles.	74
5.3	The far-field assumption: For distant sound sources, the curvature of the incident wavefront is negligible.	75
5.4	The direction of arrival of a far-field sound source is a function of the intersensor time delay.	76
6.1	A videoconferencing setup as typically found in the literature.	88
6.2	A videoconferencing setup using an array of distributed microphones.	89
6.3	A videoconferencing setup using an array of distributed sub-arrays.	91
7.1	A block diagram of an implementation of the leaky-LMS-based method for beamformer steering. The elements shown correspond to the processing undergone by the output of m_1 . Identical processing is applied to the outputs of the remaining microphones.	99
7.2	The simulated room and loudspeaker-microphone setup.	102
7.3	A comparison of the white noise gain achieved for varying DOAs.	103
7.4	A comparison of the directivity index achieved for varying DOAs.	103
7.5	A comparison of the White Noise Gain achieved under reverberant conditions.	104
7.6	A comparison of the Directivity Index achieved under reverberant conditions.	105

7.7	A comparison of the White Noise Gain achieved by the Self-Steering beamformer under different DRRs.	106
7.8	A comparison of the Directivity Index achieved by the Self-Steering beamformer under different DRRs.	106
7.9	WNG over time for varying frequencies. $\alpha = 0.02$	108
7.10	WNG over time for varying frequencies. $\alpha = 0.01$	108
7.11	WNG over time for varying frequencies. $\alpha = 0.002$	109
7.12	The array patterns corresponding to the self-steering beamformer, obtained from real-room recordings.	110
8.1	Portions of the PDFs of $\frac{N(\alpha,1)}{N(\beta,1)}$. Also shown is $\frac{\alpha}{\beta}$ (dashed line).	122
8.2	$\frac{G_{0,1}}{G_{0,2}}$ versus r_0 for $[c\tau_1, c\tau_2] = [1m, 5m]$. Range estimate error increases with r_0	124
8.3	$\left \frac{d}{dr_0} \left(\frac{G_{0,1}}{G_{0,2}} \right) \right $ with respect to $\frac{c\tau_1}{r_0}$ and $\frac{c\tau_2}{r_0}$	124
8.4	A diagram of the simulated room and setup. For precise coordinates of the microphones and loudspeakers, see Table 1.	126
8.5	Range estimates \pm one standard deviation, obtained using the Range-Finder, Naïve and SRP-based methods, for source 2.	128
8.6	Range estimates \pm one standard deviation, obtained using the Range-Finder and SRP-based methods, for source 1.	129
8.7	Range estimates \pm one standard deviation, obtained using the Range-Finder and SRP-based methods, for source 3.	130
8.8	Range estimates \pm one standard deviation, obtained using the Range-Finder, Naïve and SRP-based methods, using real recordings of a maximum-length sequence.	131
8.9	Range estimates \pm one standard deviation, obtained using the Range-Finder, Naïve and SRP-based methods, using real recordings of concatenated speech samples.	132

List of Tables

7.1	Integer multiples of the Nyquist sampling rate for which the proposed leaky-LMS-based steering approach is less computationally complex than steering with time-delays determined using the PHAT-GCC method	113
8.1	The coordinates of the microphone and source locations for the simulated room. Coordinates are in meters	126

Chapter 1

Introduction

Recent years have seen the rapid development of telecommunications technology both in terms of its advancing capabilities and in terms of its evolution towards near-ubiquity in our day-to-day lives.

When communication technology has been its most successful, it has been when it has allowed us to interact with others across boundaries of distance, location and time. We are, perhaps, all familiar with the advantages and impact of mobile (cellular) phones, text messages and E-mail. Emerging technologies promise an even more profound effect. Text-to-speech, speech-to-text and automatic translation technologies have begun and will continue to break down the barriers of disability and language.

Some of the most exciting potential applications for telecommunications technology exist in the area of education. In this thesis, we shall be seeking to design and develop techniques that support and facilitate classroom-to-classroom videoconferencing.

The potential benefits of such technology are difficult to overstate. Whereas today students and educators are largely constrained by geography, in the future they could be able to interact with other students, lecturers, teachers, politicians and public personalities in ways that would enhance the quality and depth of their education. In particular, by interacting with their peers in other (perhaps distant) countries, students would broaden their experience and understanding of differing cultures and outlooks – something that is increasingly important in a shrinking world.

However, to be widely adopted, classroom-based videoconferencing must facilitate interactions that are natural and spontaneous, whilst at a practical level being cost-efficient, reliable and effective. In addition, we must be mindful of the fact that such videoconferencing systems will, in all likelihood, be set up by people (i.e. teachers) with minimal experience in audio and video technology. Therefore, any systems or equipment used should require only a minimum amount of technical know-how for their use and maintenance. Consider the ideal. With no special effort on the part of the

participants, a talker's words are transmitted to the far-end classroom. Furthermore, the location of the active contributors is automatically determined such that a camera may be steered towards them. At the far-end they are heard clearly and intelligibly. In addition, the location information is used at the far-end so that a talker's speech is perceived to originate at his/her on-screen location, thus maintaining the sense of presence that is vital for the naturalness of such interactions.

Natural and spontaneous interaction between participants will feature a certain amount of speech overlap (i.e. multiple participants talking at the same time). In order to facilitate this, our videoconferencing system must support "full-duplex" communication - i.e. the channels through which sound is transmitted from one classroom to the other must be open in each classroom at the same time. Unfortunately, when using full-duplex communication systems there exists the potential for a "feedback-loop" to become established. Sound is detected in one room and transmitted to a second where it is produced by loudspeakers. In the second room this sound is detected by microphones and transmitted back to the first room, causing the talker to hear a delayed version of his/her own voice. This is known to be very off-putting for participants. Also, to provide the spatial audio cues required to achieve a sense of "telepresence", each classroom must contain multiple loudspeakers. This creates multiple feedback-loops, exacerbating the problem.

Achieving ideal videoconferencing presents a number of other technical challenges. Among these, how to locate an active talker and how to faithfully capture his or her speech? This latter obstacle is most commonly overcome using head or lapel-mounted microphones. In many situations, however, purchasing and maintaining such equipment, in quantities sufficient for a potentially large number of participants, is not cost-efficient. At the same time, sharing fewer microphones by, say, passing them around the room is not conducive to spontaneity or natural interaction.

We therefore require a system in which a moderate number of microphones would be sufficient to effectively capture speech from every potential contributor. However, in a classroom, audience members will be widely distributed and so we can expect a participant to be some distance (perhaps several meters) from a microphone. As we shall see, classrooms may be expected to be both noisy and reverberant (echoic) environments. By increasing the separation between talkers and microphones, we increase the attenuation of a target speech signal (due to propagation losses) and thereby increase the relative power of noise and reverberation. This leads to a reduction in speech intelligibility and perceived quality.

We therefore require methods by which we may suppress noise, reverberation and the off-putting effects of feedback-loops. Microphone-array-processing techniques, using networks of multiple remote microphones, provide a potential solution.

Array-processing techniques have their origins in the Second World War, where radar and sonar processed the outputs of multiple sensors to exploit the spatial as well as temporal and spectral information contained in the observed waves (electromagnetic and acoustic respectively). Using these techniques, the location and heading of enemy planes and submarines etc. could be determined.

Using similar methods, the location of an active talker may be found using arrays of microphones. As we shall see, microphone arrays may also be used to enhance the perceived quality and intelligibility of recorded sound, by means of processing techniques that attenuate interference such as noise and reverberation while maintaining the target speech signal.

As shall become apparent from our review of the literature, previously published techniques are typically based upon the assumptions/requirements that the microphone array geometry (that is, the relative positions of each microphone) is fixed, known *a priori* and that the distance to the target talker is large with respect to the width or extent of the array. In fact, these assumptions put us at significant practical disadvantage in situations, such as in classrooms, where it would be most advantageous to allow microphones to be moved according to the specific and changing requirements of the room or audience.

In this thesis we present two novel microphone-array-processing techniques which do not require knowledge of the array geometry and which may be applied when the target talker is close with respect to the array dimensions. The first of these is a method for estimating the distance between an active talker and multiple microphones at unknown locations. The second is a method for “steering” – a necessary preliminary step for many speech enhancement algorithms.

Classroom-based videoconferencing is by no means the only application for such technologies. Rather, it represents one of the most challenging scenarios that system designers are likely to encounter. Consequently, microphone-array-processing techniques that are effective in a classroom are highly likely to be directly applicable to, for example, office-to-office videoconferencing, “hands-free” telecommunications or talker-identification for automated speech transcription. Therefore, while we shall continue to refer to the specific case of the classroom, the microphone-array-processing techniques proposed in this thesis will have applications in rooms and enclosed spaces in general.

This thesis will continue as follows. Chapter 2 shall investigate rooms as acoustic environments. Sound propagation in rooms shall be discussed. Following from this, reverberant sound propagation shall be explored and the concept of the “room impulse response” introduced. Furthermore, the metrics to be used in this thesis for quantifying the acoustic properties of rooms shall be defined. Sources of acoustic noise shall

be discussed and the characteristics of these with particular implications for array processing shall be highlighted.

Chapter 2 shall conclude with a review of the literature concerning the effects of noise and reverberation upon the perceptual characteristics of speech. This shall confirm their adverse effects on speech intelligibility and quality, thus vindicating our requirement for noise/reverberation-suppression techniques.

In chapter 3, we will introduce the fundamentals of array-signal-processing and define the notation and much of the terminology to be used throughout the thesis. We shall develop the concept of spatial sampling and explore its implications with reference to spatial aliasing, “sidelobe” levels and wavenumber-vector resolution. The effects of frequency, direction-of-arrival and array geometry shall also be explored. Furthermore, the issues arising from discrete-time sampling shall be addressed.

In chapter 4, we review the literature concerning techniques for enhancing the perceptual characteristics of recorded speech. Starting with a review of conventional beamforming techniques, we explore the topic of “adaptive” beamformers using the “constrained-optimization” paradigm. We discuss the relationships between the popular “generalized-sidelobe-canceller” algorithm, its robust variants and the multichannel-Wiener-Filter. We also review the literature concerning dereverberation techniques.

In chapter 5, we review the literature relating to methods for determining the location of a sound source. Starting with a review of approaches to time-delay estimation, we discuss time-delay-estimate-based source-localization. Parametric and subspace-based methods are also reviewed.

In chapter 6, the inadequacies of currently-available microphone-array techniques are outlined with respect to classroom-based videoconferencing. We shall address the advantages to be obtained via a widely-distributed, ad-hoc and unknown microphone deployment, while also outlining the additional problems such an array-configuration would pose. We thus define the scenario and applications for which the novel techniques presented in this thesis are designed.

In chapter 7, we present a technique for delay-vector estimation based upon a leaky-LMS filter. This method is suitable for practical implementation in situations in which the array geometry is unknown. We also present the results of simulations and experiments that demonstrate the efficacy of this approach.

In chapter 8, we present a method for determining the distance between a sound source and the microphones in an array – the “Range-Finder” technique. Once again, this method is suitable for application to arrays of unknown geometry and its effectiveness is verified by the results of experiments using real and simulated data. We conclude in chapter 9 with a discussion of potential future work.

Chapter 2

Room Acoustic Environments

2.1 Introduction

In the following chapter we discuss room acoustic environments and their influence on the success or otherwise of classroom-based videoconferencing. We start by investigating sound propagation in rooms and briefly discuss the effects of the so-called “wave phenomena”. The combined effects of these phenomena are jointly referred to as reverberation and we present the metrics used in this thesis to quantify the degree to which reverberation is present in a room or on a recording. Noise in rooms is also discussed and this is followed by a review of the literature concerned with the perceptual impact of noise and reverberation on listeners.

2.2 Sound Propagation in Rooms

Sound waves are introduced to the medium (air) by vibrating objects, which cause the space between adjacent molecules to be compressed and expanded. The resulting disturbance travels from molecule to molecule transporting energy as it moves. Sound is a longitudinal wave – that is, its oscillations are parallel to the direction of travel.

At sea level in dry air, the speed of sound, c , may be approximately calculated as shown below, where T_C is the temperature in degrees Celsius.

$$c = (331.4 + 0.6T_C) \text{ ms}^{-1} \quad (2.1)$$

For indoor acoustic environments, we make the following assumptions. Firstly, we assume the medium to be homogenous with no significant density, temperature or pressure gradients. In addition we assume there to be no significant wind or air flow. Following from these, we therefore assume c to be known (corresponding to room temperature) and constant within the room.

A full and rigorous analysis of soundwave propagation would be excessive in the context of this discussion. Nonetheless, we find it useful to briefly discuss the behavior of soundwaves with respect to their propagation in rooms.

2.2.1 Sound Energy Density

As sound propagates through a medium the sound energy density (where sound energy density = $\frac{\text{sound energy}}{\text{area}}$) reduces. This is primarily due to “geometric spreading” of sound energy as the distance between the source and the location at which the sound is detected increases. Assuming an omnidirectional point sound source, “direct-path” (i.e. non-reflected) soundwaves will propagate outwards forming an expanding sphere. Since energy is conserved, the sound energy density will be inversely proportional to the surface area of this sphere. This surface area may be given by $4\pi r^2$, where r is the radius of the sphere. Therefore we may write

$$\text{sound energy density} \propto \frac{1}{r^2} \quad (2.2)$$

Molecular absorption (due to, for example, air viscosity) also contributes to the reduction in intensity of propagating soundwaves. However, molecular absorption reduces sound energy density by a factor that is inversely proportional to the distance travelled, as opposed to the distance squared. In air, sound energy density losses due to molecular absorption are small and, over the short ranges typical of indoor environments, may be considered negligible in comparison with those of geometric spreading.

Note that, for simplicity, the term “sound intensity” shall be taken to mean the sound energy density in the sequel.

2.2.2 Wave Phenomena

As a wave, sound will exhibit the classical wave behaviors of refraction, diffraction, interference and reflection. We briefly discuss each of these and their relevance to the indoor scenarios under investigation.

Refraction

Refraction is the phenomenon whereby the direction of a wave changes due to a change in its speed. This occurs as a wave travels from one medium to another. Refraction is more commonly demonstrated with lightwaves, such as in the classic example of the “bent” pen in a glass of water. However, as per our initial assumptions, the air-density gradients typical in rooms are such that refraction of soundwaves is insignificant.

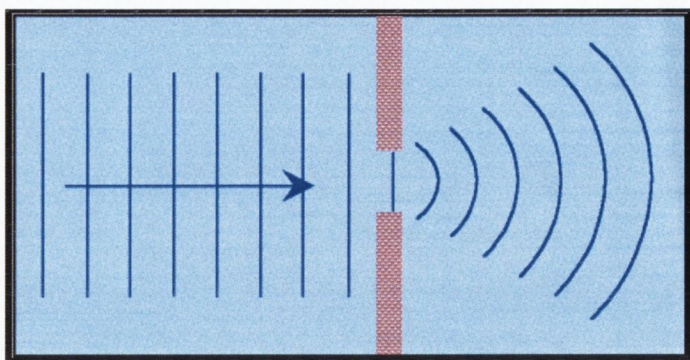


Figure 2.1: Diffraction: Waves “spread” after passing through a gap or around an obstacle. (image sourced from <http://www.gcscience.com/pwav37.htm>)

Diffraction

Diffraction refers to the “spreading out” of soundwaves after passing through a gap or around some obstacle. In general, the effects of diffraction are only noticeable where the wavelength is of the same order of magnitude as the diffracting obstacle or gap. However, the wavelengths of audible soundwaves are of a similar magnitude to the windows, doorways and furnishings common in typical rooms. We may therefore, expect diffraction to occur in the scenarios under investigation.

Reflection

When incident with a surface, soundwaves are reflected. Specular or “mirror-like” reflection occurs when the reflecting surface is smooth, figure (2.2a). Alternatively, rough surfaces will lead to diffuse reflections and a “scattering” of the reflected sound, figure (2.2b). The degree of scattering that occurs is frequency-dependent, with higher frequencies tending to be scattered more.

While most of the sound energy is reflected, some is absorbed. As with scattering, the degree of absorption that occurs is frequency-dependent. Figure (2.3) shows surface absorption with respect to frequency for common indoor surface materials.

Interference

At points in space where multiple waves meet, the resulting waveform is the algebraic sum of the component waves. This is known as interference. For two monochromatic sources generating soundwaves of equal wavelength, the “interference pattern” shown in figure (2.4) would result. The blue lines in figure (2.4) correspond to points where differences in the distance of propagation cause the instantaneous amplitude due to

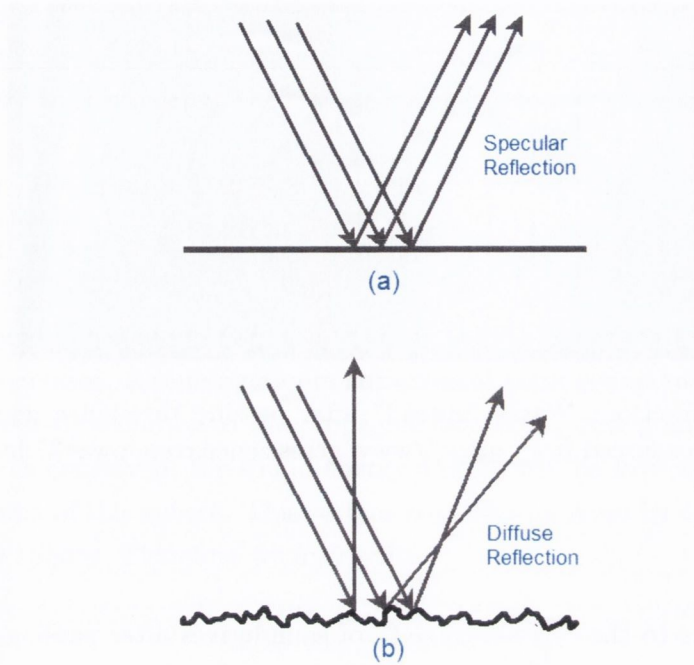


Figure 2.2: Wave reflection from smooth and rough surfaces.

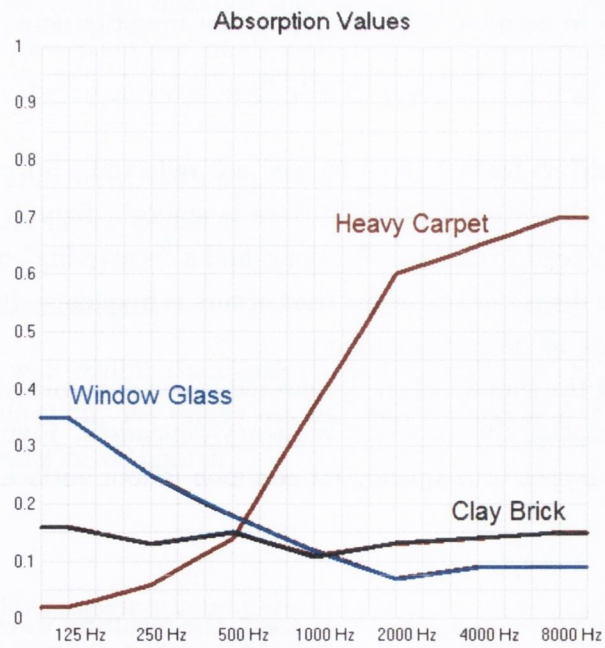


Figure 2.3: Sound absorption with respect to frequency for typical surface materials in rooms. Source: EASE 4.0 materials database, [110].

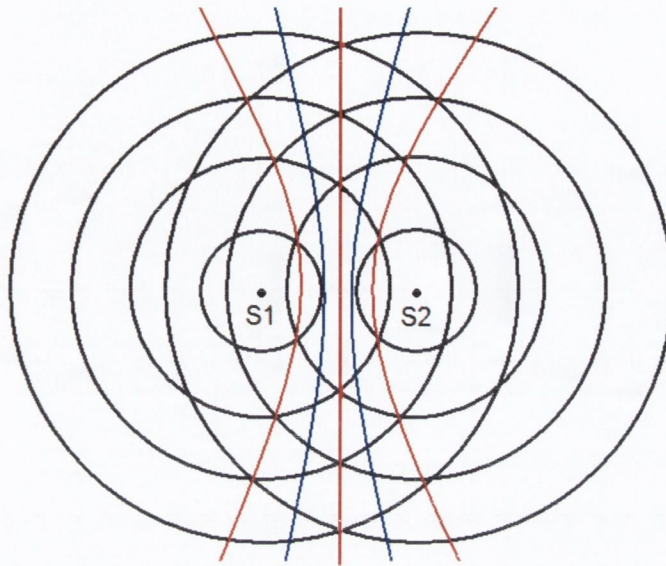


Figure 2.4: Interference: soundwaves generated by the monochromatic sources S1 and S2 interfere constructively (red lines) and destructively (blue lines). Note: for visual clarity, not all points at which constructive and destructive interference occur are shown.

one source to be π radians out of phase with the amplitude due to the other. Thus their sum is zero and we observe complete “destructive” interference. The red lines correspond to points where the instantaneous amplitude due to each source is in-phase, thus maximizing the amplitude of the resultant wave at those points. This is known as complete “constructive” interference. The remaining points in figure (2.4) would experience interference somewhere in between these two extremes.

In a reflective or “echoic” acoustic environment, multiple reflected copies of the source signal will meet at points throughout the room. However, for broadband signals, intensity variations due to interference are not perceptible as destructive interference at one frequency is compensated for by constructive interference in another.

2.3 Reverberation

The propagation losses and wave phenomena discussed in the previous section will cause an acoustic signal emitted by some sound source to be subject to “room effects” - jointly referred to as reverberation. As a result, the acoustic signal observed at any location in space will differ from the emitted sound. Furthermore, this observed sound will differ from that observed at any other location in space. In the following section we discuss reverberation and introduce the metrics used to quantify it in this thesis.

2.3.1 A Linear Time-Invariant Room Model

In the scenarios under investigation, a sound will be subject to source-directionality, whereby the intensity of the direct-path soundwave is dependent upon the orientation of the source relative to the location at which the sound is detected. Furthermore, this directionality is frequency-dependent. In addition, the emitted soundwaves are subject to room effects such as reflections and diffraction etc. Finally, the microphone itself will apply a frequency-and-orientation-dependent gain to the acoustic signal. The source, room and microphone combined represent a “system” that filters a “clean” acoustic signal to produce a microphone output. We shall model this system as being linear and time-invariant.

We must, however, qualify our assumptions of linearity and time-invariance. Regarding linearity, we make the assumption that the soundwaves of interest have amplitudes above the threshold of detection and below the saturation levels corresponding to the microphones being used to detect them. We also assume these microphones to have a linear response in this region. Regarding our assumption of time-invariance, a “room response” will, in practice, vary in time - often significantly so. Typical causes relate to changes in the room environment. Opening doors or windows and drawing blinds or curtains will alter the room response. So too will moving furniture or, indeed, the microphones themselves. Perhaps most significantly, talkers engaged in natural conversation will alter the room response as they move their heads, gesture and/or walk around.

Nonetheless, as is done throughout the literature, we shall continue to make the simplifying assumption of time-invariance whilst highlighting those instances where doing so would cause us to draw false conclusions or obtain flawed results.

2.3.2 The Room Response

Any linear, time-invariant (LTI) system may be wholly characterized by its impulse response (in the time domain) or transfer function (in the frequency domain). In a noiseless room environment, the acoustic signal observed at some point, $y(t)$, may be expressed as the convolution of the source-microphone impulse response, $h(t)$, and the “clean” acoustic signal originating at the source, $s(t)$.

$$x(t) = h(t) * s(t) \tag{2.3}$$

In the frequency domain

$$\begin{aligned} X(\omega) &= H(\omega)S(\omega) \\ &= (H_{dp}(\omega) + H_{mp}(\omega))S(\omega) \end{aligned} \quad (2.4)$$

where $H_{dp}(\omega)$ and $H_{mp}(\omega)$ denote the components of the frequency response corresponding to direct-path propagation and room effects (i.e. “multipath” propagation) respectively. Treating the effects of molecular absorption as negligible compared to those of geometric spreading, we may consider the intensity of the direct-path component of the received sound to inversely proportional to the source-microphone range squared. Therefore we may write

$$|H_{dp}(\omega)|^2 \propto \frac{1}{r^2} \quad (2.5)$$

from which we may determine that the direct-path component of received sound intensity decreases by $6dB$ per doubling of r .

Figure (2.5) depicts a source-microphone impulse response, which was obtained in a classroom as follows. A microphone was placed directly in front of a loudspeaker at a distance of $3.5m$. The loudspeaker produced a pseudorandom Maximum-Length-Sequence (MLS) of approximate duration $5.5s$, at a sampling rate of $48kHz$. The output of the microphone was recorded, also at a sampling rate of $48kHz$. This recording was then cross-correlated with the “clean” MLS to obtain an impulse response estimate.

The initial pulse in the impulse response is due to direct-path propagation from the source to the microphone and, as we would expect, corresponds to a scaled delay. The remainder of the impulse response is primarily due to reflections but also characterizes the effect of the other wave phenomena. This portion of the impulse response is composed of a series of delayed impulses that are attenuated relative to the direct-path component. This is consistent with what we might expect in a room environment where, in addition to the direct-path component, multiple soundwaves reach the microphone after being reflected.

The multipath components of the arriving sound are jointly referred to as “reverberation”. As we shall see in subsequent sections, reverberation has the potential to effect the intelligibility and perceived quality of speech. Furthermore, reverberation can adversely effect the accuracy and performance of microphone-array-processing algorithms. It is, therefore, important to be able to characterize the degree of reverberation present in a room or recording. We shall now introduce the metrics that we shall use in this thesis to do so.

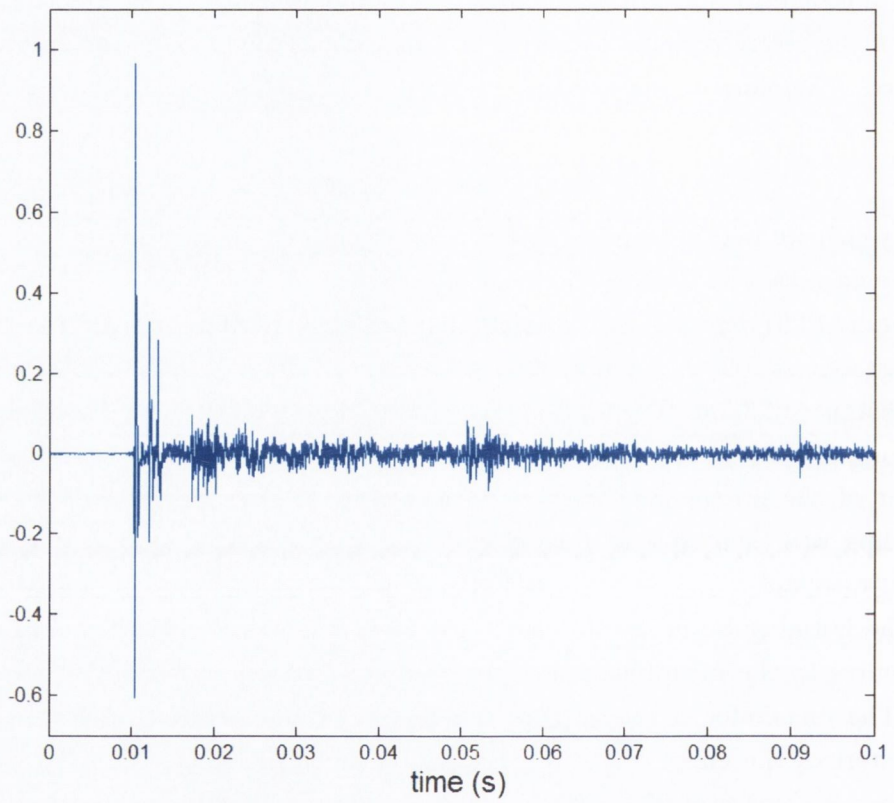


Figure 2.5: An impulse response obtained in a classroom at a distance of $3.5m$ from the source.

2.3.3 Reverberation time

In an enclosed space the reverberant component of an observed sound will decay exponentially with time. The “Reverberation Time”, (RT_{60}), is the time taken for the energy of the reverberant component to decay by $60dB$. Originally proposed by Sabine in the late 19th century, the RT_{60} remains in wide use in the field of architectural acoustics. Sabine’s reverberation equation, (2.6), describe the relationship between the RT_{60} and the physical characteristics of a room.

$$RT_{60} = 0.161 \frac{V}{SA_s} \quad (2.6)$$

where V is the volume, S the surface area and A_s the average surface absorption coefficient of a room. Sabine’s equation (which assumes a fully diffuse soundfield) describes a phenomena that will be intuitive to many readers, whereby large rooms with reflective surfaces will be more reverberant. However, because it characterizes the acoustic environment in terms of only its volume, surface area and average surface absorption, Sabine’s equation does not account for the effects of more subtle variations in room geometry or room contents. Several variations/improvements on (2.6) exist - most notably that of Eyring, [1] - but these also suffer from similar deficiencies. A more satisfactory approach is to measure the RT_{60} directly. Using the transient decay method, [2], a spectrally white acoustic signal is interrupted and the RT_{60} determined from the rate of decay of the reverberant sound intensity. To investigate the variation of the RT_{60} with frequency, the recorded signal may be bandlimited as required. A simple illustration of this technique is shown in figure (2.6).

As previously shown in figure (2.3), surface absorption and hence RT_{60} , varies with frequency. Therefore, as per what has become the convention, we shall, when specifying the RT_{60} , be referring to the RT_{60} in the region of $1kHz$ – this being generally accepted as indicative of the relevant acoustic characteristics of a room when the sound source is human speech.

Although useful for conveying a general idea of how reverberant a room may be, specifying the RT_{60} gives no idea of how reverberant a recorded sound will be. Consider, for example, a recording made in a room at a distance of $1m$ from a sound source. This recording will be perceived as being less reverberant than one made in the same room at $5m$ from the source. This is because the direct path component decays as we get farther from the source, despite the RT_{60} being the same in each instance. Nonetheless, reverberation time still finds widespread use throughout the literature, despite its relative crudeness.

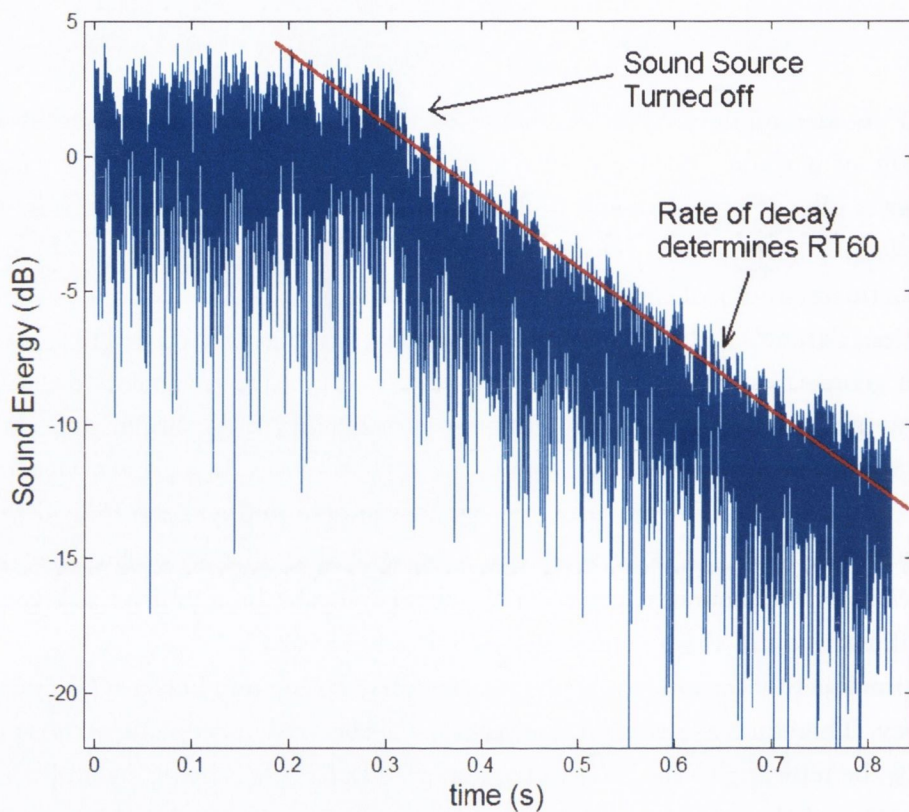


Figure 2.6: Implementing the transient decay method for determining RT_{60} . The recorded acoustic signal used to generate this figure is bandlimited in the range 1 – 2kHz.



Figure 2.7: A photograph of the experimental setup used to obtain $DRRs$ at varying distances from a sound source.

2.3.4 Direct-to-Reverberant Ratio

A more effective way of describing the degree of reverberation that obtains on a recording is to specify the direct-to-reverberant ratio (DRR) - that is, the ratio of the received sound energy due to the direct-path component and multipath reverberation. For a given bandwidth, the DRR for the output of a microphone, m_0 , may be defined as follows

$$DRR_0 = \frac{\int |H_{dp_0}(\omega)|^2 d\omega}{\int |H_{mp_0}(\omega)|^2 d\omega} \quad (2.7)$$

An investigation of $DRRs$ in real rooms proves informative. Figure (2.8) shows a plot of $DRRs$, found at a variety of locations in an office, classroom and reception hall. The $DRRs$ are plotted with respect to $\log_2(r)$. The reverberation times were determined experimentally using the transient decay method and were found to be 0.6s, 0.5s and 1.1s respectively. Source-microphone impulse responses estimates were obtained as previously described in section 2.3.2 and, from these, the $DRRs$ were estimated. Recordings were made at varying locations in each room and at varying distances relative to a single source - once again a loudspeaker. In each instance, the microphone was placed directly in-front of the loudspeaker so as to avoid complications due to the directivity of the source.

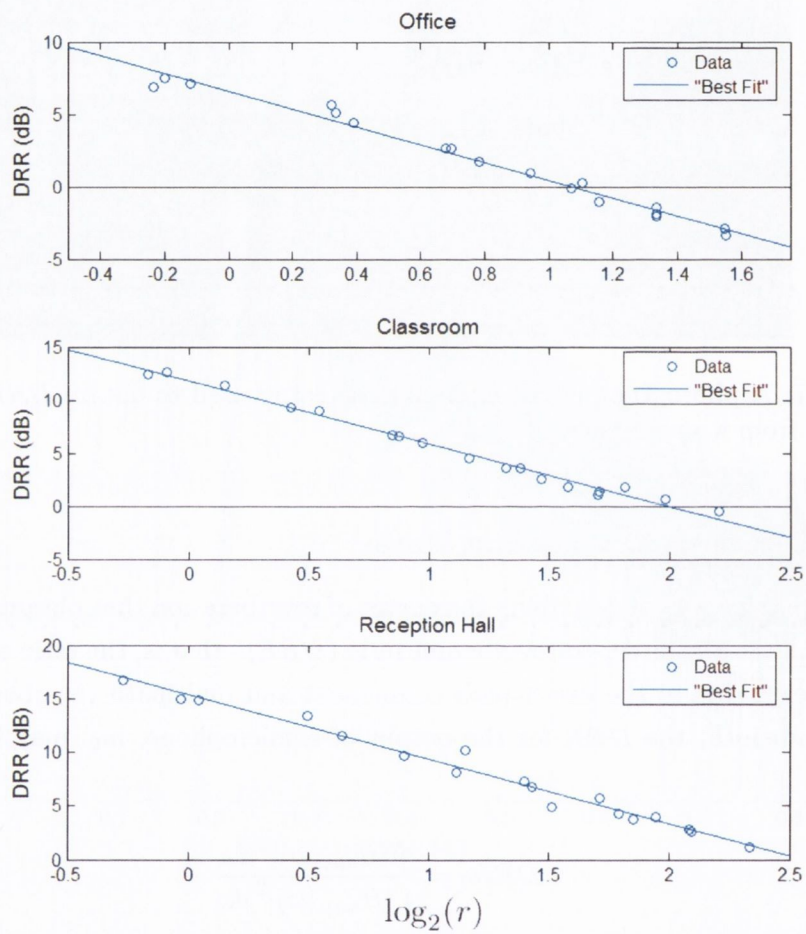


Figure 2.8: Direct-to-reverberant ratios versus $\log_2(r)$, where r is the source-microphone range. Results shown are for an office, classroom and reception hall.

Figure (2.8) also shows “best-fit” linear approximations of the data. The slopes of these fits are -6.12 , -5.99 and -5.915 decibels per doubling of range for the office, classroom and hall respectively. Given that we can expect $|H_{dp_0}|^2$ to decay at a rate of $6dB$ per doubling of the source-microphone range, these results suggest that, in a given room, $E \left\{ \int |H_{mp_0}|^2 d\omega \right\}$ is a constant that is independent of the source-microphone range.

We note that a brief inspection of the results in figure (2.8) reveals that, although it had the greatest RT_{60} , the reception hall was not the most reverberant of the rooms in which we took measurements. This further illustrates the inadequacy inherent in characterizing the degree of reverberation in a room by specifying its RT_{60} alone. Our results do, however, suggest an alternative metric. The intercept of best-fit line with the y-axis defines the spatially-averaged “ DRR -at- $1m$ ” and we shall use this metric to describe acoustic conditions in the sequel.

2.4 Noise

The term “noise” is generally taken to refer to any sound that is, in some sense, undesirable, off-putting, containing no useful information etc. Noise that is likely to be encountered in the scenarios under investigation may be separated into three broad categories - ambient noise, quantization/sensor noise and noise due to feedback.

2.4.1 Ambient Noise

In addition to (probably reverberant) speech, classrooms may be expected to contain certain levels of ambient acoustic noise. This could be as the result of exterior noise sources like passing traffic. Much noise will also originate from local sources – i.e. acoustic sources in the same room as the target talker. Examples include “fan-noise” produced by computers and air conditioning units.

In [3], the authors surveyed acoustic conditions in 32 unoccupied public school classrooms across the American state of Ohio – 8 in rural, 12 in urban and 12 in suburban areas. It is worth noting that, of these, only 4 had noise levels below those recommended for effective learning, [4].

The problem is compounded in occupied classrooms. Nominally silent audience members may also be expected to generate certain levels of noise. Coughs, sneezes, page turning, moving chairs etc. are all causes of noise and the effects of these may be amplified if, for example, the noise source is closer to a microphone than the target speech source.

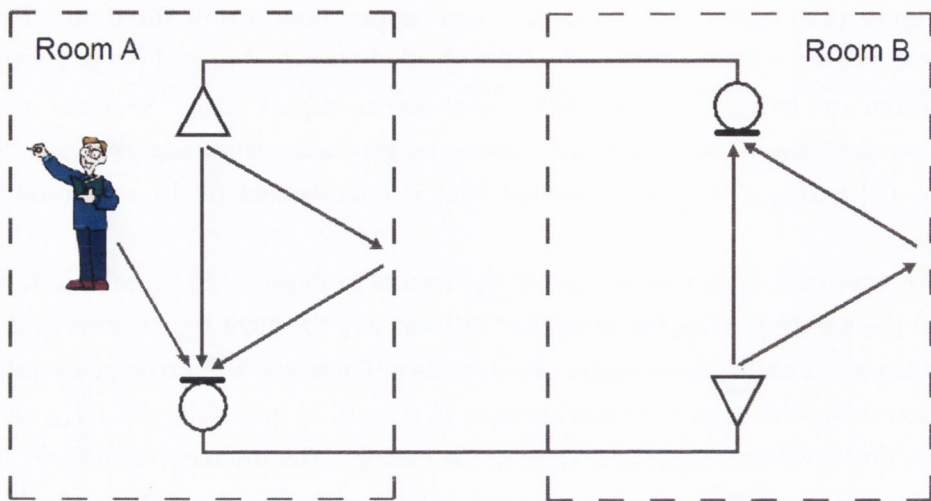


Figure 2.9: A full-duplex communication system and the resulting “feedback loop”.

2.4.2 Feedback

In the scenario under investigation, natural and spontaneous interaction between participants will feature a certain amount of speech overlap (i.e. multiple participants talking at the same time). In order to facilitate this, our videoconferencing system must support full-duplex communication. Unfortunately, when using full-duplex communication systems there exists the potential for a “feedback-loop” to become established, whereby sound is detected in one room and transmitted to a second where it is produced by loudspeakers. There, the sound is detected by microphones and transmitted back to the first room, figure (2.9). This gives rise to two forms of noise – “howl” and acoustic echoes.

Howl (or acoustic feedback) is the “whistle” that occurs when the response of the feedback-loop has a magnitude greater than 1 and a phase shift that is some integer multiple of 2π radians, at some frequency. In such scenarios signal components at that frequency are amplified with each pass through the loop, until a point of saturation is reached. Other sounds become unintelligible and the result is a deeply unpleasant listening experience.

Various methods exist for suppressing howl. Primary among these is the judicious placement of loudspeakers and microphones. However, this requirement may not be apparent to the non-professionals who are likely to be setting up a videoconferencing system in a classroom. A single-channel approach to howl-suppression is frequency shifting. At some point in the feedback loop, a frequency shifter imperceptibly in-

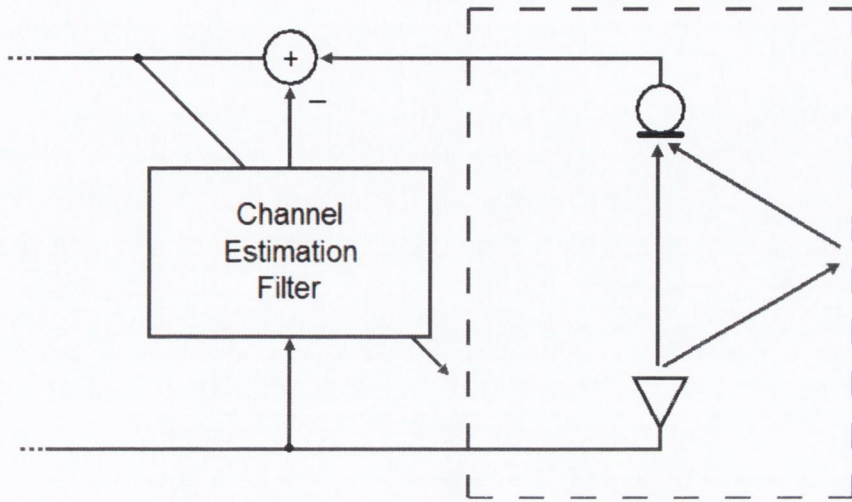


Figure 2.10: Acoustic echo cancellation: An adaptive filter estimates the loudspeaker-microphone response and subtracts an estimate of the acoustic echo from the microphone output.

creases the frequency of the signal. In this way, continuous positive feedback cannot become established at any frequency and howl is suppressed.

The term “acoustic echo” describes the phenomenon whereby the feedback loop causes a talker to hear a delayed version of his/her own voice. Traditional acoustic echo cancellation (AEC) techniques involve the use of filters that estimate the loudspeaker-microphone channel, figure (2.10). An estimate of the acoustic echo is thus subtracted from the microphone output. Typically, the filter is updated by means of a Least-Mean-Squares-type algorithm. However, a long-recognised problem in the field of AEC is the “non-uniqueness” problem that arises when multiple loudspeakers emit correlated signals, [5]. In such scenarios (which include that which we are investigating) currently-available filter update algorithms have difficulty in accurately approximating the loudspeaker-microphone channel and AEC fails.

Therefore, we must assume the presence of acoustic echo as existing methods for its suppression are inadequate for the scenario being considered.

2.4.3 Sensor/Quantization Noise

In any practical digital system, an acoustic signal is sampled and recorded with a finite degree of precision. The representation of sound as a series of discrete values gives rise “quantization noise”. In addition, a certain degree of sensor noise may be

expected to be present in the outputs of non-ideal sensors. In general, both sensor and quantization noise are modelled as additive, Gaussian random signals that are uncorrelated across the microphone outputs.

Sensor/quantization noise may be easily controlled by our choice of quantization levels and the use of microphones of sufficiently high quality. Hence, on its own, sensor/quantization noise rarely leads to audible degradation of recorded speech. However, as we shall see in chapter 4, the presence of such noise does limit the degree to which we may suppress other types of noise and so bears mention here.

2.5 Perceptual Effects of Noise and Reverberation.

In the following section, we review the literature concerning the perceptual impact of noise and reverberation on speech. In particular, we focus on the intelligibility and “quality” of speech. It is, of course, difficult to assess the perceptual and hence completely subjective characteristics of an audio signal. This difficulty is compounded by the sheer variety of acoustic environments, types of noise etc. which it is possible to encounter. As we shall see, however, there is certainly sufficient evidence in the literature to verify a supposition based upon intuition and normal everyday experience – that noise and reverberation degrade the intelligibility and quality of speech.

2.5.1 Intelligibility

Intelligibility is generally defined as being the degree to which words are correctly heard/identified/understood by listeners. Noise and reverberation both serve to reduce the intelligibility of speech. Noise degrades our comprehension of speech by corrupting its perceptually important physical characteristics. Our perception of the “loudness” of noise is determined both by the intensity and frequency of the sound – that is to say that sounds of equal intensity but different frequencies may be perceived as having different loudness. The “A-curve” or “equal loudness curve” plots intensity with respect to frequency for equal perceived loudness, [6]. This gives rise to the A-weighted SNR (ASNR) by which speech and noise signal components are assigned a weight which is inversely proportional to the A-curve, thereby placing greater emphasis on those frequencies that are perceived as being louder.

Reverberation reduces intelligibility by temporally “smearing” the speech signal. Direct-path components of recent utterances arrive at the same time as reflected components of previously uttered sounds. As a result, strong vowel sounds may, for example, mask consonants, making “Bad” indistinguishable from “Bat”/“Bap”/“Back” etc.

In [7], Bradley performs a series of intelligibility tests across ten classrooms chosen as representing the full spectrum of classroom acoustic environments. Each classroom contained an average of 24.3 12 – 13 year-olds who were asked to identify a series of words produced by a loudspeaker at the front of the room. Comparing intelligibility-test scores (*ITS*) to the acoustic conditions under which they were obtained, a regression equation was obtained which predicted test scores with a standard error of 9.6%. This equation illustrates the detrimental impact of noise and reverberation.

$$ITS = (2.26ASN R - 0.0888ASN R^2 - 13.9RT_{60} + 95) \% \quad (2.8)$$

However, not all reverberation is detrimental. It has been known since the work of Haas [8], that early arriving reflections are not perceptible as reverberation but rather are perceived in such a way as they are combined with the direct-path sound, increasing its loudness. In [8] early arriving was taken to mean <40ms after the direct-path. More recent work chose cutoffs ranging between 35 – 50ms.

Haas' work has led several authors to investigate whether, in the presence of noise, speech intelligibility is maximized by non-zero levels of reverberation – the idea being that early reflections are perceived as being direct sound and thus increase some “perceived SNR”

In [10], Bistafa and Bradley perform a theoretical investigation assuming ideal “diffuse” reverberation (i.e. propagating with equal intensity from all directions). Noise and reverberation levels were varied and analysis performed on the resulting changes in metrics previously shown to be effective predictors of mean intelligibility scores (these metrics being functions of the physical characteristics of the noisy and reverberant sound).

The results obtained revealed that in noisy environments optimal reverberation levels are non-zero (with the precise optimal level being determined by noise levels, source-microphone distances etc.). However, as pointed out in [11], the investigations in [10] contained a flaw in that the noise levels were assumed to be independent of the reverberation levels. In fact, noise will not be absorbed by room surfaces but will be reflected. Therefore, increasing reverberation levels lead to a corresponding increase in the (reflected) noise intensity. Also, just as early reflections increase the perceived loudness of speech, so too will they increase the perceived loudness of noise. Correcting for this mistake it was found in [11] that zero reverberation is optimal when the speech source is closer to the microphone/listener than the noise source. Otherwise some non-zero reverberation level is optimal.

A failure to realize that noise levels are in some ways a function of and not independent of reverberation levels may also explain contradictions elsewhere in the

literature. In [12] intelligibility scores obtained in reverberant but noiseless environments are shown to decrease from $\sim 100\%$ to $\sim 80\%$ as RT_{60} increases from $0 - 1.2s$. On the other hand, in [7], following from results obtained in noisy reverberant environments an “interaction effect” is reported whereby varying reverberation levels have a significant effect on intelligibility scores only when the $ASNR$ is low. Conversely,

"If there are minimal (noise) problems the effect of room acoustics, from very bad to near optimum changes speech intelligibility by no more than 2.5%" [7]

For the reasons we have discussed, the separate and independent treatment of noise and reverberation, as occurs in [7], leads to erroneous results. We, therefore, prefer the results in [12] showing that reverberation in the absence of noise retains the potential to significantly reduce the intelligibility of speech.

Two further phenomena relating to the perceptual impact of noise and reverberation have particular relevance for our application. The first of these is the effect on speech intelligibility due to spatial information. In [12], intelligibility scores are obtained by listeners, across a range of ages, listening to reverberant recordings played by headphones. At an RT_{60} of $0.4s$, these are $2 - 4\%$ higher for binaural (stereo) sound presentation than for the monaural case. At an RT_{60} of $1.2s$ they are $5 - 9\%$ higher. In [14], it is shown that, when noise and speech signals are generated by two spatially separated loudspeakers at right angles to each other, an increase in intelligibility is obtained compared to the case when both noise and speech are produced by the same loudspeaker. This increase is equivalent to that achieved by a $9.6dB$ increase in the SNR.

We may expect, therefore, that participants in a remote room, listening to a single-channel microphone recording, will perceive greater levels of signal degradation than participants in the same room as the talker.

Secondly, we note the particularly off-putting effect of acoustic echo. This is significantly greater than that of other noise of similar intensity. As an illustration of this consider the required acoustic echo suppression specified by GSM (Global System for Mobile communications) protocols – $46dB$ compared with an $ASNR$ of $25dB$ described as being “ideal” for ambient noise levels in classroom environments in [10].

2.5.2 Quality

Perceived quality is far less well-understood than intelligibility, mainly due to the difficulty inherent in quantifying “good” or “bad” quality. In addition individuals will have highly diverging opinions as to what constitutes “good” or “bad” quality speech.

Noise, it may safely be assumed, reduces sound quality. We note that we are aware of the existence of “acoustic conditioning” strategies whereby ambient noise is masked by less irritating, deliberately injected noise. However, acoustic conditioning is in very limited use and it may be assumed not to be present in classrooms.

Reverberation on the other hand, has long been held to improve sound quality – notably when referring to spaces used to host musical performances. In [8], Haas reports early reflections as causing speech to be perceived as more “pleasant” by listeners.

Other research, however, demonstrates a clear preference among listeners for less reverberant sound. In [15], listeners were presented with recordings of speech convolved with simulated room impulse responses and asked to rank their quality using a 9-point scale (9 being “excellent”, 1 being “unsatisfactory”). The speech bandwidth was $4kHz$. The resulting mean opinion scores were then compared with the acoustic conditions under which they were obtained and a predictor of speech quality derived. This predictor is the equation shown below

$$\text{Predicted Preference} = 7.94 - 0.46\sigma - 4RT_{60} \quad (2.9)$$

where σ is the standard deviation of the log-amplitude of the frequency response of the room. The average listener preference is, therefore, for less reverberant speech.

2.6 Discussion

In chapter 1, we outlined the requirements for a classroom-based videoconferencing application as mandated by the need to facilitate natural and spontaneous interaction while, at the same time, using moderate amounts of equipment and requiring only minimal technical know-how. Those included the use of full-duplex communication protocols, multiple loudspeakers and remote microphones. This chapter will have made apparent the significant problems that these requirements pose.

Direct-path sound attenuates as it travels. Therefore, in moving a microphone farther away from a talker, we are likely to be reducing both the SNR and DRR. Furthermore, full-duplex communication is susceptible to howl and acoustic echo. The latter is of particular concern as the traditional methods for its suppression fail when multiple loudspeakers produce correlated signals. We have seen the effects of noise, reverberation and, in particular, acoustic echo to be highly detrimental to both the intelligibility and perceived quality of speech. We have also noted that these effects are compounded when spatial information is lost, such as occurs when noisy speech is recorded by a single microphone.

What are required, therefore, are techniques by which we may suppress the noise and reverberation present in a recording while faithfully maintaining a target speech

signal. Traditional bandpass filtering will be of only limited use in this regard. Speech and ambient noise are both broadband signals and, as such, will have frequency spectra which overlap considerably. In the case of reverberation and acoustic echo, their long-term power spectra will be virtually identical to the target speech.

As we have seen, at certain levels and in certain conditions reverberation can, in fact, improve speech quality and intelligibility. However, while it is possible that, in any given scenario, the reverberation present is beneficial, it is almost certain that in some if not most cases, the reverberation will serve to degrade speech. Therefore, we require the means to suppress it.

In this thesis we will investigate methods by which we may enhance the perceived quality of recorded speech using the data provided by arrays of spatially distributed microphones. In the next chapter, we develop the fundamental theory underpinning such microphone-array-processing.

Chapter 3

Array-Processing Theory

3.1 Introduction

In this chapter, we explore the fundamental theory behind array-signal-processing. In section 3.2 we discuss our subsequent treatment of a propagating wave as a function of both space and time - a spatiotemporal signal. In section 3.3 we describe and explain the consequences inherent in sampling such a signal at discrete points in space using an array of sensors. In section 3.4 we focus on the theory behind filter-and-sum array-signal-processing, whereby sensor outputs are filtered before being combined to form an overall system output. We undertake this specific investigation due to the near-ubiquity and, hence, importance of filter-and-sum strategies in techniques for speech-enhancement and source-localization. We conclude with section 3.5, in which we outline the signal models to be used in the remainder of the thesis.

The analysis presented here is, in large part, adapted from [16]. For readers who wish to explore array-signal processing further, the author also recommends [17].

3.2 Spatiotemporal Signals

In digital signal processing, we are familiar with the concept of a waveform being a function of time. Observations of propagating waves, however, will depend not only upon the time but also the location at which they are made. For example, our observation of the soundwaves produced by a remote source, at any instant, will be a time-delayed version of that which we would have observed at the same instant and at some point closer to the source. This is due to the extra time required for the wave to propagate the farther distance. We are, therefore, concerned with spatiotemporal signals. The “wavefield”, $f(x, y, z, t)$, is such a signal. The variables x , y and z describe location with reference to a 3-dimensional Cartesian coordinate grid. For conciseness,

we shall alternatively use the vector notation whereby $(x, y, z) \equiv \vec{x}$.

Let us assume that the wavefield consists of a single, non-attenuating, monochromatic, plane wave, propagating in a direction described by the unit vector \vec{v}_0 (see figure (3.1)). Doing this shall allow us to introduce some useful terms and concepts. We define $f_0(t)$, the observation of the wavefield at the origin.

$$f_0(t) = f(0, 0, 0, t) \quad (3.1)$$

A monochromatic wave has a single frequency component, ω . We may, therefore, express $f_0(t)$ as follows

$$f_0(t) = A \exp\{j\omega t\} \quad (3.2)$$

where A is a complex scalar. A plane wave has constant phase at every point on any plane perpendicular to the direction of propagation. Furthermore, the amplitude of a non-attenuating wave remains constant as the wave propagates. Therefore, we may obtain an expression for the wavefield in terms of $f_0(t)$, as shown below

$$f(\vec{x}, t) = f_0\left(t - \frac{\vec{x} \cdot \vec{v}_0}{c}\right) \quad (3.3)$$

where c is the speed with which the wave propagates. The term $\frac{\vec{v}}{c}$ is traditionally referred to as the “slowness vector” and has units of sm^{-1} . From this

$$\begin{aligned} f(\vec{x}, t) &= A \exp\left\{j\omega\left(t - \frac{\vec{x} \cdot \vec{v}}{c}\right)\right\} \\ &= A \exp\{j(\omega t - \vec{x} \cdot \vec{k})\} \end{aligned} \quad (3.4)$$

where $\vec{k} = \frac{\omega \vec{v}}{c}$ and is known as the “wavenumber vector”. The wavenumber vector has units of radians per meter and may, therefore, be considered as a spatial analog of frequency. Following the previous convention, $\vec{k} \equiv (k_x, k_y, k_z)$

3.2.1 Multidimensional Fourier Transform

The Fourier Transform is a well-known technique that is widely used in single-dimensional digital signal processing. It may also be easily expanded for use in the multi-dimensional cases under investigation. The (non-unitary) 4-dimensional Fourier Transform of $f(x, y, z, t)$ is shown below.

$$F(k_x, k_y, k_z, \omega) = \iiint_{-\infty}^{\infty} f(x, y, z, t) \exp\{-j\omega t\} \exp\{-jk_x x\} \exp\{-jk_y y\} \exp\{-jk_z z\} dx dy dz dt \quad (3.5)$$

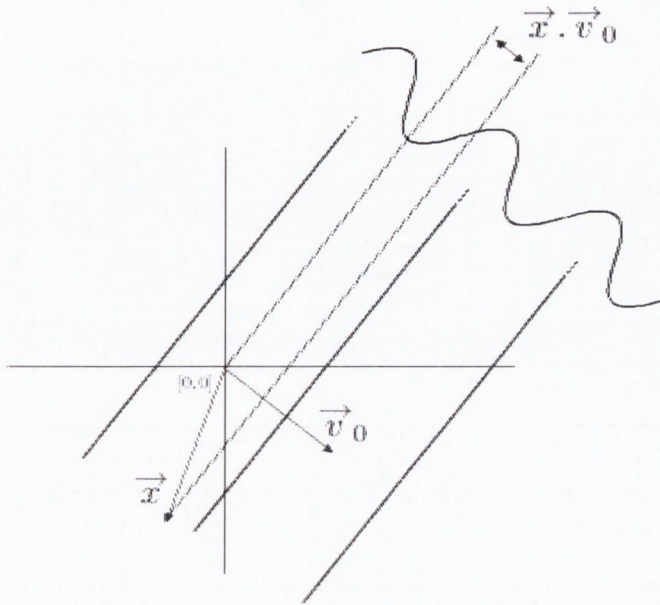


Figure 3.1: A planar, non-attenuating wave with direction of propagation \vec{v} . $f(\vec{x}, t)$, the wavefield at \vec{x} , may be expressed in terms of $f(\vec{0}, t)$, where $\vec{0}$ is the origin.

or more concisely as

$$F(\vec{k}, \omega) = \iint_{-\infty}^{\infty} f(\vec{x}, t) \exp\{-j(\omega t + \vec{k} \cdot \vec{x})\} d\vec{x} dt \quad (3.6)$$

We adopt the convention of changing the sign of the $\vec{k} \cdot \vec{x}$ term. As we are integrating between $\pm\infty$, this does not alter our result.

$$F(\vec{k}, \omega) = \iint_{-\infty}^{\infty} f(\vec{x}, t) \exp\{-j(\omega t - \vec{k} \cdot \vec{x})\} d\vec{x} dt \quad (3.7)$$

Similarly, the inverse Fourier transform may be written as follows

$$f(\vec{x}, t) = \frac{1}{(2\pi)^4} \iint_{-\infty}^{\infty} F(\vec{k}, \omega) \exp\{j(\omega t - \vec{k} \cdot \vec{x})\} d\vec{k} d\omega \quad (3.8)$$

We may interpret equation (3.8) as showing that an arbitrary wavefield may be considered as a superposition of a (possibly infinite) number of weighted exponential plane waves.

3.3 Array Signal Processing

An array is composed of multiple sensors at discrete locations in space. The array itself may, therefore, be considered a function of space, which we denote $w(\vec{x})$. We may express $w(\vec{x})$ as follows

$$w(\vec{x}) = \sum_{m=0}^{M-1} w_m \delta(\vec{x} - \vec{x}_m) \quad (3.9)$$

where M is the number of sensors, \vec{x}_m is the location of the m^{th} sensor, $w_m = w(\vec{x}_m)$ and $\delta(\vec{x})$ is defined such that

$$\int f(\vec{x}, t) \delta(\vec{x}) d\vec{x} = f_0(t) \quad (3.10)$$

We may obtain the multidimensional Fourier Transform of $w(\vec{x})$ as before

$$W(\vec{k}) = \int_{-\infty}^{\infty} w(\vec{x}) \exp\{j \vec{k} \cdot \vec{x}\} d\vec{x} \quad (3.11)$$

Inserting (3.9) yields

$$W(\vec{k}) = \sum_{m=0}^{M-1} w_m \exp\{j \vec{k} \cdot \vec{x}_m\} \quad (3.12)$$

We refer to $W(\vec{k})$ as the Array Pattern. We also define $\widehat{W}(\vec{k})$ as being the Array Pattern where $w_m = 1 \forall m$ (in general w_m need not equal 1). As such, $\widehat{W}(\vec{k})$ is a function of the location of the sensors only. Later in this chapter we shall demonstrate how fundamental quantities in array processing may be expressed in terms of $\widehat{W}(\vec{k})$, thus showing the strong influence of array geometry.

3.3.1 Spatial Sampling

Many readers will be familiar with the concept of sampling in the context of discrete-time sampling of temporal signals. Indeed, much of our previous discussion may have struck some as being highly analogous to time sampling. It will come as no surprise, then, that well-known consequences of time-sampling - those of spectral smoothing and frequency aliasing - have analogous counterparts when performing spatial-sampling.

Using the array, we may sample the wavefield at multiple discrete locations. The observed wavefield, $z(\vec{x}, t)$, may be expressed as the product of the wavefield and the array function

$$z(\vec{x}, t) = w(\vec{x}) f(\vec{x}, t) \quad (3.13)$$

Calculating the multidimensional Fourier Transform of this relationship

$$Z(\vec{k}, \omega) = \frac{1}{(2\pi)^3} \int W(\vec{k} - \vec{l}) F(\vec{l}, \omega) d\vec{l} \quad (3.14)$$

We shall restrict the following analysis to the special case of linear, equispaced arrays (LEAs; see figure (3.2a)). We assume each element of the array to be on the x -axis. As a result, $w(\vec{x})$ is zero for non-zero values of y and z . This allows us to drop the vector notation and write $w(x)$. Similarly, for variables in the wavenumber domain, we write $W(k_x)$. The ease with which such arrays may be analyzed shall allow us to obtain valuable insights. In addition, since any arbitrary array geometry may be decomposed into linear, equispaced components (even if each component contains only two sensors) these insights may be applied to the development of our understanding of more general cases. We reformulate (3.13) as shown below, where we express the finite, discrete array function $w(x)$ as the product of an infinite pulse train, $w_s(x)$, and a continuous windowing function, $w_c(x)$.

$$z(x, t) = w(x)f(x, t) = w_c(x)w_s(x)f(x, t) \quad (3.15)$$

where

$$w_s(x) = \sum_{n=-\infty}^{\infty} \delta(x - nd) \quad (3.16)$$

$$w_c(x) = \left\{ \begin{array}{ll} 1 & (\frac{1-M}{2})d \leq x \leq (\frac{M-1}{2})d \\ 0 & \text{otherwise} \end{array} \right\} \quad (3.17)$$

and d is the intersensor spacing. In the wavenumber-frequency domain

$$Z(k_x, \omega) = \frac{1}{(2\pi)^2} W_c(k_x) * W_s(k_x) * F(k_x, \omega) \quad (3.18)$$

where $*$ is the convolution operator. The Fourier Transform of a pulse train is also a pulse train

$$W_s(k_x) = \frac{2\pi}{d} \sum_{l=-\infty}^{\infty} \delta(k_x - \frac{l2\pi}{d}) \quad (3.19)$$

From this

$$Z(k_x, \omega) = \frac{1}{2\pi d} F(k_x, \omega) * \sum_{l=-\infty}^{\infty} W_c(k_x - \frac{l2\pi}{d}) \quad (3.20)$$

The observed wavefield is, therefore, seen to be a circular convolution of $W_c(k_x)$ and $F(k_x, \omega)$.

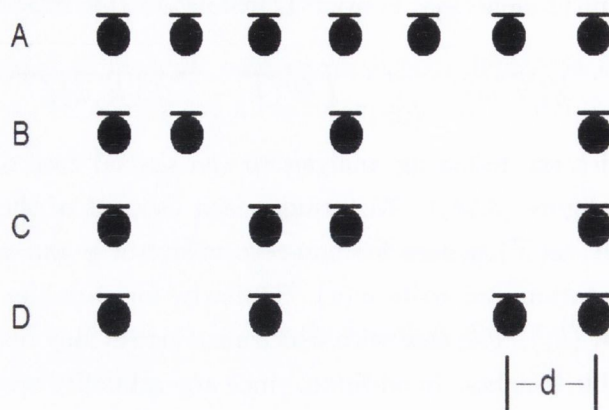


Figure 3.2: Linear arrays; A shows a 7-element LEA; B-D are “sparse” arrays. The minimum intersensor spacing is d .

3.3.2 Spatial Aliasing

Figure (3.3) plots of $|\widehat{W}(k_x)|$ for a 7-element LEA ($w_m = \frac{1}{7}\forall m$). In order to avoid spatial aliasing, $F(k_x, \omega)$ must be bandlimited such that

$$-\frac{\pi}{d} < k_x < \frac{\pi}{d} \quad (3.21)$$

From the upper bound

$$\max\{k_x\} = \max\left\{\frac{\omega v_x}{c}\right\} = \frac{\pi}{d} \quad (3.22)$$

Maximizing by letting $v_x = 1$, we get Nyquist’s result giving a lower bound for permissible intersensor spacing.

$$d = \frac{c\pi}{\omega_{\max}} = \frac{\lambda_{\min}}{2} \quad (3.23)$$

where λ_{\min} is the wavelength of the maximum frequency being sampled, ω_{\max} .

3.3.3 Wavenumber Smoothing

In addition to being periodic, $Z(k_x, \omega)$ is also “smoothed” in the wavenumber domain by convolution with $W_c(k_x)$. Following the well-known Fourier relationship, if $w_c(x)$ is a square “window”, $W_c(k_x)$ is a sinc function

$$W_c(k_x) = ((M - 1)d) \operatorname{sinc}\left(\frac{k_x(M - 1)d}{2}\right) \quad (3.24)$$

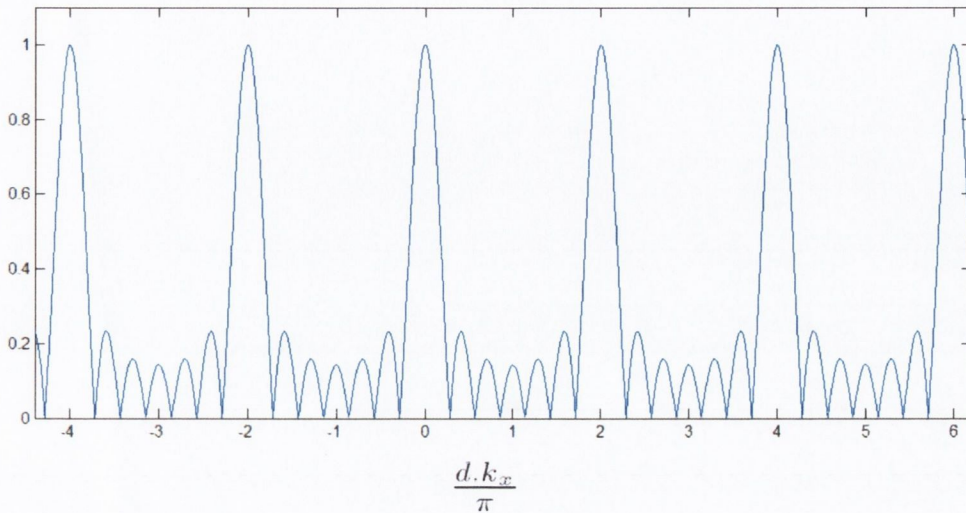


Figure 3.3: $|\widehat{W}(k_x)|$ for a 7-element LEA. The “lobe” at $\frac{dk_x}{\pi} = 0$ is referred to as the mainlobe. The other periodically repeating lobes are known as “grating” lobes - a name with its origins in experiments where light is passed through a grating to produce an interference pattern with multiple periodically repeating bright lines.

In virtually all array processing algorithms our success or otherwise will depend upon our ability to resolve an observed wavefield’s component wavenumber vectors. In signal enhancement, we must be able to isolate and enhance (or suppress) a single propagating wave. In detection or source localization, we must be able to accurately determine the wavenumber vector corresponding to an observed signal. To achieve this resolution, the degree to which $Z(k_x, \omega)$ is “smoothed” must be sufficiently small. Letting A be a real scalar, $\text{sinc}(Ak_x) \rightarrow \delta(k)$ as $A \rightarrow \infty$. From inspection of (3.24), therefore, it is apparent that resolution improves as the term $(M - 1)d$ increases - i.e. as the array width increases.

For optimal resolution, we require the dimensions of the array to be as large as possible. However, in practice array dimensions are constrained by other considerations. As we shall see, this, in turn, constrains the performance of array-processing algorithms.

One such consideration is that of hardware costs, which will place a limit on the number of sensor which may be used. However, in seeking the best placing for a finite number of sensors, an array designer may consider a trade-off whereby we increase the intersensor spacing. This increases the array width and reduces smoothing but also increasing spatial aliasing if $d > \frac{\lambda_{\min}}{2}$.

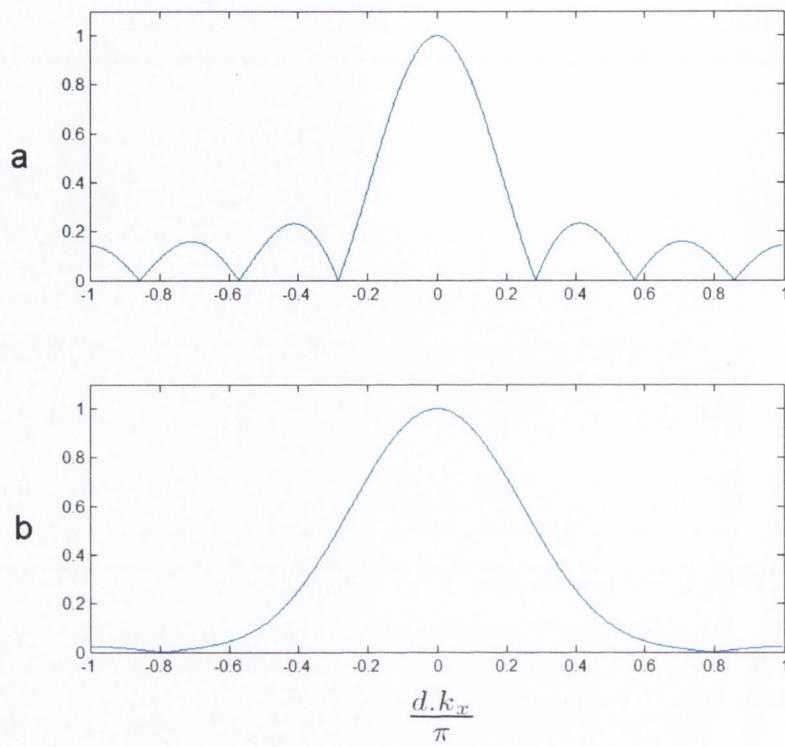


Figure 3.4: $|W(k_x)|$ for a 7-element LEA corresponding to a “square” window function, (a), and a Hamming window function, (b).

3.3.4 Weighting

We may control the smoothing exhibited by $Z(k_x, \omega)$ by varying the “weightvector”, $\{w_m\}$, in a way similar to windowing in discrete-time processing. Figure (3.4a) plots $|W(k_x)|$ corresponding to an 7-element array. The weightvector is $[\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}]$. Figure (3.4b) plots $|W(k_x)|$ corresponding to a weightvector of $[\text{.024}, \text{.093}, \text{.231}, \text{.301}, \text{.231}, \text{.093}, \text{.024}]$, which some readers may recognize as a 7-tap Hamming window. A comparison of the two plots clearly shows that by choosing the second weightvector we obtain reduced sidelobes at the expense of a wider mainlobe. Such weighting would be appropriate if we were concerned with the effects of the sidelobes while the former weighting would allow us to resolve closely spaced wavenumbers more effectively.

Alternatively, we may narrow the mainlobe at the expense of increased sidelobes. In doing so we may exploit a portion of the wavenumber axis known as the “invisible region”. At frequencies where $d < \frac{\lambda}{2}$, we are spatially over-sampled - i.e. we are sampling more frequently than is required to prevent spatial aliasing. In such situations

the values of k_x between $\frac{\omega_{\max}}{c}$ and $\frac{\pi}{d}$ define those wavenumbers that cannot correspond to propagating waves. Through appropriate weighting, we may narrow the mainlobe of $W(k_x)$. This will increase the size of the sidelobes, however this will not matter if the sidelobes are in the invisible region, where no waves of interest exist. By oversampling, we may achieve enhanced resolution known as “superdirectivity”, a fact that is exploited by many array processing algorithms.

3.3.5 Sparse Arrays

Certain non-equispace arrays may also be understood in terms of weighting. Consider the “sparse” array geometries shown in figure (3.2). Each resembles an LEA where certain sensors have been removed. A step equivalent to removing the m^{th} sensor is to let $w_m = 0$. Plots of $|\widehat{W}(k_x)|$ corresponding to the sparse arrays are shown in figure (3.5) and are compared to that of a 4-element LEA. Sparse arrays are popular in the literature as, for a fixed number of sensors, they offer a wider array extent while avoiding aliasing.

3.4 Filter-and-Sum Array Signal Processing

The large majority of array processing algorithms (to be reviewed in chapters 4 and 5) follow a Filter-and-Sum (F&S) paradigm whereby the observations at each sensor undergo temporal filtering before being combined to form the system output, figure (3.6). Given the ubiquity of F&S systems in array processing, we find it useful to characterize their behavior. For the moment we shall assume ideal, omnidirectional sensors with frequency-invariant unity gain. Furthermore, we shall assume that the response of each of the FIR filters corresponds to a simple frequency-invariant scaling and delay. We may express the output of a F&S system as the summation of multiple delayed and scaled microphone outputs.

$$z(t) = \sum_{m=0}^{M-1} w_m y_m(t - \Delta_m) \quad (3.25)$$

where $y_m(t)$ is the output of the m^{th} sensor and Δ_m is the delay applied by the m^{th} FIR filter. From equation (3.8) we see that a wavefield may be considered to be a superposition of a possibly infinite number of monochromatic, non-attenuating plane waves. We therefore characterize the system in terms of its response to a single such wave with frequency ω_0 and slowness vector $\frac{\vec{v}_0}{c}$

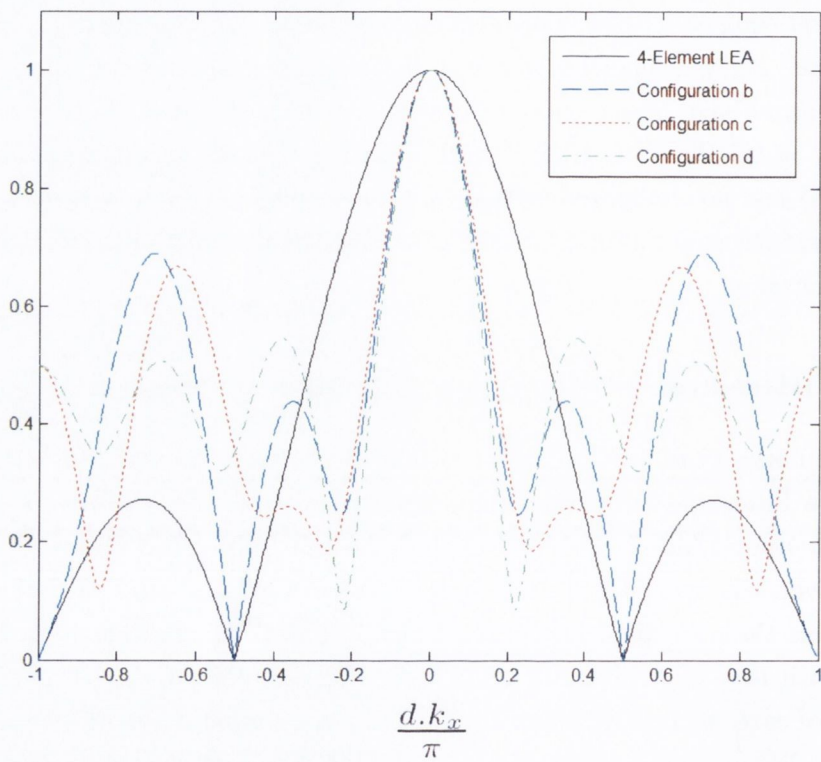


Figure 3.5: $|\widehat{W}(k_x)|$ corresponding to the sparse array geometries in figure (3.2). Compared to a 4-element LEA they achieve a narrower mainlobe at the expense of larger sidelobes.

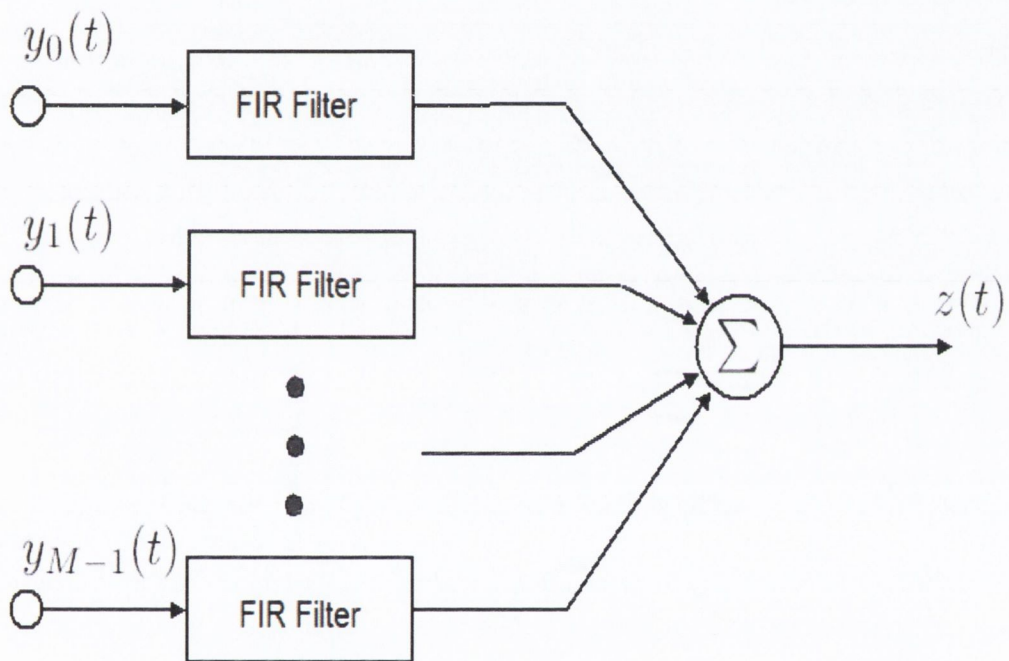


Figure 3.6: Filter and Sum: Sensor outputs undergo temporal filtering before being combined.

$$\begin{aligned}
 z(t) &= \sum_{m=0}^{M-1} w_m \exp \left\{ j\omega_0 \left(t - \Delta_m - \frac{\vec{v}_0}{c} \cdot \vec{x}_m \right) \right\} \\
 &= \left[\sum_{m=0}^{M-1} w_m \exp \left\{ j\omega_0 \left(-\Delta_m - \frac{\vec{v}_0}{c} \cdot \vec{x}_m \right) \right\} \right] \exp \{ j\omega_0 t \}
 \end{aligned} \tag{3.26}$$

The term in the square brackets may be said to be the wavenumber-frequency response of the system.

3.4.1 Steering

From inspection of (3.26), we see that we may alter the wavenumber-frequency response of an F&S system by varying the delays applied to the sensor outputs. Indeed, many array processing techniques are based upon a judicious choice of delays. For example, let us assume that $\Delta_m = -\frac{\vec{v}_0}{c} \cdot \vec{x}_m$. This, in effect, time-aligns our observations of the propagating wave in the output of each sensor. This time alignment is commonly referred to as “steering”. Following from our assumption, the wavenumber-frequency response is $\sum w_m$. Thus, by steering we may apply some desired gain to a specified propagating wave. Traditionally, steering maximizes the system response for the specified wave, thus enhancing it relative to the remainder of the sampled wavefield. This is known as “beam-steering” or “beamforming” and is *the* classical array processing technique. Alternatively we may suppress a specified wave. This is known as “null-steering” and occurs when $\sum w_m = 0$. Other steering applications include source localization, a simple implementation of which involves beamforming in multiple directions. The direction that maximizes the system output is then taken as corresponding to some active source. We shall review the applications which exploit steering more fully in chapters 4 and 5. For the moment, having highlighted its importance, we shall discuss those factors effecting our ability to steer with accuracy.

3.4.2 Discrete-Time Sampling

So far, we have considered time to be a continuous variable. In reality, we are likely to be using discrete-time sampling of the wavefield. Discrete-time sampling has important but well-understood implications. We shall not attempt to describe these fully here but, rather, shall limit our discussion to those implications that are particular to array processing.

Of these, perhaps the most significant is related to steering. In applying delays, we are restricted to a finite set of discrete times, nT - where n is some integer and T is

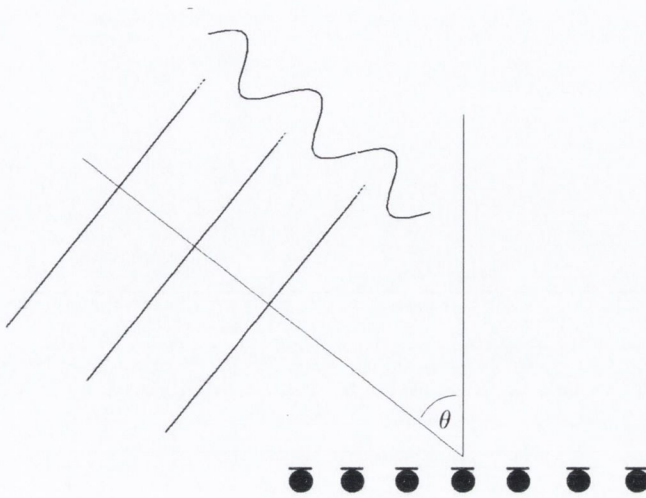


Figure 3.7: A planar wave propagating across a LEA. The angle of incidence, θ , is commonly referred to as the direction of arrival of the wave.

the temporal sampling period - which may or may not exactly equal Δ_m . We therefore introduce error, ε_m , into our steering.

$$\Delta_m = \varepsilon_m + n_m T \quad (3.27)$$

Replacing Δ_m in (3.26)

$$\begin{aligned} z(t) &= \sum_{m=0}^{M-1} w_m \exp \left\{ j\omega_0 \left(t - n_m T - \frac{\vec{v}_0 \cdot \vec{x}_m}{c} \right) \right\} \\ &= \sum_{m=0}^{M-1} w_m (\exp \{ j\omega \varepsilon_m \}) \exp \left\{ j\omega_0 \left(t - \Delta_m - \frac{\vec{v}_0 \cdot \vec{x}_m}{c} \right) \right\} \end{aligned} \quad (3.28)$$

The result is, in effect, a complex-valued and position-dependant scaling of sensor weights which distorts the response of the system.

3.4.3 Angular Resolution

In the classical array-processing scenario, we are concerned with a plane wave propagating across a LEA. In such a scenario, steering is commonly referred to in terms of the “direction-of-arrival” (DOA) of the wave - that being the angle of incidence formed by the wave with the axis of the array, figure (3.7). Let us consider a 2-element array with intersensor spacing d , steering toward a farfield source with DOA θ . To

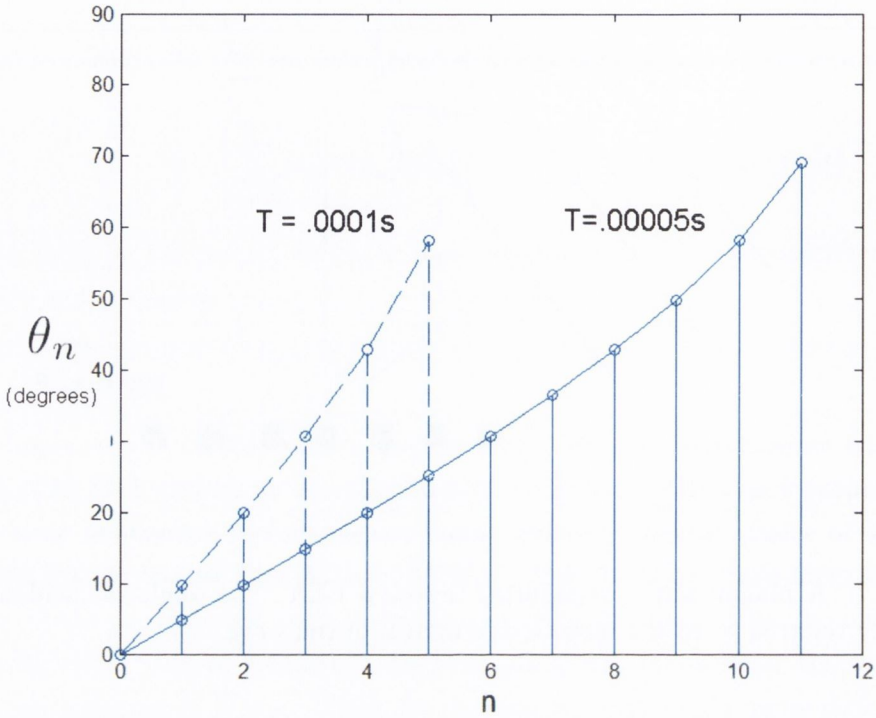


Figure 3.8: Steerable angles for a LEA where $d = 0.2m$, and $c = 340ms^{-1}$. Increasing the sampling rate (reducing T) increases the density of the steerable angles.

time-align the outputs of the sensors we apply a delay to the output of one sensor of magnitude Δ , where

$$\Delta = \frac{d \sin \theta}{c} \quad (3.29)$$

We are, however, restricted to discrete delays, nT , and hence discrete angles $\{\theta_n\}$. Substituting Δ and performing simple algebraic manipulation on (3.29)

$$\theta_n = \arcsin \left\{ \frac{cnT}{d} \right\} \quad (3.30)$$

Successful steering may only be achieved when some steerable angle θ_n is sufficiently close to the true DOA. For illustrative purposes, we let $d = 0.2m$, and $c = 340ms^{-1}$ (corresponding to an acoustic wave in air). Figure (3.8) shows the steerable angles between 0° and 90° for $T = 1 \times 10^{-4}s$ and $T = 5 \times 10^{-5}s$. We clearly see that, due to the non-linear nature of the arcsin function, the discrete steerable angles are not, themselves, evenly spaced. Rather, $(\theta_{n+1} - \theta_n)$ increases with n .

We see also, that the number of steerable angles increases as T gets smaller. The limits of n , and hence the limits on the numbers of steerable angles, may be found by maximizing and minimizing (3.30), i.e. we let $\sin \theta_n = 1$, giving us

$$n_{\max} = \text{floor} \left\{ \frac{d}{cT} \right\} \quad (3.31)$$

where $\text{floor}\{X\}$, rounds X down to the nearest integer. Remembering that we must include $n = 0$, we may therefore steer in $n_{\max} + 1$ directions, within the angular boundary $0^\circ \leq \theta \leq 90^\circ$.

Let us assume that d is just small enough to prevent spatial aliasing (i.e. $d = \frac{\lambda_{\min}}{2}$, see section 3.3.2). Following from the previous equation

$$\frac{2cn_{\max}}{\lambda_{\min}} = \frac{1}{T} = \frac{n_{\max}\omega_{\max}}{\pi} \quad (3.32)$$

In other words, when we sample spatially at the Nyquist rate, we must sample temporally at n_{\max} times the Nyquist rate to steer in $n_{\max} + 1$ directions, within the angular boundary $0^\circ \leq \theta \leq 90^\circ$. The requirement for such high sampling rates may, in many cases, be prohibitive. As an alternative, we can employ time domain interpolation or frequency-domain steering.

3.4.4 Interpolation

The classical approach to interpolating a time-signal, by some factor I , is to intersperse each sample with $I - 1$ zeros. We then pass the resulting zero-padded signals through low-pass filters to smooth out the waveforms and obtain “upsampled” signals which we may steer and add as required. Alternatively we may exploit an efficient polyphase interpolation structure which we shall now derive. To avoid confusion, we shall denote with a $\hat{}$ all signals having the upsampled sampling rate. Letting $\hat{g}(n)$ denote a suitable lowpass filter and $\hat{u}_m(n)$ be the zero-padded $y_m(n)$

$$\hat{y}_m(nI + r) = \sum_i \hat{u}_m(i) \hat{g}(nI + r - i) \quad (3.33)$$

We may obtain an equivalent but more efficient computation of the interpolated output by summing over only non-zero values of $\hat{u}_m(n)$

$$\hat{y}_m(nI + r) = \sum_l y_m(l) \hat{g}((n - l)I + r) \quad (3.34)$$

From the expression above, we see that $\hat{y}_m(nI + r)$ is, in effect, a convolution of $y_m(n)$ and every I^{th} coefficient of the lowpass filter, starting with the r^{th} coefficient. We define the following

$$g_r(n) = \hat{g}(nI + r) \quad (3.35)$$

and inserting this into (3.34)

$$\hat{y}(nI + r) = \sum_l y_m(l) g_r((n - l)I) \quad (3.36)$$

The output of a steered F&S system is given by

$$\begin{aligned} \hat{z}(nI + r) &= \sum_{m=0}^{M-1} w_m \hat{y}_m(nI + r - n_m) \\ &= \sum_{m=0}^{M-1} w_m \left[\sum_l y_m(l) g_r((n - l)I) \right] \end{aligned} \quad (3.37)$$

Denoting $r - n_m$ modulo I as $(r - n_m)_I$, we may write

$$\hat{z}(nI + r) = \sum_{m=0}^{M-1} w_m \left[\sum_l y_m(l) g_{(r-n_m)_I}(n - l) \right] \quad (3.38)$$

And thus we obtain an efficient polyphase method for interpolation steering. Downsampling by a factor I is easily achieved by setting r constant (say, for simplicity, $r = 0$). The downsampled system output may then be calculated by

$$z(n) = \sum_{m=0}^{M-1} w_m \left[\sum_l y_m(l) g_{(-n_m)_I}(n - l) \right] \quad (3.39)$$

We see, therefore, that interpolation steering may be achieved without upsampling.

3.4.5 Steering in the Frequency Domain

Steered F&S array processing applications may also be implemented in the frequency domain. For practical applications using discrete-time data, we apply a discrete Fourier transform (DFT) to short (and possibly overlapping) segments of recorded data. For simplicity and clarity, we shall assume only one segment of N samples per sensor, although we may easily extend our analysis to include multiple consecutive segments. The DFT of the data segment $y_m(n)$, $n \in \{0 : N - 1\}$, is defined as

$$Y_m(v) = \sum_{n=0}^{N-1} \tilde{w}(n) y_m(n) \exp \left\{ -\frac{j2\pi v n}{N} \right\} \quad (3.40)$$

where v is the frequency index and $\tilde{w}(n)$ is the temporal window. The DFT of $z(n)$ may be expressed as

$$Z(v) = \sum_{m=0}^{M-1} w_m Y_m(v) \exp \left\{ -\frac{j2\pi v \Delta_m}{N \cdot T} \right\} \quad (3.41)$$

The exponent applies a phase shift to $Y_m(v)$ corresponding to a delay Δ_m . Inserting (3.40)

$$Z(v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} w_m \tilde{w}(n) y_m(n) \exp \left\{ -\frac{j2\pi v n}{N} \right\} \exp \left\{ -\frac{j2\pi v \Delta_m}{N \cdot T} \right\} \quad (3.42)$$

We obtain a computationally efficient two-dimensional DFT expression whenever the following holds

$$\frac{2\pi v \Delta_m}{N \cdot T} = \frac{2\pi \eta}{M} m \quad (3.43)$$

or, equivalently, when

$$\Delta_m = \frac{\eta N T}{v M} m \quad (3.44)$$

where η is some integer. Letting $x(m, n) = w_m \tilde{w}(n) y_m(n)$

$$\begin{aligned} Z(v) &= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(m, n) \exp \left\{ -\frac{j2\pi v n}{N} \right\} \exp \left\{ -\frac{j2\pi \eta m}{M} \right\} \\ &= X(\eta, v) \end{aligned} \quad (3.45)$$

The integer η may therefore be considered as denoting the wavenumber index (just as v denotes the frequency index). A time domain output may be obtained following the inverse DFT operation, $DFT^{-1}[X(\eta, v)]$. We must be careful in obtaining the time-domain output. Remembering the well-known DFT relationship

$$DFT^{-1}[X(\eta, v)] \exp \left\{ -\frac{j2\pi \eta \Delta_m}{N \cdot T} \right\} \stackrel{DFT}{\leftrightarrow} x(n - \Delta_m)_N \quad (3.46)$$

and assuming, for illustration, that $\Delta_m = n_m T$, we may write the following

$$DFT^{-1}[Z(v)] = \sum_{m=0}^{M-1} w_m \tilde{w}(n - n_m)_N y_m(n - n_m)_N \quad (3.47)$$

The time-domain output, is subject to circular shifting. Furthermore, while the delays $\{n_m T\}$ will time-align the sensor outputs $\{y_m(n)\}$, they will mis-align the temporal windowing function $\tilde{w}(n)$. As a result, we shall observe distortion at the edges of the window. However, while we should be aware of this distortion, a variety

of techniques exist by which we may mitigate against it, allowing us to perform steered F&S operations in the frequency domain, albeit cautiously.

To steer a beam at an angle θ , η and v must follow the relationship

$$\eta = \frac{vMd}{NTc} \sin \theta \quad (3.48)$$

In general, this relationship is not an absolute requirement. In not adhering to (3.48) we are merely steering multiple distinct simultaneous beams, one corresponding to each frequency index. Presuming, however, that we wish to steer in one direction only

$$\theta = \sin^{-1} \left\{ \frac{\eta NTc}{vMd} \right\} \quad (3.49)$$

We may increase the number of steerable angles by increasing the number of sensors, M . In practice, this is rarely an available option, however, we may artificially increase M provided that the corresponding $\{w_m\}$ are zero. This is analogous to the common step of zero-padding the boundaries of time signal segments. As with time-domain zero-padding or interpolation, such a step will not increase wavenumber resolution - this would require a wider array.

3.4.6 Non-Ideal Sensors

So far, our analysis has assumed that we are using sensors that sample the wavefield at points in space of infinitely small volume. In reality, no such sensor will exist. Rather, practical sensors such as microphones, aerials and telescopes are continuous apertures with non-zero physical dimensions. As such they behave as spatial filters, scaling and integrating simultaneous observations of the wavefield within some (small) volume of space.

Furthermore, we have treated the sensors as having a frequency-invariant response. Once again, this will not be the case in practice. We must therefore consider each sensor as being a spatiotemporal filter. The output of the m^{th} sensor may be expressed as a space-time convolution of the wavefield and the sensor's spatiotemporal response, $\xi_m(\vec{x}, t)$.

$$\begin{aligned} y_m(t) &= \iint \xi_m(\vec{\chi}, \tau) f(\vec{x} - \vec{\chi}, t - \tau) d\vec{\chi} d\tau \\ &= \exp\{j(\omega_0 t - \vec{k}_0 \cdot \vec{x})\} \iint \xi_m(\vec{\chi}, \tau) \exp\{-j(\omega_0 \tau - \vec{k}_0 \cdot \vec{\chi})\} d\vec{\chi} d\tau \end{aligned} \quad (3.50)$$

The integration is equivalent to a 4-dimensional Fourier Transform. Letting $\Xi_m(\vec{k}, \omega)$ denote the Fourier transform of $\xi_m(\vec{x}, t)$,

$$y_m(t) = \Xi_m(\vec{k}_0, \omega_0) \exp\{j(\omega_0 t - \vec{k}_0 \cdot \vec{x}_m)\} \quad (3.51)$$

The response of each sensor is seen to be a function of the wavenumber of the sampled wave. For a known frequency and speed of propagation, this is equivalent to a dependence upon the direction of propagation of the wave. Indeed, real sensors are most commonly referred to in terms of their "directionality" rather than their spatial filtering properties, with some well-known examples including cardioid microphones or dipole antennas.

We now consider the effect of non-ideal sensors upon the wavenumber-frequency response of a F&S system. Indeed, this is of particular importance as many array processing techniques deliberately and explicitly incorporate directional and frequency-variant sensors to achieve some desired result.

Inserting (3.50) into (3.25)

$$\begin{aligned} z(t) &= \sum_{m=0}^{M-1} w_m \Xi_m(\vec{k}_0, \omega_0) \exp\{j(\omega_0(t - \Delta_m) - \vec{k}_0 \cdot \vec{x}_m)\} \\ &= \exp\{j\omega_0 t\} \left[\sum_{m=0}^{M-1} w_m \Xi_m(\vec{k}_0, \omega_0) \exp\{-j(\omega_0 \Delta_m + \vec{k}_0 \cdot \vec{x}_m)\} \right] \end{aligned} \quad (3.52)$$

The bracketed term expresses the wavenumber-frequency response of the system. From inspection, it becomes apparent that this equation may also be used to describe the effects of FIR filters applying frequency-variant gain. We merely need let $\xi_m(\vec{x}, t)$ denote the combined response of the m^{th} sensor and filter.

For the purposes of illustration, we assume that w_m and $\Xi_m(\vec{k}_0, \omega_0)$ are identical for all sensors and, in addition, that $\Delta_m = -\frac{\vec{v}}{c} \cdot \vec{x}_m$, giving

$$z(t) = w_m \Xi_m(\vec{k}_0, \omega_0) \widehat{W}\left(\omega_0 \left(\frac{\vec{v} - \vec{v}_0}{c}\right)\right) \exp\{j\omega_0 t\} \quad (3.53)$$

This result is significant in that it makes clear the relationship that exists between $\widehat{W}(k)$ and the wavenumber-frequency response of a F&S system and thereby emphasizes the fundamental importance of array geometry (upon which, remember, $\widehat{W}(k)$ is solely dependant) to the success or otherwise of F&S-based array processing techniques.

3.5 Signal Model

In this section we develop the signal models that shall be used in the remainder of this thesis. The derivation and analysis of the techniques comprising the novel contribution

of this work shall be done in the frequency domain. This is done for reasons of clarity and because of the insight afforded by a frequency domain analysis. For consistency and ease of comparison, our discussion and description of previously published methods and techniques shall also be with reference to the frequency domain. In practice, many of these techniques are actually implemented in the time domain, however our analysis remains valid.

3.5.1 Anechoic Signal Model

In chapters 4 and 5, we shall review previously published techniques for speech-enhancement, time-delay estimation and source localization. Many of these are based on a simplified, anechoic model of sound propagation, whereby each (ideal omnidirectional) sensor receives a delayed and simply-scaled version of the target signal, $s(t)$. Following from section 2.2.1, this scaling will be proportional to the distance between the source and the microphone. The output of the m^{th} sensor may, therefore, be expressed as

$$y_m(t) = \frac{a}{|\vec{s} - \vec{m}_m|} s\left(t - \frac{|\vec{s} - \vec{m}_m|}{c}\right) + n_m(t) \quad (3.54)$$

where $n_m(t)$ is the noise at m_m , c is the speed of propagation of the wave and a is a scalar. In the frequency domain

$$Y_m(\omega) = \frac{a}{|\vec{s} - \vec{m}_m|} \exp\left\{-j\omega\left(\frac{|\vec{s} - \vec{m}_m|}{c}\right)\right\} S(\omega) + N_m(\omega) \quad (3.55)$$

For M microphones we may define the observation vector as follows.

$$\mathbf{Y}(\omega) = [Y_0(\omega), Y_1(\omega), \dots, Y_{M-1}(\omega)]^T \quad (3.56)$$

$\mathbf{N}(\omega)$ may be similarly defined. We define the intersensor time-delay $\tau_{a,b}$ as the difference in the propagation delay between m_a and m_b .

$$\tau_{a,b} = \frac{|\vec{s} - \vec{m}_b| - |\vec{s} - \vec{m}_a|}{c} \quad (3.57)$$

Taking m_0 as a reference, we may define $\tau_m = \tau_{0,m}$. We may express $\mathbf{Y}(\omega)$ as follows

$$\mathbf{Y}(\omega) = \beta \mathbf{D}(\omega) S(\omega) \exp\left\{-j\omega\left(\frac{|\vec{s} - \vec{m}_0|}{c}\right)\right\} + \mathbf{N}(\omega) \quad (3.58)$$

where the ‘‘steering vector’’ $\mathbf{D}(\omega, \vec{s})$ is given by

$$\mathbf{D}(\omega, \vec{s}) = \left[\frac{a}{\beta |\vec{s} - \vec{m}_0|}, \frac{a}{\beta |\vec{s} - \vec{m}_1|} \exp\{-j\omega\tau_1\}, \dots, \frac{a}{\beta |\vec{s} - \vec{m}_{M-1}|} \exp\{-j\omega\tau_{M-1}\} \right]^T \quad (3.59)$$

where β is some scalar chosen such that the norm of $\mathbf{D}(\omega, \vec{s})$ is unity. We may simplify the rather cumbersome signal model in (3.58) by redefining $S(\omega)$ to include the scaling and phase shift, yielding

$$\mathbf{Y}(\omega) = S(\omega) \mathbf{D}(\omega, \vec{s}) + \mathbf{N}(\omega) \quad (3.60)$$

3.5.2 Reverberant Signal Model

Modelling a reverberant acoustic environment as an LTI system, the output of m_m may be modelled as follows

$$y_m(t) = h_m(t) * s(t) + n_m(t) \quad (3.61)$$

where h_m is the source-to-microphone impulse response. In the frequency domain

$$\begin{aligned} Y_m(\omega) &= H_m(\omega) S(\omega) + N_m(\omega) \\ &= X_m(\omega) + N_m(\omega) \end{aligned} \quad (3.62)$$

Defining $\mathbf{X}(\omega)$ and $\mathbf{H}(\omega)$ in a way similar to (3.56) we may write

$$\begin{aligned} \mathbf{Y}(\omega) &= \mathbf{X}(\omega) + \mathbf{N}(\omega) \\ &= S(\omega) \mathbf{H}(\omega) + \mathbf{N}(\omega) \end{aligned} \quad (3.63)$$

Letting $\mathbf{H}_{dp}(\omega)$ and $\mathbf{H}_{mp}(\omega)$ denote the direct-path and multipath components of $\mathbf{H}(\omega)$, respectively, we may write

$$\mathbf{Y}(\omega) = S(\omega) (\mathbf{H}_{dp}(\omega) + \mathbf{H}_{mp}(\omega)) + \mathbf{N}(\omega) \quad (3.64)$$

Note that, if the source is omnidirectional, $\mathbf{H}_{dp}(\omega)$ and $\mathbf{D}(\omega, \vec{s})$ are identical to within some complex scalar (the norm of $\mathbf{H}_{dp}(\omega)$ is not constrained to be unity). If we can assume that the source is omnidirectional, we may, in many cases, use $\mathbf{H}_{dp}(\omega)$ and $\mathbf{D}(\omega, \vec{s})$ interchangeably. Throughout this thesis, we shall use $\mathbf{H}_{dp}(\omega)$ when referring to the general case where the source is potentially directional. On occasion, however, we shall assume an omnidirectional source and use $\mathbf{D}(\omega, \vec{s})$ so as to be consistent with the literature.

3.5.3 Filter-and-Sum Processing

Using a F&S architecture, FIR filters apply a frequency-dependant weighting, $W(\omega)$, to the output of each microphone. The resulting filter outputs are then added to obtain a system output, $Z(\omega)$. This process may be expressed as shown below

$$Z(\omega) = \mathbf{W}^H(\omega) \mathbf{Y}(\omega) \quad (3.65)$$

where $\mathbf{W}(\omega) = [W_0(\omega), W_1(\omega), \dots, W_{M-1}(\omega)]^T$ and is known as the “weightvector”.

3.5.4 The Spatospectral Correlation Matrix

The spatospectral correlation matrix appears repeatedly throughout the literature and is of fundamental importance in many array processing techniques. We, therefore, find it useful to define associated notation here. In addition we describe the characteristics of such matrices that are of relevance to our future discussions.

We define the spatospectral correlation matrix of $\mathbf{Y}(\omega)$ as follows

$$\mathbf{R}_{\mathbf{Y}\mathbf{Y}}(\omega) = E \{ \mathbf{Y}(\omega) \mathbf{Y}^H(\omega) \} \quad (3.66)$$

where E is the expectation operator and H denotes the Hermitian transpose. $\mathbf{R}_{\mathbf{N}\mathbf{N}}(\omega)$ and $\mathbf{R}_{\mathbf{X}\mathbf{X}}(\omega)$ may be similarly defined. In the case of the later we note that

$$\mathbf{R}_{\mathbf{X}\mathbf{X}}(\omega) = |S(\omega)|^2 \mathbf{H}(\omega) \mathbf{H}(\omega)^H \quad (3.67)$$

Of particular interest in this thesis is the $\mathbf{R}_{\mathbf{N}\mathbf{N}}(\omega)$ associated with zero-mean, Gaussian noise signals appearing in the sensor outputs and where the noise signal in the output of any one sensor is uncorrelated with the remaining noise signals. As such, these noise signals represent samples of a spatially and temporally uncorrelated soundfield. The corresponding spatospectral correlation matrix is given by $\mathbf{R}_{\mathbf{N}\mathbf{N}}(\omega) = \delta^2 \mathbf{I}$, where \mathbf{I} is the identity matrix and δ^2 is the noise variance (assumed equal for each microphone output).

As Hermitian matrices, the spatospectral correlation matrices have the following properties; their eigenvectors are orthonormal and their eigenvalues are real and non-negative. Therefore, taking $\mathbf{R}_{\mathbf{X}\mathbf{X}}$ as an example, we may write,

$$\mathbf{R}_{\mathbf{X}\mathbf{X}}(\omega) = \sum_i \lambda_i(\omega) q_i(\omega) q_i^H(\omega) \quad (3.68)$$

where $\lambda_i(\omega)$ and $q_i(\omega)$ are the i^{th} eigenvalue and eigenvector of $\mathbf{R}_{\mathbf{X}\mathbf{X}}(\omega)$ respectively.

3.6 Discussion

In this chapter we have outlined the fundamental theory underpinning the study of array-signal-processing. We have introduced the concept of propagating waves as spatiotemporal signals – functions of both space and time. Using sensor arrays, we

sample these signals at discrete points in space, giving rise to phenomena (spatial aliasing, wavenumber smoothing etc.) analogous with those associated with temporal sampling. We explored the implications of these phenomena with respect to array geometry.

In section 3.4 we sought to characterize the behaviour of filter-and-sum systems. In doing so, we introduced the concept of steering. As shall become apparent in later chapters, accurate steering is of significant importance in microphone array processing – to the point where precise steering is typically treated as a necessary assumption in the literature. In this chapter, we have outlined the problems of steering error and finite angular resolution that occur when using time-sampled data. This analysis shall inform our discussion, in chapter 6, of the practical difficulties associated with steering.

We expanded our analysis to incorporate non-ideal sensors and derived an expression for the wavenumber-frequency response of filter-and-sum systems using such sensors. We also demonstrated how this expression may be considered as a function of the array geometry. This important result bears remembering as it means that, irrespective of the weighting strategy employed (these shall be discussed in chapters 4 and 5), system performance will, in large part, be determined by the relative positioning of the sensors. This, in turn, forms part of the motivation (which we outline fully in chapter 6) for the use of ad-hoc, distributed arrays.

Chapter 4

Speech-Enhancement Techniques

4.1 Introduction

In this chapter we review multimicrophone techniques for enhancing the perceptual characteristics of recorded speech. Adaptive and non-adaptive techniques are discussed. Starting with the data-independent delay-and-sum beamformer, we continue with a review of data-dependent approaches including minimum-variance-distortionless-response beamforming, the generalized-sidelobe-canceler and the multichannel-Wiener-filter. The relationships between these are also highlighted. We conclude with a review of dereverberation strategies. Note that, for clarity and simplicity, we omit the ω and \vec{s} arguments throughout this chapter.

4.2 Delay-and-Sum Beamforming

Inserting (3.64) into (3.65) we obtain

$$\begin{aligned} Z &= \mathbf{W}^H (S(\mathbf{H}_{dp} + \mathbf{H}_{mp}) + \mathbf{N}) \\ &= S\mathbf{W}^H \mathbf{H}_{dp} + \mathbf{W}^H (S\mathbf{H}_{mp} + \mathbf{N}) \end{aligned} \quad (4.1)$$

The system output produces an enhanced signal if the “clean” component, $S\mathbf{W}^H \mathbf{H}_{dp}$, corresponding to the direct-path sound, is boosted relative to the unwanted “reverberation-plus-noise”, $\mathbf{W}^H (S\mathbf{H}_{mp} + \mathbf{N})$. The simplest method by which we may attempt to achieve this is by maximizing $S\mathbf{W}^H \mathbf{H}_{dp}$ for a given $|\mathbf{W}|$. Following from the Cauchy-Schwartz inequality, this occurs when

$$\mathbf{W} \propto \mathbf{H}_{dp} \exp\{j\omega\tau\} \quad (4.2)$$

where τ is some time shift. This approach to target-signal enhancement is known as

delay-and-sum beamforming (D&S) and, as mentioned in chapter 3, is *the* classical array processing technique. We define a weightvector, $\mathbf{W}_{D\&S}$, to satisfy (4.2)

$$\begin{aligned}\mathbf{W}_{D\&S} &= \frac{\mathbf{H}_{dp}}{|\mathbf{H}_{dp}|^2} \exp\{j\omega\Delta_0\} \\ &= \mathbf{D} \exp\{j\omega\Delta_0\}\end{aligned}\quad (4.3)$$

$\mathbf{W}_{D\&S}$ is proportional to the steering vector, \mathbf{D} (see (3.59)). Determining an appropriate value for $\mathbf{W}_{D\&S}$ is, therefore, equivalent to estimating \mathbf{D} . In practice this is non-trivial. \mathbf{D} is a function of the source location, \vec{s} , the microphone locations, $\{\vec{m}_m\}$, and the intersensor time-delays, $\{\vec{\tau}_m\}$ (although these may be inferred from \vec{s} and $\{\vec{m}_m\}$). However, the problem may be greatly simplified by invoking the so-called “far-field” assumption, whereby whereby the distances between the microphones in the array are considered to be very small compared to the source-microphone ranges. Following from this

$$\frac{a}{\beta|\vec{s} - \vec{m}_a|} \approx \frac{a}{\beta|\vec{s} - \vec{m}_b|} \approx \frac{1}{\sqrt{M}} \quad (4.4)$$

for all \vec{m}_a and \vec{m}_b . As a result, \mathbf{D} is calculable using time-delays (or time-delay estimates - see section 3.5.1) only.

A D&S beamformer may be characterized by its array pattern, $W(\vec{k})$. Once again, for illustrative purposes we consider a LEA with its elements aligned along the x -axis. Rather than being a function of the wavenumber vector \vec{k} , the array pattern is a function of the wavenumber k_x . Alternatively, remembering that $k_x = \frac{\omega \sin\theta}{c}$, we may, for a given c (the speed of sound), express the array pattern as $W(\omega, \theta)$. Figure (4.1) shows $|W(\omega, \theta)|$ for a 7-element LEA with weighting corresponding to a D&S beamformer. The beamformer is “steered” toward a source at 0° in the far-field of the array (i.e. $\tau_m = 0$ and $\frac{a}{\beta|\vec{s} - \vec{m}_m|} = \frac{1}{\sqrt{M}}$ for all m). Propagating waves falling within the mainlobe receive unity gain while others are attenuated. This attenuation is not frequency invariant. As frequency reduces the mainlobe width increases. As a result the D&S offers poor spatial selectivity at low frequencies. The attenuation of noise and reverberation may more accurately be described as low pass filtering.

As described in section 3.3.3, we may achieve a more narrow mainlobe by widening the array. This can be achieved by adding more elements to the array, or, if we are limited in the number of microphones available, we may choose to increase d . Figure (4.2) shows $|W(\omega, \theta)|$ corresponding to a 7-element LEA where $d = 0.1m$. The consequent widening of the array has produced a narrower mainlobe (and hence greater spatial resolution) but at frequencies where $d > \frac{\lambda}{2}$, we observe “grating lobes” due to spatial aliasing. Nonetheless, we may still choose to increase d if it is felt that

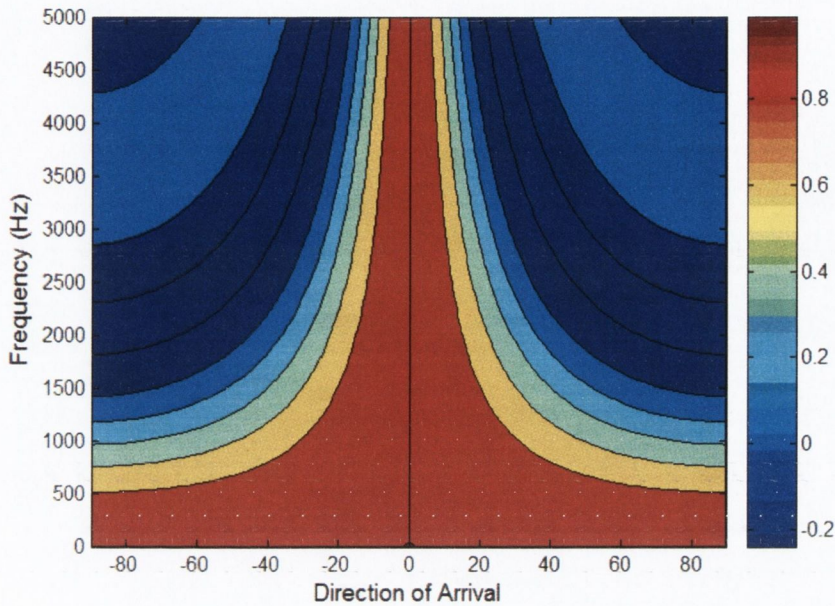


Figure 4.1: $|W(\omega, \theta)|$ for a seven-element D&S beamformer: $W_m(\omega) = \frac{1}{\sqrt{7}} \forall m, \omega$. $d = 0.034m$. The array gain has been normalized to give unity gain in the look direction.

the advantages of enhanced resolution outweigh the disadvantages arising from spatial aliasing.

4.3 Data-Dependant Signal Enhancement

4.3.1 Constrained Optimization

The weightvector applied by the D&S is independent of the noise present in the sampled wavefield and, as such, is suboptimal. In general we may achieve superior noise suppression by explicitly incorporating estimates of the soundfield statistics in the calculation of the weightvector. The resulting class of “data-dependant” beamformers have emerged as the preferred solution to signal enhancement in noisy and reverberant environments.

Virtually all data-dependant signal enhancement techniques have some basis in or bear some relationship to constrained optimization, whereby the weightvector is that which minimizes the power of system output, subject to some constraint(s). The output power of an F&S system may be expressed as $\mathbf{W}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{W}$ and the constrained optimal weightvector, \mathbf{W}_{CO} , may be calculated from

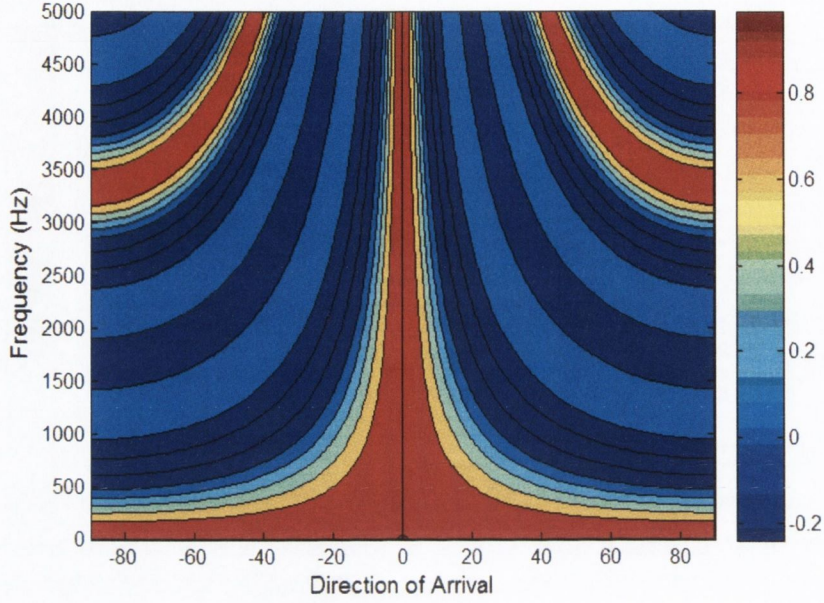


Figure 4.2: $|W(\omega, \theta)|$ for a seven-element D&S beamformer: $W_m(\omega) = \frac{1}{\sqrt{7}} \forall m, \omega$. $d = 0.1m$. The array gain has been normalized to give unity gain in the look direction.

$$\mathbf{W}_{CO} = \arg \min_{\mathbf{W}} \{ \mathbf{W}^H \mathbf{R}_{YY} \mathbf{W} \} \text{ subject to } \mathbf{W}^H \mathbf{C} = \mathbf{c} \quad (4.5)$$

where \mathbf{C} is the constraint vector/matrix and \mathbf{c} is a scalar/vector containing the constraining value(s). We may solve (4.5) using Lagrange multipliers (Appendix B), yielding

$$\mathbf{W}_{CO} = \mathbf{R}_{YY}^{-1} \mathbf{C} (\mathbf{C}^H \mathbf{R}_{YY}^{-1} \mathbf{C})^{-1} \mathbf{c}^H \quad (4.6)$$

The Minimum Variance Distortionless Response Beamformer

Perhaps the simplest implementation of constrained optimization is the Minimum Variance Distortionless Response (MVDR) beamformer, whereby $\mathbf{C} = \mathbf{D}$ and $\mathbf{c} = 1$ - we are effectively constraining the beamformer to apply unity gain to the signal corresponding to some specified steering vector.

$$\mathbf{W}_{MVDR} = \mathbf{R}_{YY}^{-1} \mathbf{D} (\mathbf{D}^H \mathbf{R}_{YY}^{-1} \mathbf{D})^{-1} \quad (4.7)$$

Substituting \mathbf{W}_{MVDR} into (4.1) yields

$$Z = S \exp\{-j\omega\Delta_0\} + \mathbf{W}_{MVDR}^H (S\mathbf{H}_{mp} + \mathbf{N}) \quad (4.8)$$

Thus the system output contains the undistorted (although delayed) target signal plus a minimized interference component. Figure (4.3) plots the array pattern corresponding to W_{MVDR} for a five-element LEA where $\mathbf{D} = [\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}]$ (i.e. we are steering towards a far-field source with a DOA of 0°) and the noise field consists of two broadband propagating noise signals with DOAs of 45° and -13° . As expected, the MVDR beamformer gives unity gain in the look direction while at the same time placing deep nulls in the path of the noise. Alternatively, we might consider setting the constraints to cancel the noise, letting $C = [D, D_{I_1}, D_{I_2}]$ (where D_{I_1} and D_{I_2} are the steering vectors corresponding to the interference sources) and $c = [1, 0, 0]$. This, however, should not be done for two reasons; firstly because such an approach would require us to take the additional step of explicitly estimating the directions from which the noise signals propagate and secondly because it would, in fact, yield inferior noise suppression. To understand why this is the case, consider figure (4.3) once again. The large sidelobes at angles from which no noise is assumed to propagate are an unavoidable consequence of placing deep nulls in the path of the propagating noise. In addition to the propagating noise, quantization and rounding error appears as noise in the microphone outputs. This noise may be considered as propagating from all directions and, as such, is magnified by the large sidelobes. This is known as “white noise gain”. We cannot suppress the propagating noise further without increasing the white noise gain and noise levels overall. Therefore, the very strong attenuation (as opposed to complete removal) of the propagating interference, as shown in figure (4.3), achieves minimal overall noise levels and, as such, is the preferred approach.

Implementing the MVDR

In implementing the MVDR, we must consider how we might effectively estimate $\mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1}$. The first approach would be to model it from assumptions made about the soundfield. By doing this, $\mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1}$ may be calculated offline. For example, we may assume that we have a single source in the presence of spectrally white noise that is uncorrelated across each of the microphone locations – i.e. the noise is spatially white also. From this $\mathbf{R}_{\mathbf{N}\mathbf{N}} = \delta\mathbf{I}$, where δ is some real scalar. If we also assume that the environment is anechoic, we need only consider $\mathbf{R}_{\mathbf{N}\mathbf{N}}$, as only noise - and not reverberation - need be suppressed. Substituting $\delta\mathbf{I}$ for $\mathbf{R}_{\mathbf{Y}\mathbf{Y}}$ in (4.7) yields a result that is, in fact, a D&S beamformer.

$$\mathbf{W}_{MVDR} = \frac{\mathbf{D}}{\delta^2 |\mathbf{D}|^2} \propto \mathbf{W}_{D\&S} \quad (4.9)$$

Alternatively we may consider our observations to contain diffuse or “spherically isotropic” noise. Diffuse noise is often held to approximate ambient noise and re-

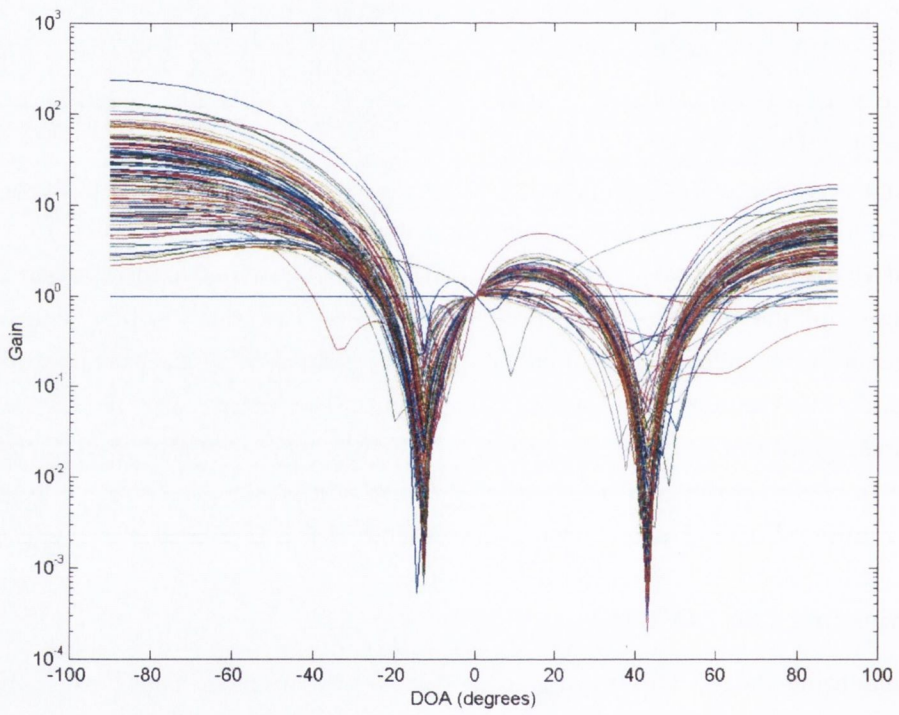


Figure 4.3: The Array Pattern for a MVDR beamformer.

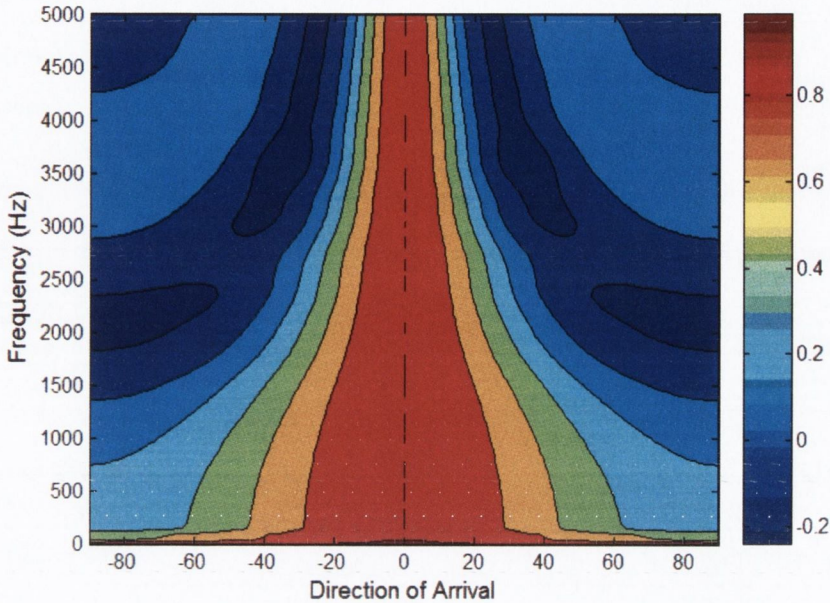


Figure 4.4: $|W(\omega, \theta)|$ for a seven-element superdirective beamformer: $d = 0.034m$. The array gain has been normalized to give unity gain in the look direction.

reverberation in small-to-medium sized rooms, [18],[19]. The corresponding correlation matrix, \mathbf{R}_{Diff} , may be defined in terms of its individual components,

$$R_{i,j_{Diff}}(\omega) = P(\omega) \text{sinc}\left(\frac{\omega |\vec{m}_i - \vec{m}_j|}{c}\right), \quad (4.10)$$

where i and j are the row and column indices respectively and where $P(\omega) = R_{i,i_{Diff}}(\omega) = R_{j,j_{Diff}}(\omega)$.

Figure (4.4) shows the array pattern obtained, for a 7-element linear, equispaced array with intersensor spacing of $0.034m$, by letting $\mathbf{D} = [\frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}]$ and replacing \mathbf{R}_{YY} in (4.7) with \mathbf{R}_{Diff} .

Compared with that achieved by D&S weighting, we obtain superior resolution at low frequencies. It is for this reason that such beamformers are commonly referred to as exhibiting “super-directivity”. Despite this, superdirective beamformers are very often unsuitable for practical implementation. The narrow mainlobe at low frequencies is achieved at the expense of massive sidelobes in the invisible region of the wavenumber domain (see section 3.3.4). No propagating waves inhabit this region but sensor/quantization noise does. To avoid applying massive gain to this noise we must take account of its presence. We do this by applying diagonal loading to \mathbf{R}_{diff} , although it is difficult to know to what degree this should be done - too little and we

suffer white noise gain, too much and we obtain performance very similar to that of a D&S beamformer.

A more satisfactory approach is to estimate $\mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1}$ directly from the observed data. We may do this in a blockwise manner as follows,

$$\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}(l) = \alpha \tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}(l-1) + \mathbf{Y}(l-1)\mathbf{Y}^H(l-1) \quad (4.11)$$

applying the matrix inversion lemma

$$\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}^{-1}(l) = \frac{1}{\alpha} \left[\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}^{-1}(l-1) - \frac{\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}^{-1}(l-1)\mathbf{Y}(l-1)\mathbf{Y}^H(l-1)\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}^{-1}(l-1)}{\alpha + \mathbf{Y}^H(l-1)\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}^{-1}(l-1)\mathbf{Y}(l-1)} \right] \quad (4.12)$$

where $\alpha \in \mathbb{R}\{0 : 1\}$, is a leakage coefficient and l is the observation-block index. As $l \rightarrow \infty$, $\tilde{\mathbf{R}}_{\mathbf{Y}\mathbf{Y}}^{-1}(l) \rightarrow \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1}$ to within a constant of proportionality. We may obtain $\tilde{\mathbf{R}}_{\mathbf{N}\mathbf{N}}^{-1}$ in a similar way using noise-only observations. We may then use the updated correlation matrices to recalculate \mathbf{W}_{MVDR} as required. Alternatively we may update \mathbf{W}_{MVDR} directly using a constrained LMS approach as derived in [20], shown below.

$$\begin{aligned} \mathbf{W}(l+1) &= \mathbf{C}(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{c}^H \\ &+ (\mathbf{I} - \mathbf{C}(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{C}^H) (\mathbf{W}(l) + \mu\mathbf{Z}^*(l)\mathbf{Y}(l-1)) \end{aligned} \quad (4.13)$$

4.3.2 Generalized Sidelobe Canceller

First developed by Griffiths and Jim, [21],[22], the generalized sidelobe canceller (GSC), is a data-dependant beamforming technique that was originally conceived as an alternative and more efficient implementation of the MVDR. The weightvector is decomposed into orthogonal components; a non-adaptive component, \mathbf{W}_c , satisfies the constraints while an adaptive component, \mathbf{W}_a , seeks to minimize the output power.

Figure (4.5) shows a block diagram of the GSC. \mathbf{W}_c defines a conventional (data-independent) beamformer that captures the target-signal-plus-noise. In parallel, the signal is processed by the blocking matrix \mathbf{B} , the purpose of which is to cancel the signal components creating a noise reference which the multiple-input canceller, MC (defined by \mathbf{W}_a) then optimally subtracts from the output of the conventional beamformer. The resulting system output may be expressed as

$$Z_{GSC} = \mathbf{W}_c^H \mathbf{Y} - \mathbf{W}_a^H \mathbf{B} \mathbf{Y} \quad (4.14)$$

In optimally subtracting the noise reference, the adaptive component is seeking to minimize the output power

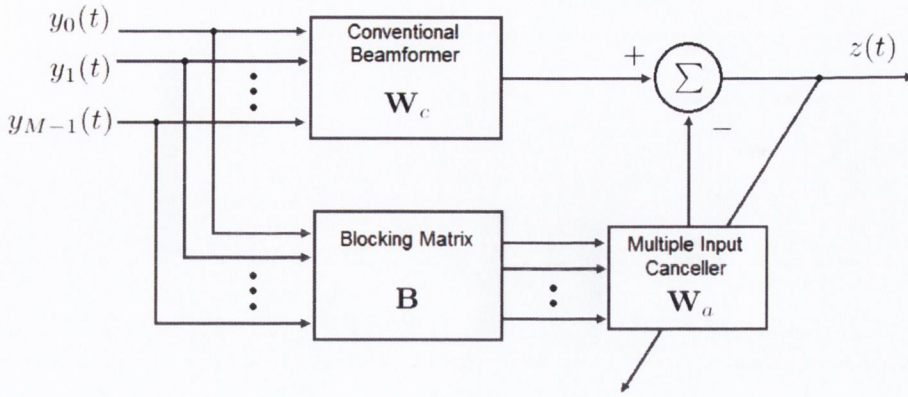


Figure 4.5: The Generalized Sidelobe Canceller.

$$\mathbf{W}_a = \arg \min_{\mathbf{W}_a} \{ (\mathbf{W}_c - \mathbf{B}^H \mathbf{W}_a)^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} (\mathbf{W}_c - \mathbf{B}^H \mathbf{W}_a) \} \quad (4.15)$$

This is an unconstrained optimization, the optimum Wiener solution to which is given by

$$\mathbf{W}_a = (\mathbf{B}\mathbf{R}_{\mathbf{Y}\mathbf{Y}}\mathbf{B}^H)^{-1} \mathbf{B}\mathbf{R}_{\mathbf{Y}\mathbf{Y}}\mathbf{W}_c \quad (4.16)$$

This yields

$$\mathbf{W}_{GSC} = \mathbf{W}_c - \mathbf{B}^H \mathbf{W}_a = [\mathbf{I} - \mathbf{B}^H (\mathbf{B}\mathbf{R}_{\mathbf{Y}\mathbf{Y}}\mathbf{B}^H)^{-1} \mathbf{B}\mathbf{R}_{\mathbf{Y}\mathbf{Y}}] \mathbf{W}_c \quad (4.17)$$

We now determine the conditions under which \mathbf{W}_{GSC} performs a constrained optimization operation. Equating (4.17) and (4.6) and multiplying both sides by $\mathbf{B}\mathbf{R}_{\mathbf{Y}\mathbf{Y}}$ reveals

$$\mathbf{B}\mathbf{C}(\mathbf{C}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{C})^{-1} \mathbf{c}^H = 0 \quad (4.18)$$

From this we see that, if the GSC is to perform constrained optimization the null-space of \mathbf{B} must contain $\mathbf{C}(\mathbf{C}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{C})^{-1} \mathbf{c}^H$. Furthermore, \mathbf{B} must be singular. If it is invertible, $\mathbf{C}(\mathbf{C}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{C})^{-1} \mathbf{c}^H = 0$ which makes no sense. Therefore, we define \mathbf{B} as a singular matrix where

$$\mathbf{B}\mathbf{C} = 0 \quad (4.19)$$

It is, however, not necessary to specify \mathbf{C} or \mathbf{c} explicitly. Remembering that \mathbf{W}_c satisfies the constraints (i.e. $\mathbf{W}_c^H \mathbf{C} = \mathbf{c}$), we may write

$$\mathbf{W}_c^H = \mathbf{c}(\mathbf{C}^H \mathbf{C})^{-1} \mathbf{C}^H \quad (4.20)$$

Thus (4.19) is equivalent to

$$\mathbf{B} \mathbf{W}_c = 0 \quad (4.21)$$

Therefore, if each row of the blocking matrix is orthogonal to the non-adaptive weight vector, the GSC will perform a constrained optimization operation.

Implementing the GSC

In the following discussion we shall describe the classical implementation of the Griffiths-Jim GSC (GJ-GSC) which considers the case of a single target source and a small array. The dimensions of the array are sufficiently small and the source sufficiently distant that the source-microphone ranges and the orientation of each microphone relative to the source may be assumed to be identical. Following from this assumption (which, we remember, is known as the “far-field” assumption) we may ignore the effects of propagation losses and source-directionality and consider the amplitude of the direct-path components of the received sound to be equal at each microphone. For simplicity, let us assume that the array has been steered such that the target-signal-components of \mathbf{Y} are time-aligned.

To satisfy a constraint of unity gain in the look direction (0°) we let $\mathbf{W}_c = \frac{1}{M}[1, 1, \dots, 1]_{1 \times M}$. The blocking matrix is defined as in [21]

$$\mathbf{B} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & 1 & -1 \end{bmatrix}_{M-1 \times M} \quad (4.22)$$

As required, \mathbf{B} is singular because it is not square. Generally speaking, \mathbf{B} may be defined in any way provided that the coefficients of each row sum to zero. In effect, each row is a null-steering beamformer. The nulls are steered toward the target source thereby removing the target signal from the noise reference $\mathbf{B}\mathbf{Y}$.

The MC is implemented using a filterbank of parallel FIR filters. The filter-tap coefficients are determined using an LMS update algorithm whereby

$$\mathbf{W}_a(l+1) = \mathbf{W}_a(l) + \mu \mathbf{B}\mathbf{Y}(l-1) [Z_{GSC}(l) - \mathbf{W}_a^H(l) \mathbf{B}\mathbf{Y}(l-1)]^* \quad (4.23)$$

where l is the update index and the “stepsize” μ is a real scalar in the range $0 : 1$.

Robust GSC

Data-Dependant beamformers following a constrained optimization paradigm suffer a noted susceptibility to target signal cancellation (TSC), [23],[24]. In the case of the GSC, this occurs when target signal information is passed by the blocking matrix to the MC (this is known as “target signal leakage”). The MC then uses this information to remove the target signal from the GSC output.

Target signal leakage can be a result of steering errors, sensor-gain errors or reverberation, when multipath replicas of the target signal propagate with DOAs other than the look direction and are, therefore, not cancelled by \mathbf{B} . The vast majority of research into the GSC concerns the design of GSC variants that are robust against these problems. Several researchers ([25] to [33]) have obtained robust performance using a variety (and usually a combination) of techniques.

Robust MC Robust MCs are implemented using either leaky, [25],[26],[27], or norm-constrained, [28],[29],[31], LMS filters. The effect of such constraints may be interpreted as follows. Target signals falling close to but just outside the nulls of the blocking matrix are not fully cancelled but are, nonetheless, greatly attenuated. Similarly, multipath reflections are subject to increased attenuation due to propagation and surface absorption. To remove target signal from Z_{GSC} we must first apply significant gain to the attenuated target-signal components in the noise reference. By constraining \mathbf{W}_a , we prevent this from occurring.

Unfortunately, as \mathbf{W}_{GSC} is data-dependant, the GSC response is determined more by signal and interference statistics than it is by the controllable parameters of robust beamformers such as leakage coefficients and norm constraints. For this reason, no particular parameter set can be seen to consistently and predictably correspond to a particular level of robustness.

Robust B Blocking Matrices may be altered to improve robustness. Typically, this is achieved by specifying additional constraints to broaden the null that is steered by the blocking matrix. One example is the derivative constraint [21], whereby the derivative of the blocking matrix’s array pattern is constrained to be zero in the look-direction. Assuming, once again, that the look direction is 0° , this is realized using a blocking matrix with rows of $[1, -2, 1, 0, \dots, 0]$.

Chen et al, [32], propose a simple and sturdy design incorporating additional amplitude constraints (extra nulls about the look direction) in \mathbf{B} . This is achieved by means of a multistage blocking matrix. In [23], a methodology is outlined whereby, given prior knowledge of the approximate noise and reverberation statistics, we may

design a blocking matrix to achieve a specified balance between noise suppression and TSC.

Nordholm and Claesson, [25], suggest a generalized approach. A temporal high-pass filter is designed according to the requirements of its null-width (stopband) and restrictions on the number of coefficients that may be used. The blocking matrix then applies weights to the sensor outputs that correspond to the coefficients of the temporal filter. This approach is expanded for wideband applications by designing appropriate highpass filters for each frequency bin and applying these frequency-dependent weights via FIR filters at the microphone outputs. In this way we attain an effective null-width that is uniform across all frequencies of interest, [26]. A similar effect is achieved by a slightly different approach in [33]

As an alternative to null-steering, Hoshuyama et al ([27] to [30]) implement a blocking matrix using coefficient-constrained LMS filters (CCLMS). Under this approach, the filter-tap coefficients of each CCLMS adapt such that, following the filter-and-subtract operation illustrated in figure (4.6), the power of each output of the blocking matrix is minimized. In this way, the correlation between $\mathbf{W}_c^H \mathbf{Y}$ and the noise reference is also minimized. The coefficient constraints are chosen such that only signals with a DOA within a specified range (which would presumably include the target signal) may be removed from the blocking matrix output.

This method may be shown to be theoretically superior to null-steering blocking matrices, [34]. This is because blocking occurs only at those frequencies which compose the target signal. At other frequencies no blocking occurs, thereby removing the constraints on the permissible noise reduction. However, other researchers have reported this approach to be unsuitable in the presence of non-stationary signals or noise, [32].

Although robust GSC techniques are demonstrably successful in preventing target signal leakage due to steering errors, they are ineffective in the presence of reverberation when multipath replicas of the target signal propagate across the array with DOAs well outside even a broadened null. Furthermore, by constraining the MC and widening the null steered by \mathbf{B} , we necessarily limit noise cancellation, thereby diminishing the advantages of the GSC over data-independent approaches. As a result, practical implementations of the GSC limit adaptation to noise-only periods with consequent reductions in the ability of the system to respond to changes in the noise field, [30].

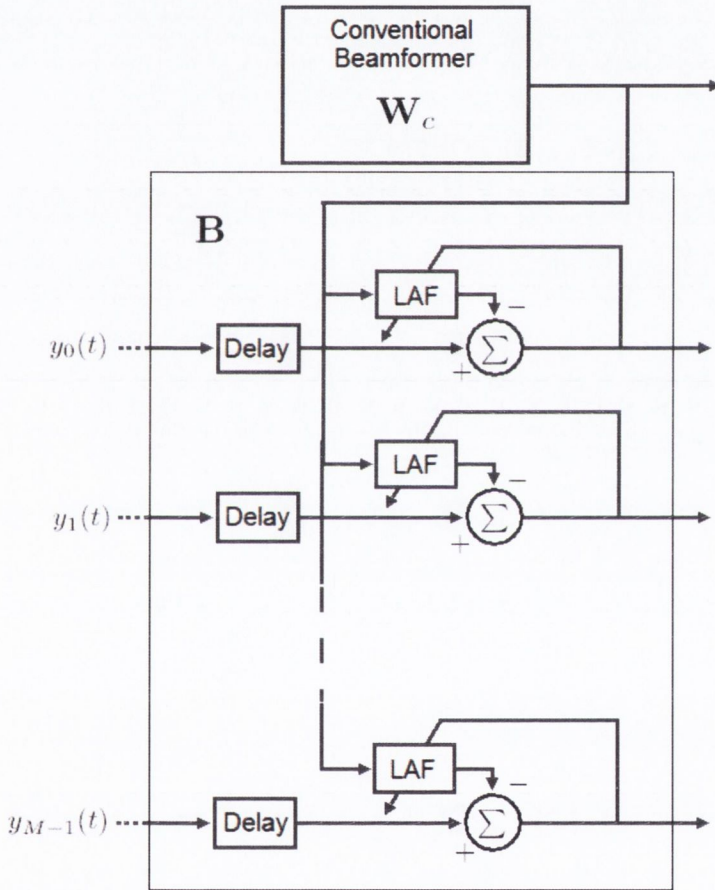


Figure 4.6: Leaky adaptive filters (LAFs) in the blocking matrix of a GSC.

4.3.3 The Multichannel Wiener Filter

As an alternative to spatial filtering we may attempt to suppress noise by processing multiple channels containing noise corrupted signals in such a way as to optimally approximate an estimate of the clean signal. The multichannel Wiener filter (MWF) returns a MMSE approximation of the clean signal (see appendix A). Note that in this case, “clean” is taken to mean “noise-free” as distinct from “noise-and-reverberation-free”. This is a departure from previous speech enhancement strategies where we wish to preserve only the direct-path component of the recorded sound.

$$\mathbf{W}_{MWF} = (\mathbf{R}_{\mathbf{X}\mathbf{X}} + \mathbf{R}_{\mathbf{N}\mathbf{N}})^{-1} \mathbf{R}_{\mathbf{X}\mathbf{X}} \mathbf{v}_0 \quad (4.24)$$

where $\mathbf{v}_0 = [1, 0, \dots, 0]_{1 \times M}^T$

$\mathbf{R}_{\mathbf{N}\mathbf{N}}$ may be estimated during noise-only periods and, assuming that the noise is stationary, we may estimate $\mathbf{R}_{\mathbf{X}\mathbf{X}}$ via $\mathbf{R}_{\mathbf{X}\mathbf{X}} = \mathbf{R}_{\mathbf{Y}\mathbf{Y}} - \mathbf{R}_{\mathbf{N}\mathbf{N}}$, [35],[36],[37]. \mathbf{W}_{MWF} may then be calculated either directly, or by means of a computationally efficient generalized singular-value decomposition (GSVD) approach, [35]. Further reductions in computational complexity are obtained using a subband implementation of the GSVD method, [36]. Alternative approaches employ prerecorded calibration signals which are then used to converge LMS filters, [38],[39],[40].

Decomposing \mathbf{W}_{MWF} (see appendix A) is informative and reveals that the multichannel Wiener filter may be described as the product of a scalar component, W_{p2} , and a vector component, \mathbf{W}_{p1} , which is in fact, equivalent to the solution to a constrained optimization problem (compare the form of \mathbf{W}_{p1} with equation (4.6))

$$\mathbf{W}_{MWF} = \underbrace{\left[(\mathbf{H}^H \Gamma_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{H})^{-1} (\Gamma_{\mathbf{N}\mathbf{N}}^{-1} \mathbf{H}) \mathbf{H}_0^* \right]}_{\mathbf{W}_{p1}} \underbrace{\left[\frac{|S|^2 \mathbf{H}^H \mathbf{H}}{|S|^2 \mathbf{H}^H \mathbf{H} + P_N} \right]}_{W_{p2}} \quad (4.25)$$

Consider W_{p2} ; as the SNR at a given frequency reduces the effect of W_{p2} is to suppress the signal components at that frequency, thus distorting the signal. Since the signal and noise can be assumed to overlap, at least partially, in the frequency domains, noise suppression occurs at the expense of target signal distortion, [41]. The speech-distortion-regularized MWF (SDR-MWF, \mathbf{W}_{SDW}) allows us to strike a balance between the twin imperatives of noise attenuation and speech signal fidelity, [37].

$$\mathbf{W}_{SDW} = (\mathbf{R}_{\mathbf{X}\mathbf{X}} + \kappa \mathbf{R}_{\mathbf{N}\mathbf{N}})^{-1} \mathbf{R}_{\mathbf{X}\mathbf{X}} \mathbf{v}_0 \quad (4.26)$$

where $\kappa \in \mathbb{R}\{0 : \infty\}$. As $\kappa \rightarrow \infty$, all emphasis is placed on noise suppression and $\mathbf{W}_{SDW} \rightarrow 0$. Conversely, if $\kappa = 0$, all emphasis is placed on minimizing the signal

distortion and $\mathbf{W}_{SDW} = \mathbf{v}_0$. Note that, as in [38], a similar effect may be achieved by varying the gain applied to the signal and noise calibration signals.

4.3.4 Blind Source Separation

Blind source separation (BSS) refers to the problem of separating mixtures of statistically independent signals using arrays of unknown geometry. Such techniques would have obvious application to, for example, enhancing recordings made in the presence of a local noise source. While a thorough discussion of BSS is beyond the scope of this thesis, a brief overview of the topic is presented here. For a more detailed review of BSS techniques, readers are referred to [42] and [43].

Let us consider the two-source, two microphone problem. In such a scenario the observation vector may be expressed as

$$\begin{aligned} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} &= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \\ \mathbf{X} &= \mathbf{H}_{12}\mathbf{S} \end{aligned} \quad (4.27)$$

To separate the signals we must calculate the matrix \mathbf{W}_{12} such that

$$\mathbf{S} = \mathbf{W}_{12}\mathbf{X} \quad (4.28)$$

or equivalently

$$\mathbf{W}_{12} = \frac{1}{H_{11}H_{22} - H_{12}H_{21}} \begin{bmatrix} H_{22} & -H_{12} \\ -H_{21} & H_{11} \end{bmatrix} \quad (4.29)$$

It is apparent that BSS is achieved by weighting and combining the outputs of multiple microphones. As such, the problem is strongly related to beamforming. Indeed, in [44] it is shown that the performance of a MVDR beamformer represents an upper bound on that of frequency domain BSS techniques.

A wide variety of techniques exist for calculating \mathbf{W}_{12} (an overview of these is provided in [45]). All of these may be seen to suffer from one or more typical drawbacks. From (4.29), it may be seen that \mathbf{W}_{12} is a function of the source-microphone frequency responses. In reverberant environments, the corresponding impulse responses are of the order of hundreds of milliseconds long. BSS techniques implemented in the time domain require similarly long filters which can prove prohibitive in terms of computational complexity [45].

BSS techniques implemented in the frequency domain offer reduced computational complexity but suffer from a what is commonly referred to as the “permutation problem”, whereby the signals in each frequency bin, while separated, are not assigned to

a known and unique source. Special methods are required to determine which of the signals across each frequency bin must be recombined to obtain a broadband signal corresponding to a single unique source. These are reviewed in [46], which seeks to present an implementation of a frequency-domain BSS technique that comprehensively addresses the permutation and other problems while offering robust performance in reverberant environments. While the source separation achieved is described by the authors as being “fairly good”, the acoustic environment in which the recordings were obtained were only mildly reverberant (the microphones were placed approximately 1.2m from a sound source in a room with RT60 of 130ms). Other studies by the same authors show BSS performance to be “highly limited” in more reverberant environments such as we are likely to encounter in a classroom [47].

4.4 Dereverberation

Like noise, reverberation reduces the intelligibility of recorded speech (see section 2.5.1). Its exact effect on the perceived quality of the speech is more difficult to characterize given the subjective nature of the term. Much research has therefore focused upon techniques by which we may “dereverberate” a signal, albeit with limited success.

Several techniques have been proposed that exploit knowledge of the source-microphone impulse responses. Gillespie and Atlas, [51], present a method by which we may determine the appropriate FIR filter tap coefficients such that the overall room-plus-filters impulse response optimally approximates an all-pass filter. Multichannel Matched-Filtering, [52],[53], convolves recordings with time reversed versions of the source-microphone impulse responses before adding the resulting filter outputs. This simultaneously time aligns and maximizes the gain applied to the direct-path components in the microphone outputs. In practice, we are unlikely to have access to impulse responses and so techniques that do not require this explicit knowledge are of greater practical interest.

Beamforming may be applied suppress reverberation. Unfortunately, beamforming does not achieve satisfactory dereverberation in situations where a significant proportion of the reverberation propagates from directions of arrival close to that of the direct component, [54]. Also, for practical talker-microphone distances, the far-field requirement restricts the array width, which, in turn, severely limits beamforming performance at low frequencies. Using wider arrays and performing near-field beamforming has been shown to lead to enhanced dereverberation, [55], but does so at the expense of increased steering-vector-estimation complexity.

Cepstral techniques may be applied to the dereverberation problem, [49]. The

cepstrum, $C_X(\omega)$, may be characterized as follows, [50].

$$C_X(\omega) = \log \{|X(\omega)|\} \quad (4.30)$$

Consider a transmission path whose impulse response consists of a direct path component and a single attenuated echo, delayed by τ_e

$$\begin{aligned} X(\omega) &= S(\omega)(1 + \alpha \exp \{-j\omega\tau_e\}) \\ C_X(\omega) &= C_S(\omega) + \log \{1 + \alpha \exp \{-j\omega\tau_e\}\} \end{aligned} \quad (4.31)$$

where $C_S(\omega) = \log\{S(\omega)\}$. The cepstra of the received sound is equal to the superposition of $C_S(\omega)$ and a periodic component with period $\frac{2\pi}{\tau_e}$, which may be removed by comb filtering. The clean speech signal may then be recovered.

Unfortunately, while effective in the presence of a single echo, cepstral techniques have greater difficulty in the presence of multiple echoes, as are typical in reverberant room environments. Several authors have sought to combine beamforming with cepstral techniques, [56],[57], whereby lowpass filtering is applied in the cepstral domain. Problems still remain however if the signal and reverberation cepstra overlap in the cepstral domain.

Recently, much research has focused on techniques based on processing of the “residual” obtained using a linear predictive coding (LPC) approach, [58],[59],[60]. Two assumptions underpin such approaches; firstly, that the LP coefficients of clean and reverberant speech are identical and secondly, that the LPC residual consists of periodic glottal pulses followed by additional peaks due to reverberation. In [61] it is shown that the first assumption is true only in a spatially averaged sense (that is to say that the mean of the LP coefficients corresponding to reverberant speech recorded at multiple, spatially distributed locations will tend towards those of clean speech). Also, for high-pitched speech, such as from women or children, the second assumption does not hold and so these techniques have limited application.

4.5 Discussion

In this chapter we have reviewed multimicrophone techniques for enhancing the perceptual characteristics of recorded speech. These may be described as being either data-dependent or data-independent techniques. The former of these two broad categories will, in theory, offer noise suppression that is, in some sense, optimal and generally superior to that achieved using data-independent methods.

For data-dependent, constrained optimization approaches (of which the GSC is the most versatile and efficient implementation) the “optimal” output is obtained as that

which minimizes output power subject to some constraints (a typical example being unity gain for signals originating at some target location). The GSC is highly effective in low-reverberation environments and will adapt quickly to changes in noise statistics. However, the GSC is subject to target signal cancellation in highly reverberant environments or where mis-steering has occurred.

For MCWF approaches the optimal output is that which is closest, in a MMSE sense, to a (noise free) reference signal. The MCWF requires no steering and does not fail in the presence of strong reverberation. It does, however, require any noise present to be stationary such that a valid noise reference may be obtained by means of infrequent updates. Non-stationary noise results in reduced performance.

The varying modes of failure of the MCWF and GSC make it difficult to make a definitive and quantitative comparison of their relative performances as this will depend heavily upon the specific acoustic conditions encountered. Indeed, it is conceivable, that in the presence of strong reverberation and highly non-stationary noise, a data-independent beamformer such as the D&S would provide superior performance in practice. Nonetheless, these methods, if employed appropriately, can achieve a reduction in noise power and a consequent improvement in acoustic signal quality.

Chapter 5

Time-Delay Estimation and Source-Localization

5.1 Introduction

In the following section we shall review the literature concerning the problems of time-delay estimation and source localization. The two problems are distinct but strongly related, to the point where the terms are used interchangeably in the literature. The confusion arises principally because time-delay estimation is a necessary first step for many source-localization algorithms but is further compounded by inconsistencies in the way in which the source localization problem is defined.

In the strictest sense, the term “source localization” refers to the problem of finding a unique and discrete estimate of the location of the source with respect to some coordinate system. Throughout the literature, however, varying and usually less strict definitions are used. Following from these, source localization may also be taken to refer to estimating the location of source to within some surface of revolution or bearing line. Alternatively, we may simply wish to determine which of a group of candidate locations is closest to the source.

Time-delay estimation is, on the other hand, concerned with estimating the inter-sensor time-delays. For any two microphones, m_a and m_b , the time-delay, $\tau_{a,b}$, is the difference in between the source-microphone propagation delays (see section 3.5.1). For M microphones we may define a set of time-delays, $\{\tau_m\}$.

$$\{\tau_m\} = \{0, \tau_1, \dots, \tau_{M-1}\} \quad (5.1)$$

For a given array geometry and known speed of sound, $\{\tau_m\}$ is wholly dependent upon the source location \vec{s} . Therefore, each source location *estimate* (SLE, $\vec{\tilde{s}}$), will have a corresponding and implicit set of time-delay estimates (TDEs, $\{\tilde{\tau}_m\}$). However, these

may not be unique to \tilde{s} and so not every $\{\tilde{\tau}_m\}$ will have a corresponding, discrete \tilde{s} . Furthermore, rather than being calculated geometrically, as in (3.57), TDEs are found by a variety of means that are independent of the source or microphone locations.

We shall continue with a review of time-delay estimation techniques and follow this with a review of the literature concerning source localization using TDE, parametric and subspace-based methods.

5.2 Time-Delay Estimation Techniques

5.2.1 Cross-Correlation

The cross correlation (CC) method is one of the earliest and simplest techniques for time-delay estimation. The CC method is based upon the ideal anechoic signal model (section 3.5.1) and assumes that the noise present in each channel is uncorrelated with the signal and noise in other channels. For two microphones

$$\begin{aligned} y_a(t) &= x_a(t) + n_a(t) \\ y_b(t) &= x_b(t) + n_b(t) \\ &= \alpha x_a(t - \tau_{a,b}) + n_b(t) \end{aligned} \quad (5.2)$$

The intersensor time delay is taken as that time lag which maximizes the cross-correlation function between the microphone outputs.

$$\tilde{\tau}_{a,b} = \arg \max_{\tau} \{\psi_{a,b}(\tau)\} \quad (5.3)$$

where $\psi_{a,b}(\tau) = E\{y_a(t)y_b(t + \tau)\}$.

5.2.2 Generalized Cross-Correlation

Generalized cross-correlation techniques (GCC) represents an extension and improvement of the CC method, [62]. Following from the Wiener-Kinchin theorem, the Fourier transform of $\psi_{a,b}(\tau)$ is the cross-spectrum, $\Psi_{a,b}(\omega)$,

$$\Psi_{a,b}(\omega) = E\{Y_a(\omega)Y_b^*(\omega)\} \quad (5.4)$$

Applying a frequency variant weighting, $\Phi(\omega)$.

$$\Psi_{GCC}(\omega) = E\{\Phi(\omega)Y_a(\omega)Y_b^*(\omega)\} \quad (5.5)$$

In practice, $\Psi_{GCC}(\omega)$ is estimated using the DFTs of blocks of time-sampled microphone output data. Furthermore, the instantaneous rather than expected value is used. This yields

$$\begin{aligned}\widehat{\Psi}_{GCC}(v) &= \Phi(v)Y_a(v)Y_b^*(v) \\ \widehat{\psi}_{GCC}(n) &= IDFT\{\widehat{\Psi}_{GCC}(v)\}\end{aligned}\quad (5.6)$$

where v is the frequency-bin index, n is the sample index and $Y_a(v)$ is the DFT of $y_a(k)$. Following from the CC method

$$\frac{\tilde{\tau}_{GCC}}{T} = \arg \max_n \left\{ \widehat{\psi}_{GCC}(n) \right\} \quad (5.7)$$

where T is the temporal sampling period.

The weighting function $\Phi(v)$ may be chosen according to the acoustic conditions that obtain. In [62] a maximum-likelihood GCC (GCC-ML) is derived for use in anechoic environments where the noise spectra are known *a priori*.

$$\Phi_{ML}(v) = \frac{|Y_a(v)||Y_b(v)|}{|Y_a(v)|^2 |N_b(v)|^2 + |Y_b(v)|^2 |N_a(v)|^2} \quad (5.8)$$

The performance of the GCC-ML is optimal in the sense that, when the underpinning assumptions hold, the estimate variance approaches the lower Cramer-Rao bound. However, when our assumptions regarding the signal model do not hold - in particular when reverberation is present - the GCC-ML is suboptimal and we observe a potentially significant performance degradation, [63],[64].

The phase transform GCC (GCC-PHAT) is an alternative, suboptimal weighting that has been observed to give improved TDEs in reverberant environments, [62],[65].

$$\Phi_{PHAT}(v) = \frac{1}{|Y_a(v)Y_b^*(v)|} \quad (5.9)$$

The GCC-PHAT weighting function flattens the cross-spectrum thus retaining only the phase information. Phase information that is correlated across all frequency bands (i.e. the phase shift due to the direct-path delay) is thereby emphasized. However, an additional effect of $\Phi_{PHAT}(v)$ is to place equal emphasis on all frequency bands, including those with low SNRs. The GCC-PHAT will, therefore, offer inferior performance in the presence of noise, [62].

Other weighting functions seek to enhance the TDEs by exploiting the spectral characteristics of speech. In [66], $\Phi(v)$ is given by

$$\Phi(v) = 20 \log_{10}(\overline{S}(v)) \quad (5.10)$$

where $\bar{S}(v)$ a smoothed, averaged speech spectrum, thus emphasizing the contribution of those frequency components where the speech-energy (and hence SNR) is high. A pitch-based delay estimator is proposed in [67], whereby $\Phi(v)$ is selected according to the degree to which the observed signal spectrum corresponds to a harmonic speech model. Frequency bands corrupted by noise/reverberation will deviate from the explicit speech model and will, therefore, be de-emphasized.

Typically, GCC-based estimators use short segments of data ($\sim 10 - 30ms$). In practice, the frequency with which we must update the TDEs is often very much lower. In many applications, therefore, we will have access to many individual TDEs with which to make a single, time-averaged, estimate. Beyond simple averaging, we may construct a histogram of the TDEs and select the $\tilde{\tau}_{GCC}$ as that corresponding to the maximum. Alternatively we may assign a weight to each TDE based upon a set of “reliability criteria” including the observed power and the ratio of local maxima of $\hat{\psi}_{GCC}(n)$, [68],[69],[70].

5.2.3 Least-Mean-Squares Methods

In addition to its many applications to microphone array noise suppression, the least-mean-squares (LMS) algorithm may also be applied to time-delay estimation, [71],[72]. Like the CC/GCC methods, LMS-based approaches assume an ideal signal model. From (5.2)

$$y_b(t) = x_a(t) * \delta(t - \tau_{a,b}) + n_b(t) \quad (5.11)$$

Remembering that $n_a(t)$ and $n_b(t)$ are assumed uncorrelated with the signal and each other, we may determine the delay by means of a channel-identification approach, whereby we find the filter that processes $y_a(t)$ to optimally approximate $y_b(t)$ in a MMSE sense. The response of this filter is taken to equal $\delta(t - \tilde{\tau}_{a,b})$. In practice, we will be working with discrete-time-sampled data and using a L -tap FIR filter, \mathbf{f} , to model the channel. This is illustrated in figure (5.1). As the system is necessarily bandlimited the coefficients of \mathbf{f} will be samples of $\text{sinc}(\frac{\pi t - \tilde{\tau}_{a,b}}{T})$.

$$\mathbf{f}(n+1) = \mathbf{f}(n) + \mu \mathbf{y}_a(n) [y_a(n - \zeta) - \mathbf{f}^T(n) \mathbf{y}_a(n)] \quad (5.12)$$

where $0 < \mu < 1$, $\mathbf{y}_a(n) = [y_a(n), \dots, y_a(n - \zeta + 1), \dots, y_a(n - L + 1)]^T$, $\mathbf{f}(n) = [f_1(n), \dots, f_\zeta(n), \dots, f_L(n)]^T$ and $f_l(n)$ is the l^{th} coefficient of the FIR filter. ζ is selected such that $f_\zeta(n)$ is a tap near the middle of $\mathbf{f}(n)$, thereby allowing us to accommodate negative as well as positive intersensor time-delays.

If the delay may be assumed to be an integer multiple of the sampling period or if the sampling rate used provides sufficient temporal resolution, we may determine the

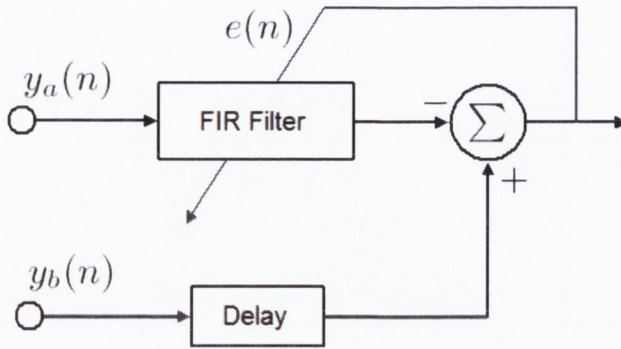


Figure 5.1: Time Delay Estimation using the LMS algorithm.

time-delay estimate as shown below.

$$\frac{\tilde{\tau}_{LMS}}{T} = \left[\arg \max_l \{f_l(n)\} \right] - \zeta \quad (5.13)$$

Alternatively, we may, by a variety of numerical methods, apply the interpolation formula to increase the temporal resolution before finding an improved TDE as above, [73],[74],[75]. In [76], only the maximum filter-tap weight is updated and all other weights are determined by reference to a look-up table of sinc-function samples. In this way, the filter-tap coefficients are effectively constrained to be samples of a sinc function. This is shown to lead to improved convergence.

In [77], it is shown that the true delay (whether an integer or non-integer multiple of T) may be expressed as a function of the filter-tap weights. By this method, known as direct-delay-estimation (DDE) we may achieve superior accuracy and convergence over the conventional or look-up-table approaches. In [78], the authors combine DDE with a unity-norm constraint on \mathbf{f} (the norm of samples of the delay sinc function may be shown to equal unity) resulting in additional performance improvements over DDE alone.

5.2.4 Adaptive Eigenvalue Decomposition

Adaptive eigenvalue decomposition (AED) techniques - proposed and developed by Benesty et al [79]-[83] - explicitly incorporate multichannel effects to obtain TDEs that are robust to reverberation. Following from the commutative properties of convolution

$$x_a(t) * h_b = s(t) * h_a * h_b = x_b(t) * h_a \quad (5.14)$$

Using a L -tap FIR filter, \mathbf{f}_m , to model h_m

$$\mathbf{x}_a^T(n)\mathbf{f}_b - \mathbf{x}_b^T(n)\mathbf{f}_a = 0 \quad (5.15)$$

or alternatively, letting $\mathbf{x}(n) = [\mathbf{x}_a^T(n), \mathbf{x}_b^T(n)]^T$ and $\mathbf{u} = [\mathbf{f}_b^T, -\mathbf{f}_a^T]^T$,

$$\mathbf{x}^T(n)\mathbf{u} = 0 \quad (5.16)$$

Premultiplying by $\mathbf{x}(n)$ and applying the expectation operator yields

$$\mathbf{K}_{\mathbf{xx}}\mathbf{u} = 0 \quad (5.17)$$

where $\mathbf{K}_{\mathbf{xx}} = E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$. Thus \mathbf{u} is the eigenvector of $\mathbf{K}_{\mathbf{xx}}$ corresponding to the eigenvalue 0. In the presence of uncorrelated noise, no eigenvalue equals 0 and so we seek the eigenvector corresponding to the minimum eigenvalue. This may be achieved by means of generalized singular value decomposition or the iterative approach shown below, [80].

$$\mathbf{u}(n+1) = \frac{\mathbf{u}(n) - \mu\mathbf{u}^T(n)\mathbf{x}(n)\mathbf{x}(n)}{|\mathbf{u}(n) - \mu\mathbf{u}^T(n)\mathbf{x}(n)\mathbf{x}(n)|} \quad (5.18)$$

under the constraint $|\mathbf{u}(n)| = 1$. From, $\mathbf{u}(n)$, the corresponding TDE may be calculated as the difference in the time-lags of the first arriving impulse of the $\mathbf{f}_a(n)$ and $\mathbf{f}_b(n)$ components.

Convergence of (5.18) requires that the Z-transforms of \mathbf{f}_a and \mathbf{f}_b do not share common zeros, [79]. Unfortunately, common zeros are likely, particularly as the impulse responses/filter lengths become longer, [65]. However, given multiple microphones we may construct a multi-microphone blind channel-estimation approach where, compared with the two-channel case, the likelihood of all of the impulse responses sharing a zero is significantly reduced. Given M channels we can construct a vector of concatenated L -tap filters, \mathbf{f}_C and concatenated blocks of microphone output data, \mathbf{x}_C ,

$$\mathbf{x}_C(n) = [\mathbf{x}_0^T(n), \mathbf{x}_1^T(n), \dots, \mathbf{x}_{M-1}^T(n)]^T \quad (5.19)$$

$$\mathbf{f}_C(n) = [\mathbf{f}_0^T(n), \mathbf{f}_1^T(n), \dots, \mathbf{f}_{M-1}^T(n)]^T \quad (5.20)$$

The converged value of \mathbf{f}_C is that which minimizes the cost function, Ω ,

$$\Omega(n+1) = \sum_{a=0}^{M-2} \sum_{b=a+1}^{M-1} \frac{\mathbf{x}_a^T(n)\mathbf{f}_b(n) - \mathbf{x}_b^T(n)\mathbf{f}_a(n)}{|\mathbf{f}_C(n)|} \quad (5.21)$$

The cost function may be minimized using a LMS-type algorithm such as that derived in [83], shown below.

$$\mathbf{f}_C(n+1) = \frac{\mathbf{f}(n) - 2\mu[\mathbf{K}_{\mathbf{x}_C\mathbf{x}_C}(n+1)\mathbf{f}_C(n) - \Omega(n+1)\mathbf{f}_C(n)]}{|\mathbf{f}(n) - 2\mu[\mathbf{K}_{\mathbf{x}_C\mathbf{x}_C}(n+1)\mathbf{f}_C(n) - \Omega(n+1)\mathbf{f}_C(n)]|} \quad (5.22)$$

In [82], a filter update equation based upon the Newton optimization method is derived and is shown to achieve faster convergence than (5.22) but with greater computational complexity. In [81], a frequency-domain normalized-LMS approach is proposed and demonstrated to represent a balance between the requirements for fast convergence and low computational cost.

Despite the elegance of the AED approach, blind estimation of the source-microphone impulse responses is a non-trivial problem that is complicated by the spectral sparseness of $s(t)$. Although the impulse response estimates in $\mathbf{u}(n)$ are accurate enough for time delay estimation, they are not sufficiently accurate to be applied to, for example, dereverberation.

5.3 TDE-based Source Localization

In the previous section we reviewed techniques for obtaining time-delay estimates. In the following section we shall discuss the methods by which, when given $\{\tilde{\tau}_m\}$ and the set of microphone locations, $\{\vec{m}_m\}$, we may find \vec{s} .

5.3.1 Viète's Solution

Given $\{\tilde{\tau}_m\}$ and $\{\vec{m}_m\}$, the problem of determining \vec{s} may be considered a practical application of a classical geometry problem - Apollonius' problem of tangent circles, [84]. In the two-dimensional case, illustrated in figure (5.2), \vec{s} may be found as the centre of the circle that is externally tangent to the circles with centres $\{\vec{m}_m\}$ and respective radii $\{\tilde{\tau}_m\}$.

Although soluble with compass and straight edge, the numerical solution proposed by Viète, [85], is of greater practical interest because, in addition to the obvious advantages of numerical implementation, it may easily be expanded to the 3-dimensional case (tangent spheres). Viète's solution is the system of simultaneous equations shown below.

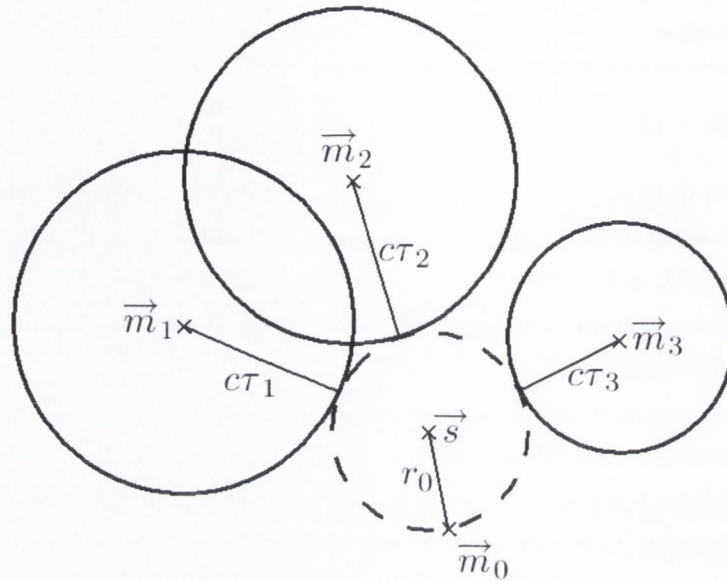


Figure 5.2: Source localization as an instance of Apollonius' problem of tangent circles.

$$\begin{aligned}
 (s_x - m_{0,x})^2 + (s_y - m_{0,y})^2 + (s_z - m_{0,z})^2 &= (r_0)^2 & (5.23) \\
 (s_x - m_{1,x})^2 + (s_y - m_{1,y})^2 + (s_z - m_{1,z})^2 &= (r_0 + c\tau_1)^2 \\
 &\vdots & \\
 (s_x - m_{M-1,x})^2 + (s_y - m_{M-1,y})^2 + (s_z - m_{M-1,z})^2 &= (r_0 + c\tau_{M-1})^2
 \end{aligned}$$

where $r_0 = |\vec{s} - \vec{m}_0|$ and $m_{0,x}$ is the x -component of \vec{m}_0 . There are four unknowns, $\{s_x, s_y, s_z, r_0\}$, and so we require a minimum of four equations/microphones. However, this is in itself insufficient to ensure that we will obtain a unique and discrete value for \vec{s} , as the system of equations may, nonetheless, be under-determined. Consider, for example, a scenario in which the source and microphones are colinear. It is, perhaps, intuitive and may be verified mathematically that in such a case $\{\tau_m\}$ will be constant for all non-negative values of r_0 and hence an infinite number of solutions exist. Furthermore, due to estimation errors, inserting $\{\tilde{\tau}_m\}$ into (5.23) may yield an inconsistent system of equations with no solution.

We require methods that overcome these shortcomings. As we shall see, this can generally be achieved in one of three ways (or a combination thereof). The first of these is to redefine the problem such that \vec{s} need not be a unique or discrete point



Figure 5.3: The far-field assumption: For distant sound sources, the curvature of the incident wavefront is negligible.

in space. Alternatively, we may find $\tilde{\vec{s}}$ by means of a minimum-mean-square-error (MMSE) fitting of the data. Finally, we may assume $\tilde{\vec{s}}$ to be an element of a subset of candidate source locations thereby allowing us to restrict our search to the subset of corresponding candidate $\{\tau_m\}$.

5.3.2 Direction-of-Arrival Estimation

Direction-of-arrival (DOA, θ) estimation techniques may be applied with as few as two microphones and follow from the assumption that the sound source is in the far-field of the array. In such cases, as illustrated in figure (5.3), the curvature of the direct-path wavefront may be considered negligible - hence we assume a planar wavefront. The DOA for each microphone pair is therefore a function of the corresponding TDE, figure (5.4).

$$\tilde{\theta} = \arcsin\left(\frac{c\tilde{\tau}_{a,b}}{d}\right) \quad (5.24)$$

This method has rotational symmetry, therefore $\tilde{\vec{s}}$ may be any point on the cone with apex $\left(\frac{\vec{m}_b + \vec{m}_a}{2}\right)$ aperture $(180 - 2\theta)^\circ$ and its axis along the line $(\vec{m}_b - \vec{m}_a)$ (although the closer we get to the apex of this cone the less valid our far-field assumption becomes).

When the geometry of the array is two-dimensional, it is possible to obtain DOA

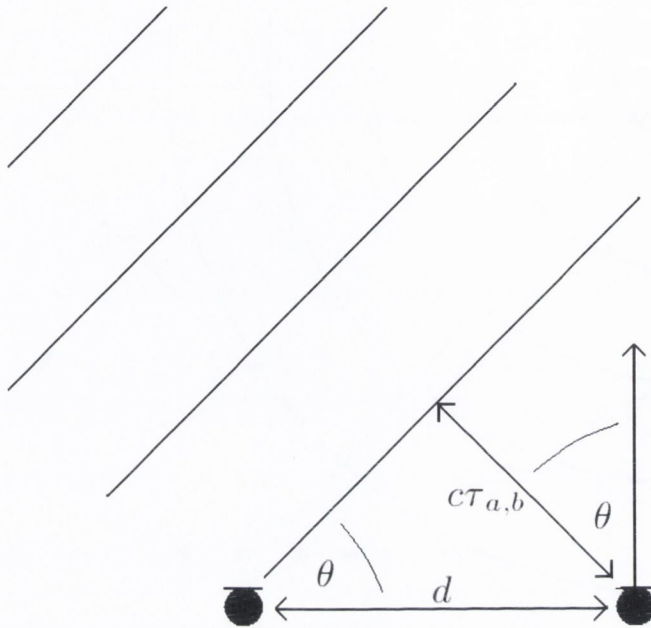


Figure 5.4: The direction of arrival of a far-field sound source is a function of the intersensor time delay.

estimates in each dimension. These estimates of the azimuth ($\tilde{\theta}$) and elevation ($\tilde{\phi}$) characterize a “bearing-line” along which the source will lie.

For many applications, a DOA/bearing-line is all that is required to characterize the spatial characteristics of the source. These include any scenario in which the range to the source is irrelevant (e.g. automatic camera steering) or may be inferred from other data. However, for other applications, an obvious extension of the bearing-line approach allows us to obtain a more refined SLE as that point where bearing lines intersect, [86],[87]. When bearing lines cannot be expected to converge, we may use the method proposed in [88] and [89], where \tilde{s} is found as a weighted average of the points on each bearing line that are the shortest perpendicular distance from the remaining bearing-lines. The weighting is chosen according to the probability distribution of the TDEs used to calculate the θ and ϕ values.

DOA-based approaches are popular in the literature due to the simplicity of their implementation, their low computational cost and the minimal estimation time-lag (important when tracking a moving source). However, a significant drawback of these methods lies in their susceptibility to error as a result of steering-quantization (see section 3.4.3). For practical applications, the elements of $\{\tilde{\tau}_m\}$ will be quantized. As a result, $\{\tilde{\theta}, \tilde{\phi}\}$ is also limited to a subset of discrete values. It is required, therefore, that d be increased or T reduced (either by increasing the sampling rate or interpolation)

until $\{\tilde{\theta}, \tilde{\phi}\}$ may be found with some specified accuracy.

5.3.3 Least-Square-Error Fitting

DOA-based approaches assume that the source is in the far-field, thus rendering them unsuitable for application to the general case in which the source may be in the near-field or interior of the array. In [90], a least-squares source location estimator is derived for implementation with an ad-hoc but known deployment of sensors. Without loss of generality, we let the reference microphone be at the origin of our coordinate system, i.e. $\vec{m}_0 = [0, 0, 0]^T$.

$$|\vec{s} - \vec{m}_m|^2 = |\vec{s}|^2 + 2\vec{s} \cdot \vec{m}_m + |\vec{m}_m|^2 \quad (5.25)$$

From our definition of τ_m , and remembering that $|\vec{m}_0| = 0$

$$|\vec{s} - \vec{m}_m| = c\tau_m - |\vec{s}| \quad (5.26)$$

Inserting (5.26) into (5.25)

$$\vec{s} \cdot \vec{m}_m + c\tau_m |\vec{s}| = \frac{1}{2} \left(|\vec{m}_m|^2 - |c\tau_m|^2 \right) \quad (5.27)$$

For M microphones we can construct the following matrix expression

$$\mathbf{A}\mathbf{b} = \mathbf{c} \quad (5.28)$$

where

$$\mathbf{A} = \begin{bmatrix} m_{1,x} & m_{1,y} & m_{1,z} & c\tau_1 \\ m_{2,x} & m_{2,y} & m_{2,z} & c\tau_2 \\ \vdots & \vdots & \vdots & \vdots \\ m_{M-1,x} & m_{M-1,y} & m_{M-1,z} & c\tau_{M-1} \end{bmatrix} \quad (5.29)$$

$$\mathbf{c} = \frac{1}{2} \left[\left(|\vec{m}_1|^2 - |c\tau_1|^2 \right), \left(|\vec{m}_2|^2 - |c\tau_2|^2 \right), \dots, \left(|\vec{m}_{M-1}|^2 - |c\tau_{M-1}|^2 \right) \right]^T \quad (5.30)$$

$$\mathbf{b} = [s_x, s_y, s_z, |\vec{s}|]^T \quad (5.31)$$

If we treat s_x, s_y, s_z and $|\vec{s}|$ as independent variables then an unconstrained least squares fitting of the data is obtained via

$$\tilde{\mathbf{b}} = [\tilde{s}_x, \tilde{s}_y, \tilde{s}_z, \tilde{r}_s]^T = (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{A}^T \mathbf{c} \quad (5.32)$$

We have replaced $|\vec{s}|$ with \tilde{r}_s to emphasize the fact that, in general, $\tilde{s}_x^2 + \tilde{s}_y^2 + \tilde{s}_z^2 \neq \tilde{r}_s^2$. As such, $\tilde{\mathbf{b}}$ is inconsistent with $\tilde{\mathbf{s}}$ being a discrete point in space. We may resolve this contradiction by placing constraints on $\tilde{\mathbf{b}}$. Doing so also enhances the accuracy of $\tilde{\mathbf{b}}$ and hence yields a superior estimate $\tilde{\tilde{\mathbf{s}}}$, [91],[92].

$$\tilde{\mathbf{b}} = \underset{\mathbf{b}}{\arg \min} (\mathbf{A}\mathbf{b} - \mathbf{c})^T (\mathbf{A}\mathbf{b} - \mathbf{c}) \quad \text{subject to} \quad \mathbf{b}^T \mathbf{\Upsilon} \mathbf{b} = 0 \quad (5.33)$$

where $\mathbf{\Upsilon} = \text{diag}\{1, 1, 1, -1\}$. Solving using Lagrange multipliers yields an iterative solution by which we may obtain successively improved $\tilde{\mathbf{b}}$

$$\tilde{\mathbf{b}}_{l+1} = [\mathbf{I} - \rho(\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{\Upsilon}] \tilde{\mathbf{b}}_l \quad (5.34)$$

where ρ is a small constant and l is the update index.

In [93], a least-squares source locator is derived for the special case where the speed of propagation is unknown. The authors' final solution may be expressed as

$$\mathbf{A}\mathbf{b}(\vec{s}) = \boldsymbol{\tau} \quad (5.35)$$

where

$$\mathbf{A} = \begin{bmatrix} -(\vec{m}_1 - \vec{m}_0) & |\vec{m}_1 - \vec{m}_0| & -\tau_1 \\ \vdots & \vdots & \vdots \\ -(\vec{m}_{M-1} - \vec{m}_0) & |\vec{m}_{M-1} - \vec{m}_0| & -\tau_{M-1} \end{bmatrix} \quad (5.36)$$

$$\mathbf{b}(\vec{s}) = \left[\frac{s_x - m_{0,x}}{c|s_x - m_{0,x}|}, \frac{s_y - m_{0,y}}{c|s_y - m_{0,y}|}, \frac{s_z - m_{0,z}}{c|s_z - m_{0,z}|}, \frac{1}{2c|\vec{s} - \vec{m}_0|}, \frac{c}{2|\vec{s} - \vec{m}_0|} \right]^T \quad (5.37)$$

$$\boldsymbol{\tau} = [\tau_1, \tau_2, \dots, \tau_{M-1}]^T \quad (5.38)$$

However, this technique is of only limited advantage in indoor scenarios for which $c \approx 340ms^{-1}$ represents an accurate estimate.

5.4 Parametric Methods

The source-localization techniques reviewed in the previous section followed a two-step process. First, TDEs were found after which a SLE was obtained. In the following section, we review single-step techniques whereby $\tilde{\mathbf{s}}$ is found following the evaluation of a parametric equation.

5.4.1 Maximum-Likelihood Method

The best-known parametric source-localization technique is based on a Maximum-Likelihood (ML) approach. Assuming an anechoic environment

$$\mathbf{Y}(\omega) = \mathbf{D}(\omega, \vec{s})S(\omega) + \mathbf{N}(\omega) \quad (5.39)$$

we may develop an expression for M microphones and P sources.

$$\mathbf{Y}(\omega) = \mathbf{D}_{1:P}(\omega, \vec{s})\mathbf{S}_{1:P}(\omega) + \mathbf{N}(\omega) \quad (5.40)$$

where

$$\vec{s} = [\vec{s}_1, \vec{s}_2, \dots, \vec{s}_P]^T \quad (5.41)$$

$$\mathbf{D}_{1:P}(\omega, \vec{s}) = [\mathbf{D}(\omega, \vec{s}_1), \mathbf{D}(\omega, \vec{s}_2), \dots, \mathbf{D}(\omega, \vec{s}_P)] \quad (5.42)$$

$$\mathbf{S}_{1:P}(\omega) = [S_1(\omega), S_2(\omega), \dots, S_P(\omega)]^T \quad (5.43)$$

If the noise statistics are known, then for a given parameter set, $\eta(\omega)$, we may obtain an expression for the conditional probability distribution of $\mathbf{Y}(\omega)$. We assume temporally and spatially white noise - i.e. that the noise signals defined by the elements of $\mathbf{N}(\omega)$ are uncorrelated with each other and correspond to zero-mean Gaussian noise with variance δ^2 .

$$Pr(\mathbf{Y}(\omega)/\eta(\omega)) = \frac{1}{(\pi\delta^2)^M} \exp \left\{ -\frac{|\mathbf{Y}(\omega) - \mathbf{D}_{1:P}(\omega, \vec{s})\mathbf{S}_{1:P}(\omega)|^2}{\delta^2} \right\} \quad (5.44)$$

where $\eta(\omega) = \{\vec{s}, \mathbf{S}_{1:P}(\omega)\}$. Since $\ln\{B\}$ is monotonically increasing with B , as an alternative to maximizing $Pr(\mathbf{Y}/\eta(\omega))$ we may instead minimize $-\ln\{Pr(\mathbf{Y}/\eta(\omega))\}$. Ignoring irrelevant constant terms,

$$\tilde{\eta}(\omega) = \arg \min_{\eta(\omega)} \left\{ |\mathbf{Y}(\omega) - \mathbf{D}_{1:P}(\omega, \vec{s})\mathbf{S}_{1:P}(\omega)|^2 \right\} \quad (5.45)$$

Minimizing this expression with respect to $\mathbf{S}_{1:P}(\omega)$ yields the estimate,

$$\tilde{\mathbf{S}}_{1:P}(\omega) = (\mathbf{D}_{1:P}^H(\omega, \vec{s})\mathbf{D}_{1:P}(\omega, \vec{s}))^{-1} \mathbf{D}_{1:P}^H(\omega, \vec{s})\mathbf{Y}(\omega) \quad (5.46)$$

Inserting (5.46) into (5.45) and integrating across the bandwidth of the signal(s) to obtain a unified, frequency-independent estimate.

$$\vec{s}_{ML} = \arg \min_{\vec{s}} \left\{ \int |\mathbf{G}(\omega, \vec{s}) \mathbf{Y}(\omega)|^2 d\omega \right\} \quad (5.47)$$

where

$$\mathbf{G}(\omega, \vec{s}) = \mathbf{I} - \mathbf{D}_{1:P}(\omega, \vec{s}) (\mathbf{D}_{1:P}^H(\omega, \vec{s}) \mathbf{D}_{1:P}(\omega, \vec{s}))^{-1} \mathbf{D}_{1:P}^H(\omega, \vec{s}) \quad (5.48)$$

Note that for practical implementations of the frequency domain ML source locator, we will need to use DFTs which will introduce a circular shift. However, where the segments are sufficiently long, the resulting edge effects will not seriously degrade our estimates, [94].

While the ML source locator may be concisely expressed, obtaining a solution to (5.47) requires a P -dimensional optimization - a non-trivial problem in itself. Iterative approaches, following the steepest descent, Newton-Raphson method [95] and Gauss-Newton [96] methods have been proposed. However, in these, the objective function to be minimized contains multiple local minima making an accurate initial estimate critical, [97].

Furthermore, while convergence to the optimum solution may be improved by updating each \vec{s}_p sequentially, [94],[98], in general the ML approach is too computationally complex to be practical.

5.4.2 Steered Response Power Techniques

For a single source, an expression equivalent to (5.47) may be written as

$$\vec{s} = \arg \max_{\vec{s}} \left\{ \int |\mathbf{D}^H(\omega, \vec{s}) \mathbf{Y}(\omega)|^2 d\omega \right\} \quad (5.49)$$

Comparing the expression above to the D&S beamformer (see section 4.2), it becomes apparent that in a single source, spatially and spectrally white noise, scenario the maximum-likelihood SLE is obtained by steering a D&S beamformer to all locations and selecting \vec{s} as that location which returns the maximum output power. In fact, this approach is one of the earliest source-localisation strategies - commonly known as the steered-response-power (SRP) approach.

For a broadband signal using an arbitrary F&S beamformer, we may obtain \vec{s}_{SRP} as follows

$$\vec{s}_{SRP} = \arg \max_{\vec{s}} \left[\int \left| \sum_{m=0}^{M-1} W_m^*(\omega) Y_m(\omega) \exp\{j\omega\tau_m(\vec{s})\} \right|^2 d\omega \right] \quad (5.50)$$

In the expression above we write $\tau_m(\vec{s})$ to explicitly denote the dependence of inter-sensor time-delays on the source location.

Expanding (5.50) and using the concepts of cross-correlation introduced in section 5.2.2 provides a key insight into SRP techniques.

$$\begin{aligned} \vec{s}_{SRP} &= \arg \max_{\vec{s}} \left[\sum_{i=0}^{M-1} \sum_{m=0}^{M-1} \int W_i^*(\omega) W_m(\omega) Y_i(\omega) Y_m^*(\omega) \exp\{j\omega(\tau_m(\vec{s}) - \tau_i(\vec{s}))\} d\omega \right] \\ &= \arg \max_{\vec{s}} \left[\sum_{i=0}^{M-1} \sum_{m=0}^{M-1} \int \Phi_{i,m}(\omega) Y_i(\omega) Y_m^*(\omega) \exp\{j\omega(\tau_m(\vec{s}) - \tau_i(\vec{s}))\} d\omega \right] \end{aligned} \quad (5.51)$$

Observing that the integral in the expression above is equivalent to evaluating the inverse Fourier transform at $t = 0$, we obtain,

$$\vec{s}_{SRP} = \arg \max_{\vec{s}} \left[\sum_{i=0}^{M-1} \sum_{m=0}^{M-1} \psi_{i,m}(\tau_m(\vec{s}) - \tau_i(\vec{s})) \right] \quad (5.52)$$

SRP methods may, therefore, be considered a multimicrophone extension of the GCC approach outlined in section 5.2.2 and, indeed, the expression in 5.52 is equivalent to that obtained by several researchers seeking to enhance SLE accuracy by exploiting the redundancies inherent in having multiple GCC estimates, [99],[100],[101].

In general, SRP-based source localization approaches are held to be more robust to noise and, in particular, reverberation than their TDE-based counterparts, [97],[102]. This, however, comes at the expense of increased computational complexity (although this may be compensated for somewhat by our choice of search strategy, [103],[104]). This motivates hybrid techniques in which we perform the summation in (5.52) over selected, as opposed to all, microphone pairs, [105],[106],[107]. With judicious microphone-pair selection, this leads to a reduction in computational cost while maintaining the noise and reverberation robustness characteristic of SRP techniques.

As with GCC-based approaches, it has been shown to be advantageous to modify $W_m(\omega)$ (and hence $\Phi_{i,m}(\omega)$) according to the acoustic conditions. With this motivation, the SRP-PHAT, [97], seeks to emulate the reverberation-robust performance of the GCC-PHAT by applying a weightvector whereby

$$W_m(\omega) = \frac{1}{|Y_m(\omega)|} \quad (5.53)$$

In [106] and [107], further performance enhancements are obtained by the inclusion of a ‘‘spatial observability function’’ - a weighting which takes into account factors such as the room geometry, source and sensor directionality and the distance between the

array and the SLE to emphasize/de-emphasize the contribution of a microphone pair according to the reliability of its estimate.

5.5 Subspace-Based Techniques

Subspace-based source localization techniques exploit the properties of the correlation matrix, $\mathbf{R}_{\mathbf{Y}\mathbf{Y}}$, to achieve high-resolution SLEs. Subspace techniques include the Estimation of Signal Parameters by Rotational Invariance Technique, (ESPRIT), and the Decomposition of the Time Reversal Operator (DORT) technique. However, DORT requires active (that is, capable of emitting sound) sensors and ESPRIT assumes a specific array geometry. We therefore limit our review of subspace methods to a discussion of the MULTiple SIGNAL Classification (MUSIC) approach.

5.5.1 MUSIC

Following from the anechoic signal model and assuming spatially and spectrally uncorrelated noise, letting $\mathbf{R}_{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}}(\omega) = E\{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}^H\}$ we may write

$$\mathbf{R}_{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}}(\omega) = \mathbf{D}_{1:P}(\omega, \vec{s}) \mathbf{R}_{\mathbf{S}\mathbf{S}}(\omega) \mathbf{D}_{1:P}^H(\omega, \vec{s}) + \delta^2 \mathbf{I} \quad (5.54)$$

where $\mathbf{R}_{\mathbf{S}\mathbf{S}}(\omega) = E\{\mathbf{S}_{1:P}(\omega)\mathbf{S}_{1:P}(\omega)\}$. As a Hermitian matrix $\mathbf{R}_{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}}(\omega)$ may be expressed as follows.

$$\mathbf{R}_{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}}(\omega) = \sum_{m=1}^M \lambda_m(\omega) \mathbf{q}_m(\omega) \mathbf{q}_m^H(\omega) \quad (5.55)$$

where $\mathbf{q}_m(\omega)$ and $\lambda_m(\omega)$ are the m^{th} eigenvector (column vector) and eigenvalue of $\mathbf{R}_{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}}(\omega)$ respectively. Let us denote the eigenvalues in order of decreasing size. Assuming then that $M > P$ and that the signal dominates the noise, we may separate the signal and noise eigenvectors.

$$\begin{aligned} \mathbf{R}_{\mathbf{Y}_{1:P}\mathbf{Y}_{1:P}}(\omega) &= \sum_{m=1}^P \lambda_m(\omega) \mathbf{q}_m(\omega) \mathbf{q}_m^H(\omega) + \sum_{m=P+1}^M \lambda_m(\omega) \mathbf{q}_m(\omega) \mathbf{q}_m^H(\omega) \\ &= \mathbf{Q}_S(\omega) \mathbf{\Lambda}(\omega) \mathbf{Q}_S^H(\omega) + \delta \mathbf{Q}_N(\omega) \mathbf{Q}_N^H(\omega) \end{aligned} \quad (5.56)$$

where

$$\begin{aligned} \mathbf{Q}_S(\omega) &= [\mathbf{q}_1(\omega), \dots, \mathbf{q}_P(\omega)] \\ \mathbf{Q}_N(\omega) &= [\mathbf{q}_{P+1}(\omega), \dots, \mathbf{q}_M(\omega)] \\ \mathbf{\Lambda}(\omega) &= \text{diag}\{\lambda_1(\omega), \dots, \lambda_P(\omega)\} \end{aligned} \quad (5.57)$$

A further property of Hermitian matrices is that their eigenvectors are orthonormal. Thus $\mathbf{Q}_N(\omega)$ is orthogonal to $\mathbf{Q}_S(\omega)$, and assuming that $\mathbf{D}\mathbf{R}_{SS}\mathbf{D}^H$ is full-rank, it follows that

$$\begin{aligned}\mathbb{R}\{\mathbf{Q}_S(\omega)\} &= \mathbb{R}\{\mathbf{D}_{1:P}(\omega, \vec{s})\} \\ \mathbb{R}\{\mathbf{Q}_N(\omega)\} &= \mathbb{N}\{\mathbf{D}_{1:P}(\omega, \vec{s})\}\end{aligned}\quad (5.58)$$

where $\mathbb{R}\{\}$ and $\mathbb{N}\{\}$ denote the range and null-space respectively. Hence $\mathbf{Q}_N(\omega)$ and $\mathbf{D}(\omega, \vec{s})$ are orthogonal.

$$\mathbf{Q}_N^H(\omega)\mathbf{D}_{1:P}(\omega, \vec{s}) = 0 \quad (5.59)$$

To exploit this, we define the MUSIC “spatial spectrum”, $U(\vec{s})$, which will exhibit peaks at those \vec{s}_p corresponding to the true \vec{s} .

$$U(\vec{s}) = \frac{\mathbf{D}_{1:P}^H(\omega, \vec{s})\mathbf{D}_{1:P}(\omega, \vec{s})}{\mathbf{D}_{1:P}^H(\omega, \vec{s})\mathbf{Q}_N(\omega)\mathbf{Q}_N^H(\omega)\mathbf{D}_{1:P}(\omega, \vec{s})} \quad (5.60)$$

Unfortunately, whilst capable of resolving multiple simultaneous sources with a very high degree of accuracy, MUSIC has a noted susceptibility to error in the presence of multiple correlated sources, such as we would find in reverberant environments, [97],[108]. Several techniques have been proposed to reduce this susceptibility, albeit with very limited success. “Incoherent” approaches involve the application of the MUSIC algorithm within non-overlapping frequency bins, followed by a weighted averaging of the results, [109]. “Coherent” methods include the use of “focusing matrices”, [110], to translate the signal spaces for all frequency bands onto a single signal subspace while “Spatial Smoothing” methods, [111],[112], average the correlation matrices obtained from overlapping subarrays, thereby suppressing reverberant components that are not strongly correlated in each subarray output.

5.6 Discussion

In this chapter we have reviewed approaches to the related problems of time-delay estimation and source localization. In the application under consideration, sound sources are likely to be at some distance (perhaps several meters) from the microphones. Furthermore, classrooms, being large rooms, are likely to be highly reverberant. As a result, the robustness against reverberation of the various techniques we have discussed is of significant interest.

In the case of time-delay estimation, adaptive eigenvalue decomposition techniques are explicitly designed for use in reverberant environments and have been demonstrated to offer excellent performance under such conditions. However, estimation of impulse responses requires the use of very long filters which, in turn, leads to slow convergence. While by no means an insurmountable problem when the sound source is stationary (in both a physical and statistical sense), this poses significant problems if the impulse response is time varying as would occur if, for example, a talker were moving or turning his/her head. To the author's knowledge, AED has not been tested under such conditions. Generalized cross-correlation methods suffer no such convergence problems and can return time-delay estimates using sound samples as short as a few tens of milliseconds. This, combined with its demonstrated reverberation robustness, make the GCC-PHAT the benchmark time-delay estimator for reverberant environments.

The likely presence of significant levels of reverberation in classrooms causes us to reject MUSIC and other subspace-based methods as being unsuitable for source localization. In contrast, SRP and TDE-based techniques have demonstrated to offer robust performance in reverberant environments (although the success or otherwise of TDE-based approaches will depend upon the methods by which the TDEs were obtained). In general, SRP techniques offer superior performance. However, hybrid techniques, such as the SRP-PHAT, allow us to achieve performance approaching that of SRP methods without exceeding a given tolerance for computational complexity.

We note, that to obtain an estimate of the source location we must have knowledge of the location of the microphones in the array, $\{\vec{m}\}$. This requirement may be intuitive to many readers and is explicit in the formulation of TDE-based source localization strategies. While the requirement may not be as apparent in the cases of parametric and subspace-based approaches, it exists nonetheless. ML and SRP techniques, as well as the MUSIC algorithm, are functions of $\mathbf{D}(\omega, \vec{s})$ which, from inspection of (3.59) is itself a function of $\{\vec{m}\}$. We will return to this point again in chapter 6.

We note also that, given knowledge of $\{\vec{m}\}$, a set of TDEs may easily be inferred from a source location estimate (see equation (3.57)). In this way, source localization techniques may be equally applied to the problem of time-delay estimation.

Chapter 6

Classroom-Based Videoconferencing: A Problem Overview

6.1 Introduction

In chapter 4, we reviewed previously published techniques for enhancing the quality and intelligibility of recorded speech. Among these, the class of speech enhancers known as beamformers are, perhaps, the most widely investigated and have proven popular for their effectiveness, versatility and ease of implementation. If we can assume an omnidirectional source, implementing a beamformer requires us to first estimate the steering vector $\mathbf{D}(\omega, \vec{s})$. We may do this using source localization algorithms. Maximum-Likelihood, Steered-Response-Power and MUSIC methods estimate $\mathbf{D}(\omega, \vec{s})$ directly as that which maximizes some objective function. Alternatively, $\mathbf{D}(\omega, \vec{s})$ may be inferred from knowledge of the array geometry and some estimate of the source location obtained using TDE-based source localization methods (section 5.3). Source location estimates may also be considered as an end unto themselves as they may be used for automatic camera steering, providing spatial audio cues etc.

Throughout the literature, practical implementations of source localization (and, hence, beamforming) strategies are based upon one or both of two common simplifying assumptions; firstly that the array geometry is known and secondly that the source is in the “far-field” of the array. In this chapter we outline these assumptions and highlight the instances in which they are made. We go on to show that these assumptions, while simplifying in a certain sense, actually impose design constraints that are disadvantageous in practice. We outline the potential benefits of a scenario in which the array geometry is unknown and the sources are in the near-field. We also outline

the specific technical challenges that such a scenario would pose and which the novel algorithms, presented in this thesis, are designed to address.

6.2 A Typical Videoconferencing Setup

The videoconferencing setup shown in figure (6.1) is typical of those found throughout the literature, [113]-[116]. In this setup, the microphones are arranged in a narrow array which is most often planar or linear to allow for easy mounting on a wall or surface. This array is then positioned at a distance from potential participant locations - often at the front of the room. This type of setup is motivated by two main assumptions/requirements - namely that the array geometry is known and that the source is in the “far-field” of the array.

6.2.1 The Far-Field Assumption

In many cases it is assumed that the distance between the sound source and the microphones is far greater than the separation between the microphones themselves (such as is the case in figure (6.1)). This is known as the far-field assumption. The origins of this assumption may be traced back to the Second World War where, for example, individual array elements could only be as far apart as the greatest dimensions of a ship while an enemy submarine would (hopefully) be very much farther away.

No such inherent constraints exist for indoor applications using microphone arrays. Nonetheless, the far-field assumption persists as a common and oftentimes significant simplification. Perhaps most importantly, the far-field assumption allows us to treat the source as being omnidirectional. In reality, this may not be the case but because each microphone is at roughly the same azimuth and elevation relative to the source, any variations in source gain due to changes in the relative azimuth/elevation (i.e. directivity) may be ignored. Assuming an omnidirectional source allows us to replace $\mathbf{H}_{dp}(\omega)$ with $\mathbf{D}(\omega, \vec{s})$ (section 3.5.2)

As seen in chapter 4, the steering vector, $\mathbf{D}(\omega, \vec{s})$, is of fundamental importance in non-adaptive and constrained optimization-based speech enhancement techniques (see sections 4.2 and 4.3 respectively). In practice, the steering vector must be estimated - a non-trivial problem. As a consequence of the far-field assumption, we may obtain $\mathbf{D}(\omega, \vec{s})$ from TDEs alone (4.2). This greatly simplifies practical implementations of the previously mentioned speech-enhancement techniques.

The far-field assumption also allows us to treat the curvature of a wavefront propagating across the array as negligible. As such, it underpins the DOA or bearing-line source localization strategies described in section 5.3.2.

6.2.2 Known Array Geometry

Many array processing techniques assume knowledge of the relative microphone locations, $\{\vec{m}\}$. In particular, source-localization requires knowledge of the array geometry and is, otherwise, impossible to implement. This is the case regardless of the sense in which the problem is defined (see section 5.1). The requirement that the array geometry be known is apparent for TDE-based source localization strategies. In these, microphone coordinates are explicitly treated as known variables in the source localization functions. This requirement is equally (if perhaps implicitly) present for parametric and subspace-based approaches, which are based upon estimation of the steering vector, $\mathbf{D}(\omega, \vec{s})$. We cannot infer \vec{s} from $\mathbf{D}(\omega, \vec{s})$ without knowledge of $\{\vec{m}\}$.

A known array geometry is also a significant advantage for practical implementations of techniques requiring steering - i.e. beamforming. Typically, for practical applications using time-sampled data, the TDEs are limited to being integer multiples of the sampling period. Consequently, the source locations to which we may precisely steer are also limited to a number of discrete locations. When the true source location does not coincide with one of these, the resulting missteering leads to a reduction in beamformer performance.

To increase the number of steerable locations, designers may reduce the sampling period by increasing the sampling frequency or by using interpolation filters (see section 3.4), albeit with a consequent increase in the computational load. Given knowledge of the geometry of the array, we may determine the maximum sampling period that will allow us to steer to all potential source locations with some specified accuracy.

In addition, when the relative microphone locations are known, the delays required for steering may be inferred for all microphones from an incomplete set of TDEs, thereby reducing computational complexity.

6.2.3 Disadvantages

As we have seen, the setup in figure (6.1) provides the potential advantages associated with a known array geometry and a far-field source. However, such a setup also has some inherent disadvantages and we outline these in this section.

Increased microphone-talker separation

Under the far-field assumption, each microphone will be at a large distance (perhaps several meters) from the source. By increasing microphone-talker separation, we increase the attenuation of the direct-path speech component due to propagation losses.

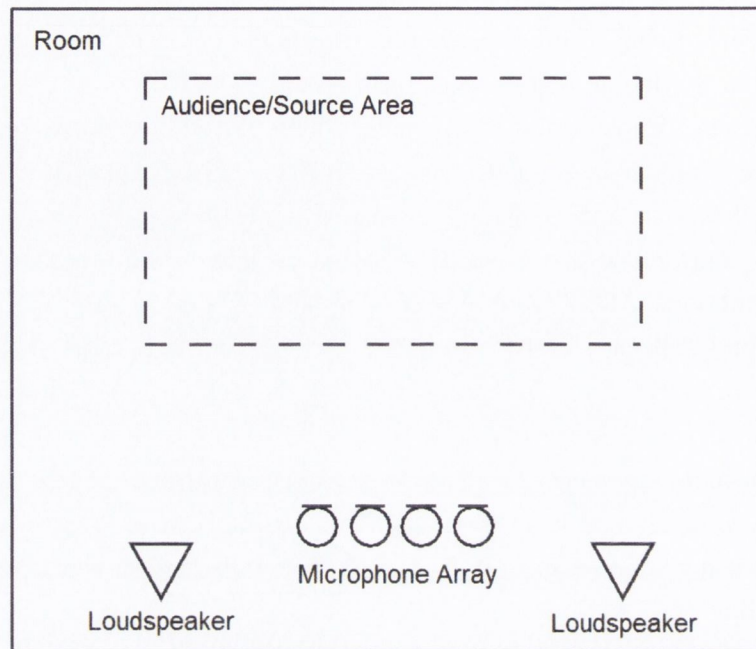


Figure 6.1: A videoconferencing setup as typically found in the literature.

As a result, the SNR and DRR of the recorded sound is reduced, thereby increasing the degree of noise and reverberation suppression required.

Reduced microphone-loudspeaker separation

By placing microphones at the front of the room, we increase their proximity to the loudspeakers. This increases the severity of any acoustic echo that may be present and is, therefore, not to be desired.

Narrow array width

As a consequence of the far-field requirement, the width/extent of the array is small. As shown in chapter 3, the performance of array-processing algorithms may generally be said to be increasing with increasing array extent. While this is most evident in the case of the D&S beamformer (see section 3.3.3), it remains equally true for all F&S-based approaches including data-adaptive speech enhancers and source-localization techniques (3.4.6). Therefore, by constraining the array extent, the far-field assumption constrains system performance also.

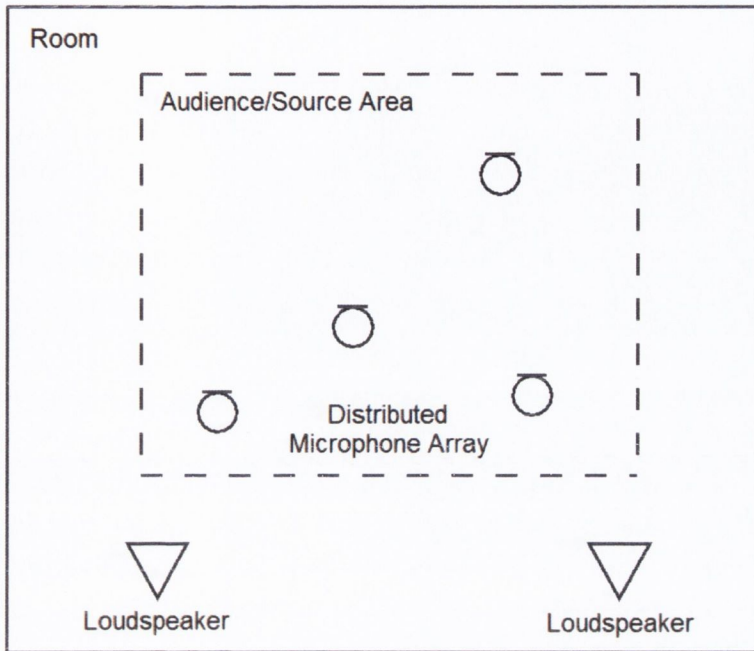


Figure 6.2: A videoconferencing setup using an array of distributed microphones.

Poor Flexibility/Scalability

The setup in figure (6.1) lacks scope for adapting to variations in audience numbers or distribution. For example, the setup is biased against rearward audience members in that their contributions will be less intelligible than those from participants closer to the microphones. It would therefore be unsuitable for long classrooms. A possible solution would entail the use of multiple arrays located around the room. However, this may be difficult to achieve while still satisfying the far-field requirement.

6.3 The Proposed Videoconferencing Setup

It is proposed to employ a videoconferencing setup such as that shown in figure (6.2), in which microphones are distributed throughout the audience area. As a consequence, an active talker is more likely to be close to at least one microphone. This microphone will, in turn, be farther from the loudspeakers, leading to improved speech intelligibility and reduced acoustic echo.

It is also proposed that, rather than being fixed, the microphone array be deployed

in an ad-hoc manner. In this way, individual microphones may be positioned according to the conditions specific to any given scenario, allowing the array configuration to adapt to varying audience distributions or numbers.

Figure (6.3) illustrates a special case of the distributed microphones array approach, whereby the array is composed of distributed sub-arrays. In certain instances, such a setup may be preferred to that shown in figure (6.2) and so it merits specific mention here. Primarily, a distributed sub-array setup would be motivated by a requirement to suppress spatially localized sources of noise (i.e. where the noise has an associated direction of origin, as opposed to reverberation which arrives at a microphone from many multiple directions or sensor noise which is uncorrelated in space). Where microphones are closely spaced, the respective source-microphone impulse responses (corresponding to the noise source) are strongly correlated with respect to the intensity and relative delay of the significant reflected components etc. Where the noise components in the microphone outputs are very similar, we may achieve a high degree of noise attenuation by means of simple methods such as null steering (section 3.4.1). However, as microphone separation increases, impulse response correlation (and hence the achievable noise suppression) reduces.

While providing flexibility and superior speech capture, an ad-hoc, changing and distributed deployment of microphones makes accurate estimation of microphone locations cumbersome, time-consuming and prone to error. Therefore, to be practical, we must assume that **the array geometry is unknown**. This presents several technical challenges and we seek to address two of these - steering and source-microphone range estimation - in this thesis.

6.3.1 Steering

Steering – whereby the target signal components in the outputs of multiple microphones are time-aligned by the application of appropriate delays or phase-shifts – is of fundamental importance in array-processing. Beamformers, in particular, require accurate steering (section 4.2). When missteering occurs, the speech enhancement offered by beamformers degrades. This is particularly true of the generalized sidelobe canceller (a data-dependent beamformer), for which inaccurate steering can lead to significant target-signal cancellation (section 4.3.2).

Before steering an array, the appropriate delays/phase shifts must first be determined. Time-delay estimates (TDEs) may be obtained directly, using cross-correlation, adaptive eigenvalue decomposition or LMS-based techniques (section 5.2). Alternatively, we may employ maximum-likelihood or subspace-based source-localization techniques (sections 5.4 and 5.5 respectively). Using these, TDEs may be inferred from an

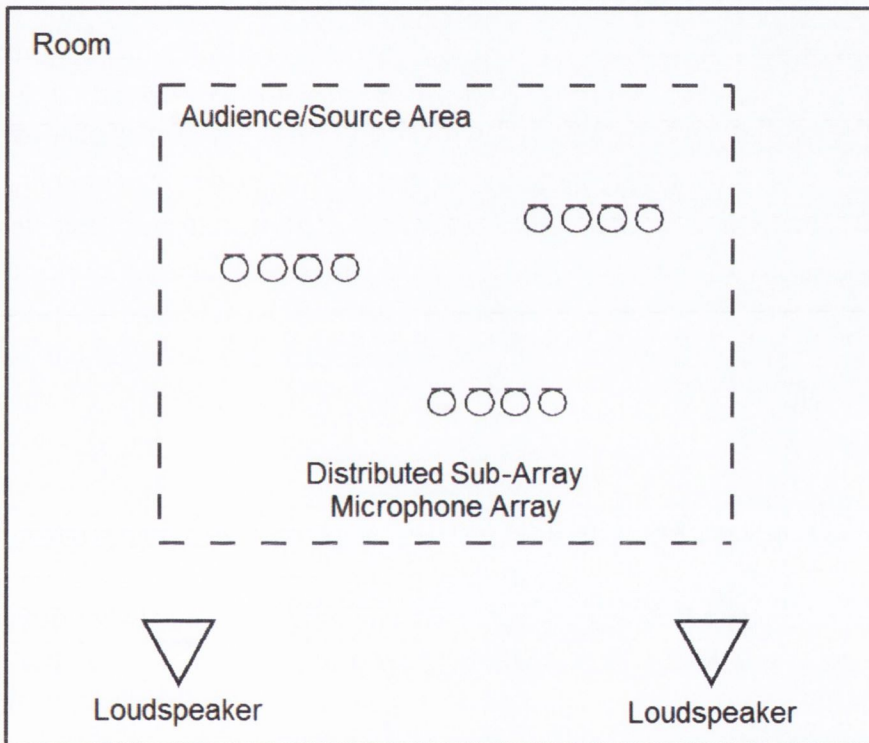


Figure 6.3: A videoconferencing setup using an array of distributed sub-arrays.

estimate of the source location and knowledge of the relative microphone positions. In the scenarios under investigation, such an approach would be limited to application with distributed sub-arrays where we may reasonably assume knowledge of the relative locations of the microphones in any given sub-array.

In each of the approaches we have mentioned, the delay-vector is estimated as that which minimizes or maximizes some function. However, not every possible delay-vector is realizable in practice. Typically, therefore, the estimated delay vector is arrived at by testing each of a subset of candidate delay-vectors. As previously mentioned, if we possess knowledge of the array geometry we may further restrict the candidate delay vectors to a subset sufficient to steer to all possible source locations with satisfactory accuracy. If we can also assume that the source is in the far-field, we may restrict the candidate delay-vectors to that subset sufficient to steer toward all directions-of-arrival (these are, in general, very much fewer than those required to steer to all discrete locations). However, when the array geometry is unknown, such simplifications are not possible. In a such a scenario, we would need to test all realizable delay vectors which, in turn, must be very large in number – possibly requiring interpolation filters or very high sampling rates – to attempt to ensure full coverage of all possible source locations. This increases our computational requirements and without knowing the array geometry we cannot be sure that full coverage is achieved or to what degree of accuracy.

Also, when we may assume that the source is in the far-field of the array, we may consider the soundwaves originating at that source as being simple plane waves. We may, therefore, easily infer intersensor time-delay estimates from an incomplete set of TDEs - allowing for further reductions in computational complexity. However, where the source is potentially in the near-field (as will be the case in a distributed sub-array scenario), no such inference may be made and we must obtain TDEs for each microphone individually.

In chapter 7, we derive a method for precise beamformer steering, based upon a multichannel, leaky LMS filter. Our technique is suitable for practical implementation in moderately reverberant environments and using arrays of unknown geometry. The proposed method is shown to be more computationally efficient than a conventional steering approach in which the TDEs are determined using the well-known PHAT-GCC technique. We also present the results of experiments using real and simulated data, to demonstrate the efficacy of our approach.

6.3.2 Source-Microphone Range Estimation

In the absence of knowledge of the relative microphone locations, it is impossible to perform source-localization using array-processing techniques. However, for many applications, knowledge of the source-microphone range is sufficient and the precise source location need not be known. Some of the simplest of these would use range estimates to select the microphone(s) that is closest to a target sound source or farthest from some noise source, thereby providing a degree of speech-quality enhancement. Range estimates could also be used to inform more sophisticated speech enhancement strategies. Since the DRR of recorded speech decreases with increasing source-microphone separation (see figure 2.8), range estimates could be considered when deciding whether or not to apply a dereverberation algorithm to a microphone output.

Range estimates may also be applied to the problem of source identification or “speaker segmentation”, whereby recorded sound is labeled or “tagged”, according to which of a number of spatially separated sources/talkers is active. In [106] and [107], estimates of the source-microphone range are incorporated as one of a number of criteria used to assess the reliability of time-delay estimates obtained by distributed microphone pairs.

In chapter 8 we shall derive an algorithm by which we may determine the distance between a sound source and the microphones in an array of unknown geometry. As we shall demonstrate, our approach is robust against reverberation.

Chapter 7

A Leaky-LMS-Based Method for Precise Beamformer Steering

7.1 Introduction

In this chapter we present a novel method for beamformer-steering, based upon the LMS filter. Traditionally, steering is performed in two stages. In the first of these, the appropriate steering delays are estimated using time-delay-estimation techniques (section 5.2). These are then used to determine the characteristics of the all-pass filters which should be used to correctly steer the array. To ensure that the time-delay estimates so obtained achieve some required degree of accuracy, the sampling period used must be sufficiently small. Very often this requires the use of increased sampling rates or interpolation filters. Our method, however, achieves precise steering even where the appropriate steering delays are non-integer multiples of the sampling rate.

LMS-based approaches have been previously proposed for the problem of time-delay estimation (section 5.2.3). Such estimates could then be used for steering. In contrast with these, our approach requires no intermediate, explicit estimation of the intersensor time-delays. Furthermore and unlike the previously proposed LMS-based methods, we verify the efficacy of our approach under reverberant conditions.

7.2 Leaky-LMS-Based Beamformer Steering

Our technique employs a filter-and-sum architecture. The outputs of each microphone are processed by an all-pass filter which applies the delay required for steering. To ensure accurate steering, the filter-tap coefficients are periodically updated and, hence, vary in time. For simplicity, we shall assume that updates occur every sampling period. The filter-tap coefficients corresponding to the m^{th} microphone, are characterized by

the $L \times 1$ vector $\mathbf{w}_m(n)$ where n is the time sample index and L is the filter length.

$$\mathbf{w}_m(n) = [w_{m,1}(n), w_{m,2}(n), \dots, w_{m,L}(n)]^T \quad (7.1)$$

In reverberant but noise-free acoustic conditions, the output of the m^{th} filter, $z_m(n)$, is given as follows,

$$z_m(n) = \sum_{k=1}^L w_{m,k}(n)x_m(n-k) = \mathbf{w}_m^T(n)\mathbf{x}_m(n-1) \quad (7.2)$$

where $\mathbf{x}_m(n) = [x_m(n), x_m(n-1), \dots, x_m(n-L+1)]^T$. The system output, $z(n)$, is given by

$$z(n) = \sum_{m=1}^{M-1} z_m(n) = \mathbf{w}^T(n)\mathbf{x}(n-1) \quad (7.3)$$

where

$$\begin{aligned} \mathbf{w}(n) &= [\mathbf{w}_1^T(n), \mathbf{w}_2^T(n), \dots, \mathbf{w}_{M-1}^T(n)]^T \\ \mathbf{x}(n) &= [\mathbf{x}_1^T(n), \mathbf{x}_2^T(n), \dots, \mathbf{x}_{M-1}^T(n)]^T \end{aligned} \quad (7.4)$$

In practice, beamformers are unlikely to be implemented in “noise free” conditions. However, where noise is temporally sparse (i.e. not continually present – say due to occasional coughing etc.), recordings will be noise free for large periods. These periods may be identified using a pre-processing stage to discriminate between noisy and noiseless recordings.

7.2.1 Updating the Weightvector

The filter taps are updated using the well-known “leaky” LMS algorithm. Given appropriate parameter selection, the LMS filter will converge the filter-tap coefficients to those values which will minimize the expected mean squared difference between $z(n)$ and some reference signal.

$$\mathbf{w}(n) = \arg \min_{\mathbf{w}(n)} \{ \mathbf{w}^T(n)\mathbf{x}(n-1) - x_0(n-\eta) \} \quad (7.5)$$

In our implementation, the reference signal being approximated is $x_0(n-\eta)$, where η is some integer in the range $1 < \eta < L$. This may seem unusual. To estimate $x_0(n-\eta)$, using a LMS approach, we must have access to $x_0(n-\eta)$ and if that is the case why are we trying to estimate it? However, as we shall see, this formulation does

allow us to achieve useful results. Our implementation of a multichannel leaky LMS algorithm is shown below.

$$\mathbf{w}(n+1) = \alpha \mathbf{w}(n) + \mu \mathbf{x}(n-1) [x_0(n-\eta) - \mathbf{w}^T(n) \mathbf{x}(n-1)] \quad (7.6)$$

where the leakage coefficient, α , and the “stepsize”, μ , are scalar constants and $0 < \alpha < 1$. We discuss the selection of appropriate values for α and μ in section 7.2.5.

So as to best understand the relevant characteristics of the converged filters, our analysis shall be in the frequency domain. A frequency-domain equivalent of the expression in (7.6) is given by

$$\mathbf{W}(\omega, l+1) = \alpha \mathbf{W}(\omega, l) + \mu \mathbf{X}(\omega, l-1) [X_0(\omega, l) - \mathbf{W}^H(\omega, l) \mathbf{X}(\omega, l-1)]^* \quad (7.7)$$

where $*$ denotes complex conjugate and l is the update index. Note that to maintain the equivalence between (7.7) and (7.6) we must alter our definitions (originally given in section 3.5.2) of $\mathbf{X}(\omega)$ and the weightvector, $\mathbf{W}(\omega)$, by omitting $X_0(\omega)$ and $W_0(\omega)$ respectively.

$$\begin{aligned} \mathbf{X}(\omega) &= [X_1(\omega), X_2(\omega), \dots, X_{M-1}(\omega)]^T \\ \mathbf{W}(\omega) &= [W_1(\omega), W_2(\omega), \dots, W_{M-1}(\omega)]^T \end{aligned} \quad (7.8)$$

However, we also note that, because m_0 is the reference microphone, $W_0(\omega)$ will correspond to some delay that will be known *a priori* and, hence, need not be calculated. The vectors $\mathbf{N}(\omega)$ and $\mathbf{H}(\omega)$ (section 3.5.2) may be similarly redefined.

Omitting the frequency index, ω , for clarity and assuming that the coefficients of \mathbf{W} are initialized to zero.

$$\begin{aligned} E\{\mathbf{W}(l+1)\} &= \sum_{k=1}^l \mu \alpha^{l-k} [E\{\mathbf{X}(k-1) X_0^*(k-1)\} \\ &\quad - E\{\mathbf{X}(k-1) \mathbf{X}^H(k-1)\} E\{\mathbf{W}(k)\}] \end{aligned} \quad (7.9)$$

Let us assume that the system is stationary, i.e. $E\{\mathbf{X}(k-1) \mathbf{X}^H(k-1)\} = E\{\mathbf{X}(k) \mathbf{X}^H(k)\} = \mathbf{R}_{\mathbf{X}\mathbf{X}}$ and $E\{\mathbf{X}(k-1) X_0^*(k-1)\} = E\{\mathbf{X}(k) X_0^*(k)\} = \mathbf{R}_{\mathbf{X}X_0}$. As $l \rightarrow \infty$, then for an appropriate choice of μ , the weightvector converges in expectation and $E\{\mathbf{W}(l+1)\} = E\{\mathbf{W}(l)\} = \mathbf{W}_\infty$, (see [16] for a discussion of stepsize bounds and filter convergence behaviour).

$$\mathbf{W}_\infty = \frac{\mu \alpha}{(1-\alpha)} (\mathbf{R}_{\mathbf{X}X_0} - \mathbf{R}_{\mathbf{X}\mathbf{X}} \mathbf{W}_\infty) \quad (7.10)$$

Following algebraic manipulation

$$\mathbf{W}_\infty = \left(\mathbf{R}_{\mathbf{X}\mathbf{X}} + \frac{(1-\alpha)}{\mu\alpha} \mathbf{I} \right)^{-1} \mathbf{R}_{\mathbf{X}\mathbf{X}_0} \quad (7.11)$$

We note that the scaled identity matrix in (7.11) is equivalent to $\mathbf{R}_{\mathbf{N}\mathbf{N}}$ where \mathbf{N} is a noise observation vector whose elements correspond to spatially and temporally uncorrelated noise, with equal variance $\frac{(1-\alpha)}{\mu\alpha}$. Therefore, comparing (7.11) with (4.24), we see that \mathbf{W}_∞ is equivalent to the optimum multichannel Wiener filter that would be obtained if the microphone outputs contained uncorrelated white noise. Thus, employing a leaky LMS filter is mathematically equivalent to injecting spatially and temporally uncorrelated noise into the microphone outputs.

7.2.2 Decomposing the Wiener Solution

Letting $\delta^2 = \frac{(1-\alpha)}{\mu\alpha}$ and $P_s = E\{|S|^2\}$, (7.11) becomes

$$\mathbf{W}_\infty = (P_s \mathbf{H}\mathbf{H}^H + \delta^2 \mathbf{I})^{-1} (P_s \mathbf{H}\mathbf{H}_0^*) \quad (7.12)$$

Applying the matrix inversion lemma,

$$(P_s \mathbf{H}\mathbf{H}^H + \delta^2 \mathbf{I})^{-1} = \frac{1}{\delta^2} \left[\mathbf{I} - \frac{\mathbf{H}\mathbf{H}^H}{\mathbf{H}^H \mathbf{H} + \frac{\delta^2}{P_s}} \right] \quad (7.13)$$

Inserting this result into (7.12) yields

$$\mathbf{W}_\infty = \frac{P_s}{\delta^2} \left[\mathbf{I} - \frac{\mathbf{H}\mathbf{H}^H}{\mathbf{H}^H \mathbf{H} + \frac{\delta^2}{P_s}} \right] \mathbf{H}\mathbf{H}^H \mathbf{v}_0 \quad (7.14)$$

and following some simple algebraic manipulation we obtain

$$\mathbf{W}_\infty = \frac{P_s \mathbf{H}\mathbf{H}_0^*}{P_s |\mathbf{H}|^2 + \delta^2} \quad (7.15)$$

7.2.3 Flattening the Filter Responses

We make the following approximation

$$\mathbf{H} \approx [\gamma_1 H_0 \exp(-j\omega\tau_1), \dots, \gamma_{M-1} H_0 \exp(-j\omega\tau_{M-1})]^T \quad (7.16)$$

where the γ terms represent real, scalar constants. This approximation is a valid one in situations where, for example, a high proportion of the received signal energy is due to direct-path propagation. Following from (7.16), (7.15) becomes

$$\mathbf{W}_\infty = \frac{|H_0|^2}{|\mathbf{H}|^2 + \frac{\delta^2}{P_s}} [\gamma_1 \exp(-j\omega\tau_1), \dots, \gamma_{M-1} \exp(-j\omega\tau_{M-1})]^T \quad (7.17)$$

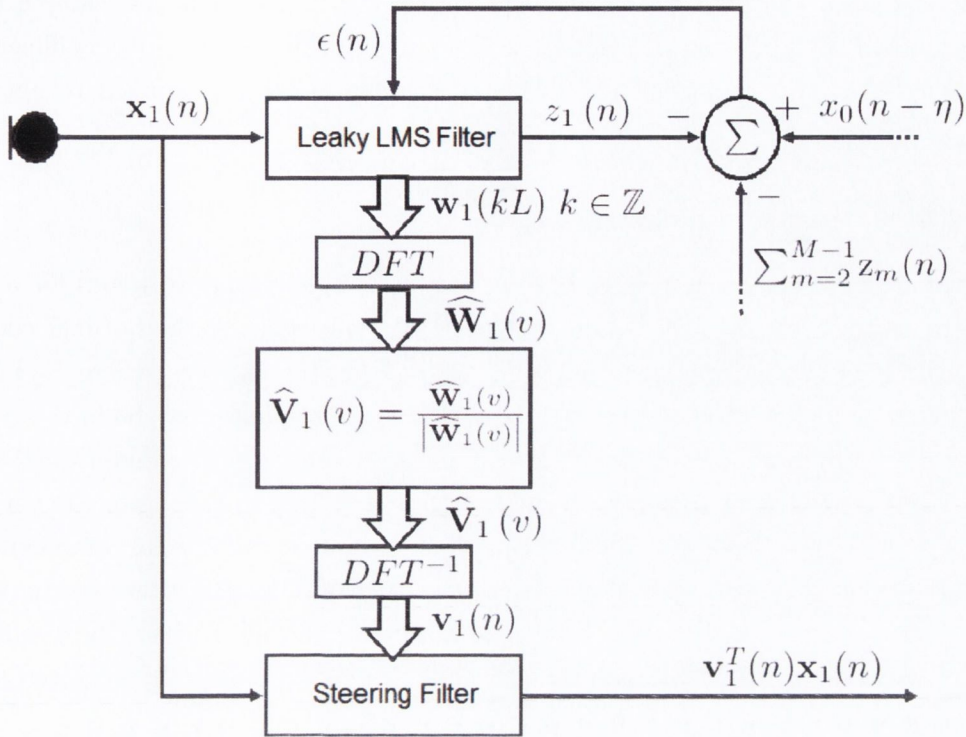


Figure 7.1: A block diagram of an implementation of the leaky-LMS-based method for beamformer steering. The elements shown correspond to the processing undergone by the output of m_1 . Identical processing is applied to the outputs of the remaining microphones.

The individual elements of the \mathbf{V} vector, V_m , are found by flattening the frequency responses of the elements of \mathbf{W}_∞ .

$$V_m = \frac{W_{\infty m}}{|W_{\infty m}|} \quad (7.18)$$

$$\mathbf{V} = [1, V_1, V_2, \dots, V_{M-1}]^T \quad (7.19)$$

Thus, we obtain a vector containing phase-shifts that will correctly steer the array.

$$\mathbf{V} = [1, \exp(-j\omega\tau_1), \dots, \exp(-j\omega\tau_{M-1})]^T \quad (7.20)$$

7.2.4 Implementation

Figure (7.1) shows a block diagram of a practical implementation of the leaky-LMS-based approach to beamformer steering. An FIR filter processes the output of each

microphone. The filter-tap coefficients are updated with each new sample using a leaky-LMS filter. Every L samples, a DFT is applied to the filter coefficients and spectral flattening is performed. Finally, an inverse DFT is applied to obtain the time-domain coefficients of the steering filter, $\mathbf{v}_m(n)$.

7.2.5 Parameter Selection

We now address the question of how we might select appropriate values for α and μ . From inspection of (7.15) we see that the m^{th} component of the optimal converged weightvector, \mathbf{W}_∞ , is proportional to $H_m H_0^* = \mathcal{F}\{h_m(t) * h_0(-t)\}$ (where $\mathcal{F}\{\}$ denotes the Fourier transform). However, a typical impulse response may be in the region of several hundred milliseconds long. Therefore, to accurately approximate $h_m(t) * h_0(-t)$ an FIR filter must be of similar length. In practice, FIR filter lengths are very much shorter, due to constraints on the permissible level of computational complexity. As a result of the consequent disparity between the lengths of the FIR filters and the impulse responses, the weightvector in (7.15) may not be achievable. Rather, the weightvector will converge to that achievable weightvector which minimizes the expected squared error.

In early investigations it was observed that, for certain values of α and μ , the weightvector did not converge to a value consistent with that predicted by (7.15). Rather, the converged weightvector would tend to emphasize the contribution of microphones closest to the reference microphone (and hence having outputs most similar to that of the reference microphone) while suppressing the contributions of the remaining microphones. This was observed even in scenarios featuring closely-spaced microphones and a far-field source.

Due to the small contribution of some microphones to the overall system output, large errors may be tolerated in the corresponding weightvector elements. These errors are then emphasized by (7.18) leading to inaccurate steering. We therefore require the weightvector to converge to a value for which all microphones make an approximately equal contribution to the overall output. This may be achieved by the selection of appropriate values for α and μ .

Consider a scenario in which spatially uncorrelated white noise of variance δ^2 is present in the output of each microphone. Applying a weightvector to approximate X_0 we may express an error signal, ε , as follows

$$\varepsilon = \mathbf{W}^H(\mathbf{X} + \mathbf{N}) - X_0 \quad (7.21)$$

Multiplying each side by its conjugate and applying the expectation operator.

$$E \left\{ |\varepsilon|^2 \right\} = \mathbf{W}^H \mathbf{R}_{\mathbf{X}\mathbf{X}} \mathbf{W} + \mathbf{W}^H \mathbf{R}_{\mathbf{X}\mathbf{X}_0} + \mathbf{R}_{\mathbf{X}\mathbf{X}_0}^H \mathbf{W} + \delta^2 |\mathbf{W}|^2 \quad (7.22)$$

From inspection of (7.22) it is apparent that $E \left\{ |\varepsilon|^2 \right\}$ increases with $\delta^2 |\mathbf{W}|^2$. Therefore, the presence of white noise effectively constrains the norm of the weightvector. Selecting $\alpha < 1$ is, as we have shown, equivalent to injecting spatially uncorrelated, spectrally white noise into the microphone outputs. Therefore, leaky implementations of the LMS filter also constrain the norm of the weightvector. This constraint “tightens” as $\frac{(1-\alpha)}{\mu\alpha}$ (or in the case of actual noise, δ^2) increases. Therefore, by selecting sufficiently small values for α and μ , we may prevent any one component of \mathbf{W} from becoming too large. The filter responds to this by increasing the contributions of the remaining microphones until the contribution of each microphone is approximately equal.

7.3 Simulations and Experiments

7.3.1 Simulations

A series of simulations were performed to examine the performance of a beamformer steered using the proposed leaky-LMS-based technique. Our simulated environment, was a simple rectangular room with uniform surface absorption coefficient of 0.3 and dimensions $[5.25m, 6.95m, 2.44m]$, figure (7.2). 7 sources were positioned at a range of 2m around a linear, equispaced 7-element array with intersensor spacing of 0.034m. The sources were at the same height as the array and were placed at 15° angular intervals between 0° and 90°. The azimuth of the first-arriving wavefront, relative to the array is called the direction of arrival (DOA).

In this setup, the sources may be considered to be in the far-field of a narrow array. However, this is done only so that we may present our results with respect to two variables (DOA and frequency) as opposed to four (x -coordinate, y -coordinate, z -coordinate and frequency) and we note, once again, that as the proposed method for LMS-based steering requires no knowledge of the array geometry, it is equally applicable to scenarios in which the array is distributed and the source is in the near-field.

Source-microphone impulse responses were then using a “raytracing” algorithm with random reverberant tails, [117]. The direct-to-reverberant ratio was found to be approximately 5dB at each microphone in the array. The sampling frequency used was 10kHz. These impulse responses were then convolved with a Maximum-Length-Sequence (MLS) of length 3.3s to obtain simulated “recordings”. These recordings were then used to converge a weightvector with a time-domain, leaky LMS filter. The

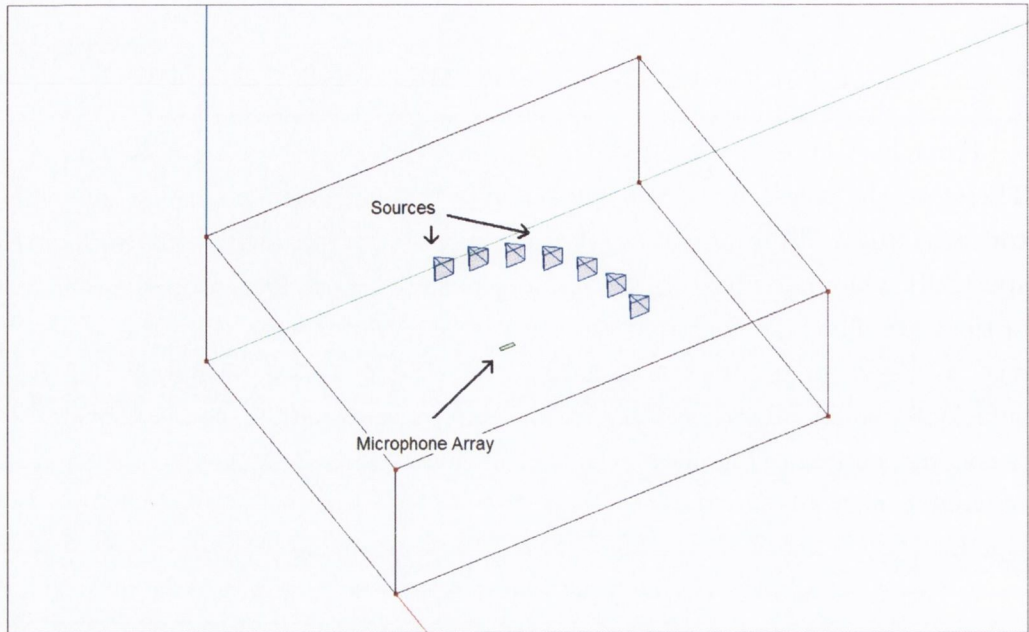


Figure 7.2: The simulated room and loudspeaker-microphone setup.

parameters used were $\mu = 2 \times 10^{-4}$, $\alpha = 0.9968$ and $L = 16$. From this weightvector, the coefficients of the steering filters were obtained as previously described.

The steering filters were then used to steer a delay-and-sum (D&S) beamformer. The performance of this beamformer was compared to that of a D&S steered to 0° (“Broadside”) and 90° (“Endfire”). We note that, given our choice of sampling rate and intersensor spacing, application of delays which are integer multiples of the sampling period allows us to steer in the Broadside and Endfire directions only. In other words, using traditional steering techniques a beamformer processing data from our array is incapable of being steered in any directions other than 0° or 90° .

The results, shown in Figures (7.3) and (7.4), clearly demonstrate that a beamformer steered using our technique outperforms a mis-steered D&S beamformer, whilst maintaining almost identical performance to that of a correctly steered D&S beamformer regardless of the DOA of the source.

This procedure was repeated using a concatenated speech sample (containing 2 male and 2 female speakers and of approximately 13s in length) in place of the MLS. It has been previously noted that when attempting to enhance speech signals, most of the degrees of freedom available to the filter go towards minimizing the error at the dominant low frequencies whilst leaving the weaker high frequencies largely un-enhanced (i.e. large errors are tolerated at high frequencies) [119]. Steering errors at high frequencies are then magnified by (7.18). To compensate for this and achieve a

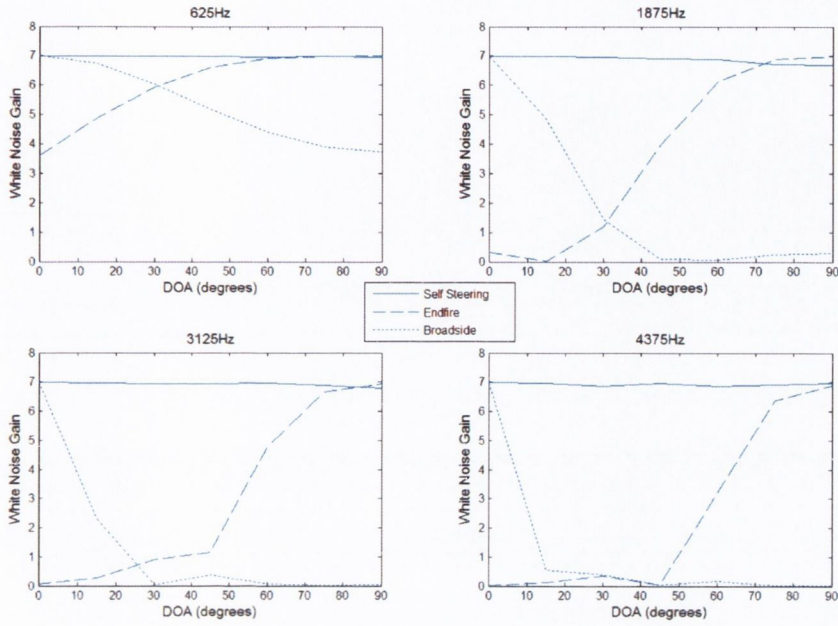


Figure 7.3: A comparison of the white noise gain achieved for varying DOAs.

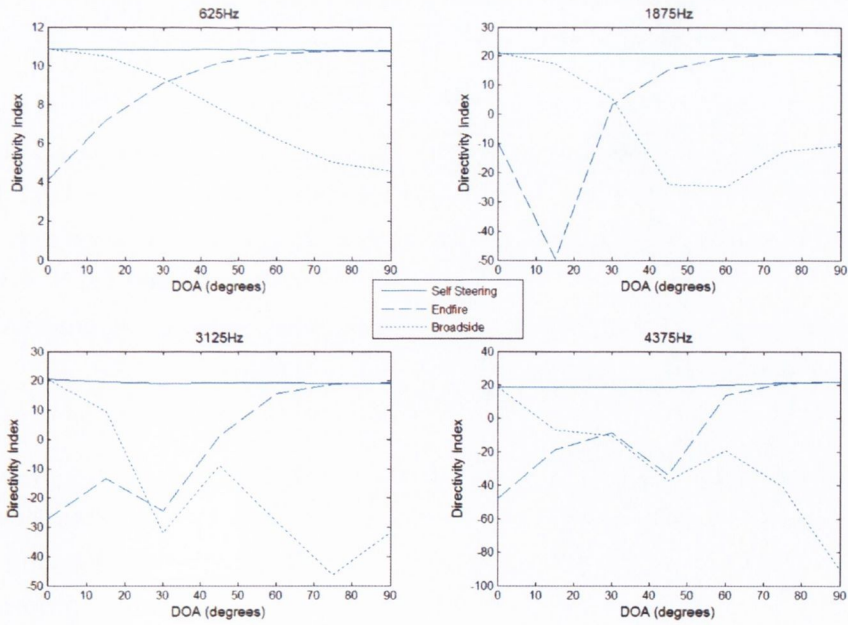


Figure 7.4: A comparison of the directivity index achieved for varying DOAs.

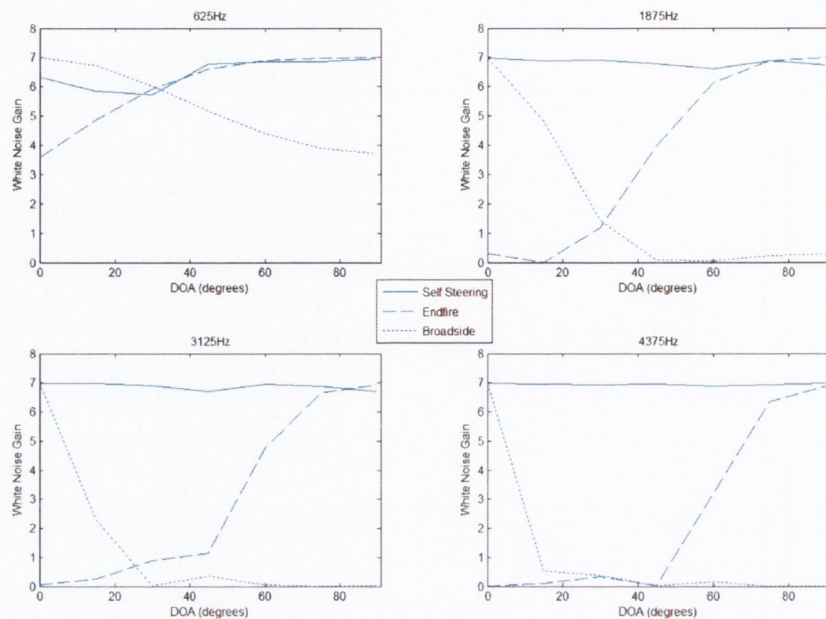


Figure 7.5: A comparison of the White Noise Gain achieved under reverberant conditions.

more even distribution of the filter's degrees of freedom across the full bandwidth, we applied a whitening filter with a frequency response equal to the square root of the inverse of the long term power spectrum of speech, as described in [119].

Using the same parameters and procedure as before, steering filters were obtained and applied to a D&S beamformer. Once again, the performance of this beamformer, with respect to $WNG(\omega)$ and $DI(\omega)$, was compared with that of a D&S beamformer steered in the Broadside and Endfire directions. The results are shown in figures (7.5) and (7.6). While not as good as that achieved using simulated MLS recordings, the performance obtained using speech recordings is, nonetheless, comparable to that of a correctly-steered D&S, regardless of the DOA of the source.

Varying Reverberant Conditions

A second series of simulations was performed to analyze the performance of the LMS-steered beamformer under different direct-to-reverberant ratios. Varying DRRs were achieved by scaling the direct-path components of the simulated impulse responses. Using the simulated voice recordings, steering filters were then obtained using the same procedure and update parameters as previously described. Once again, these steering filters were used to steer a D&S beamformer, the performance of which was

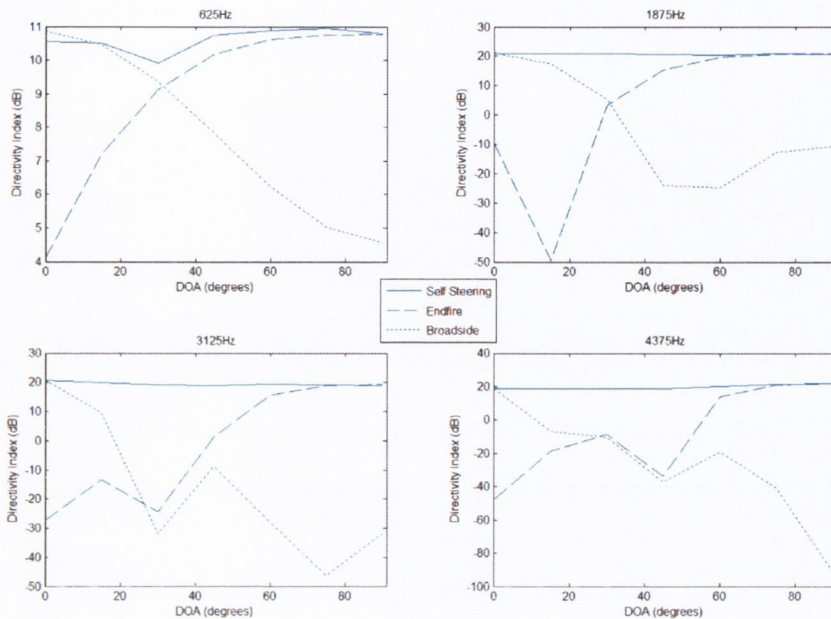


Figure 7.6: A comparison of the Directivity Index achieved under reverberant conditions.

then analyzed with respect to $WNG(\omega)$ and $DI(\omega)$. The results are shown in figures (7.7) and (7.8). From our results we see that, while performance does worsen as the DRR reduces (a result of the decreasing plausibility of the assumption in equation (7.16)), it remains (unexpectedly) good at a DRR as low as $0dB$.

This occurs as a result of our use of a linear array. For soundwaves propagating across a linear array, the relative time delay between the sound detected by microphone depends only upon the azimuth, and not the elevation, of the DOA of the soundwave. Due to the geometry of our simulated environment, the direct-path component of the signal, as well as strong 1st order reflections from the floor and ceiling, will propagate across the array with the same azimuth. As a result, a sufficiently large proportion of each source-microphone frequency response is identical subject to a phase shift corresponding to the relative intersensor time delays.

We also observe that in the case of the source at 0° , the performance is good across all DRRs. This, again, is due to room geometry. A large proportion of the reflected energy propagates across the array from the two long walls at 0° and -180° relative to the array and therefore has the same azimuth as the direct path component of the signal.

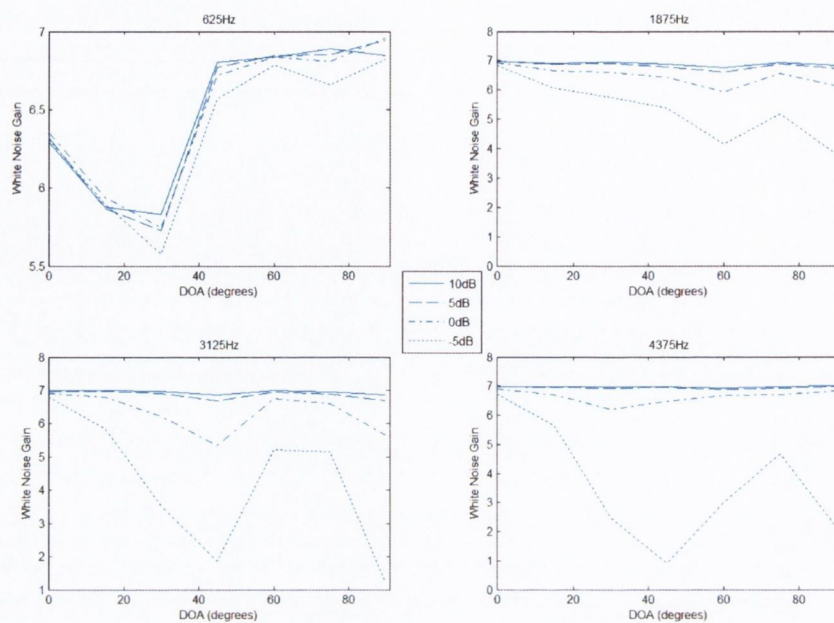


Figure 7.7: A comparison of the White Noise Gain achieved by the Self-Steering beamformer under different DRRs.

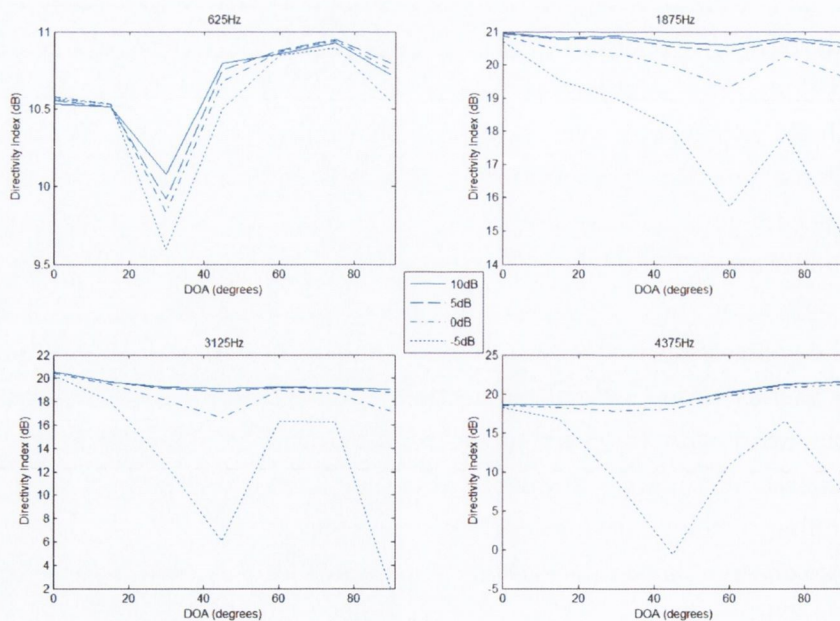


Figure 7.8: A comparison of the Directivity Index achieved by the Self-Steering beamformer under different DRRs.

Dynamic Performance

We analyzed the dynamic performance of our approach using the concatenated speech recordings and differing values of α and μ . $\alpha = 0.9992, 0.9984$ and 0.9968 . $\mu = 0.02, 0.01$ and 0.002 . The DRR was $5dB$. The results in figures (7.9-7.11) show the variation in White Noise Gain over time. The results shown are for a source at 45° but are illustrative of the results obtained for all the source locations.

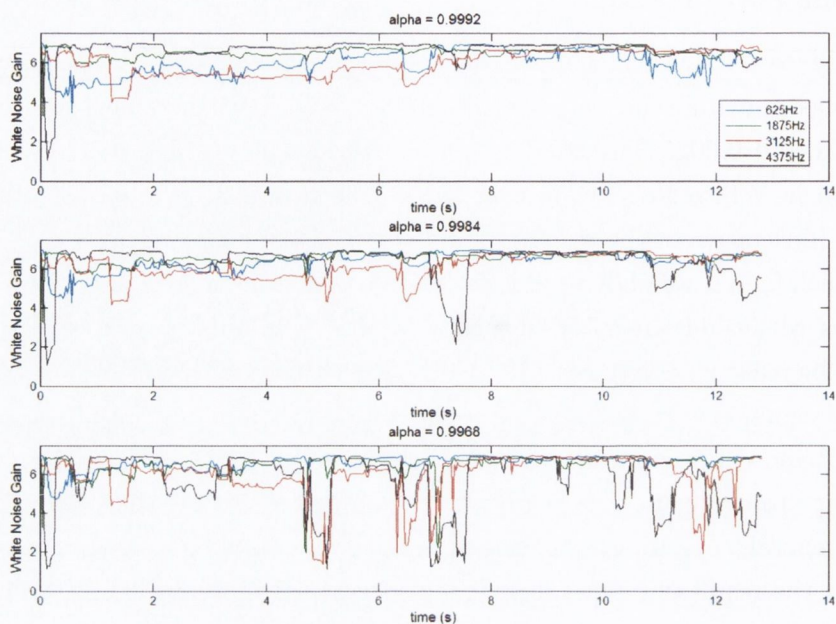
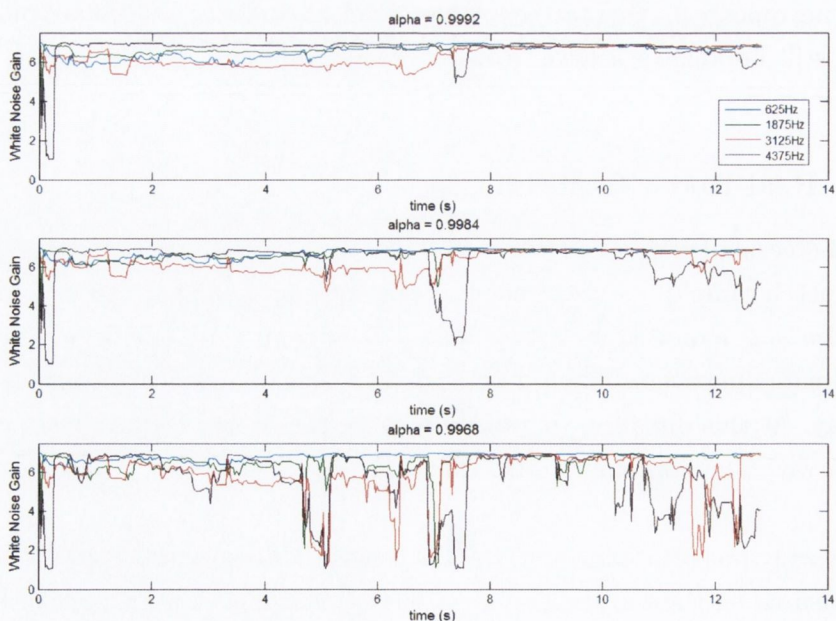
In each figure, we observe our results to be variable in time with the WNG subject to drops within one or more frequency bands. The number and severity of these is seen to increase as α reduces. These drops occur at times and frequencies where the received signal energy is low. Due to the often harmonic nature of speech, large portions of bandwidth will frequently contain low energy and we require long segments of recordings to overcome this. Comparing the figures, we see that, when $\mu = 0.01$, we obtain generally superior performance to that when $\mu = 0.02$. This we may account for as being the result of having insufficient virtual noise. From figure (7.11) we observe that initial convergence is slow for $\mu = 0.002$. This slow convergence leads to subsequent poor performance at certain times and frequencies, particularly as α reduces. We may, to a certain extent, overcome these subsequent performance problems by allowing adaptation only in bands where received energy is sufficiently high. However, our requirement for long segments of recordings would render our approach incapable of tracking quickly moving targets. Given a stationary or slow moving source, however, we will still be able to achieve satisfactory performance by selecting an appropriate value for μ .

7.3.2 Real-Room Experiments

The self-steering beamformer was tested using real recordings of an MLS and concatenated speech samples. A six-element linear, equispaced array with intersensor spacing of $0.034m$ was mounted on a theodolite, $1.3m$ from the floor, in an empty office of approximate dimensions $[3.2m, 4.2m, 2.6m]$. A single loudspeaker was placed $1m$ from the array. At this distance, the source may be considered to have been in the farfield of our array. The approximate DRR at the array was $6.7dB$. The sampling rate was $10kHz$.

The array was rotated to vary the DOA of the loudspeaker output. The exact DOA was measured on the theodolite. Recordings of voiced and MLS signals were taken for DOA's of $0^\circ, 30^\circ$ and -45° . From these recordings, steering filters were obtained as before. The parameters used were $\mu = 0.01$ and $\alpha = 0.9992$. The speech recordings were also prewhitened as before.

Determining the true delay-vector was problematic, due to the difficulty of mea-

Figure 7.9: WNG over time for varying frequencies. $\alpha = 0.02$ Figure 7.10: WNG over time for varying frequencies. $\alpha = 0.01$

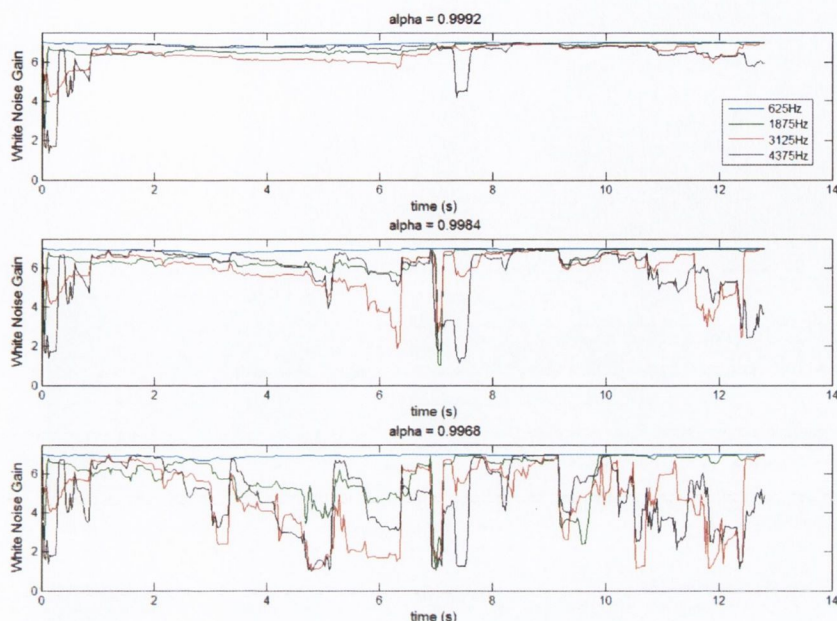


Figure 7.11: WNG over time for varying frequencies. $\alpha = 0.002$

suring the exact distance between the source and the microphones. As a result we do not use $WNG(\omega)$ or $DI(\omega)$ in our analysis of the beamformers' performance. Instead, in figure(7.12), we show the array patterns (i.e. gain with respect to frequency and DOA) obtained using each set of recordings.

From inspection of the array patterns obtained, we see that the resolution offered at low frequencies is very poor. This is as a result of the narrow array-width and we can expect the resolution to improve as this width increases. Low-frequency resolution notwithstanding, the mainlobe of the array pattern is pointed/steered toward the appropriate DOA in each case. This confirms that the self-steering beamformer we have presented does, indeed, work in real environments for both white and voiced sources.

7.4 Computational Complexity

In this section we compare the computational complexity of the widely- used PHAT-GCC method for time delay estimation (section 5.2.2) to that of the leaky-LMS-based approach presented in this chapter. To do this we make a number of simplifying assumptions. In the case of the leaky-LMS-based approach, we consider only those calculations required to obtain the coefficients of the steering filters. In the case of

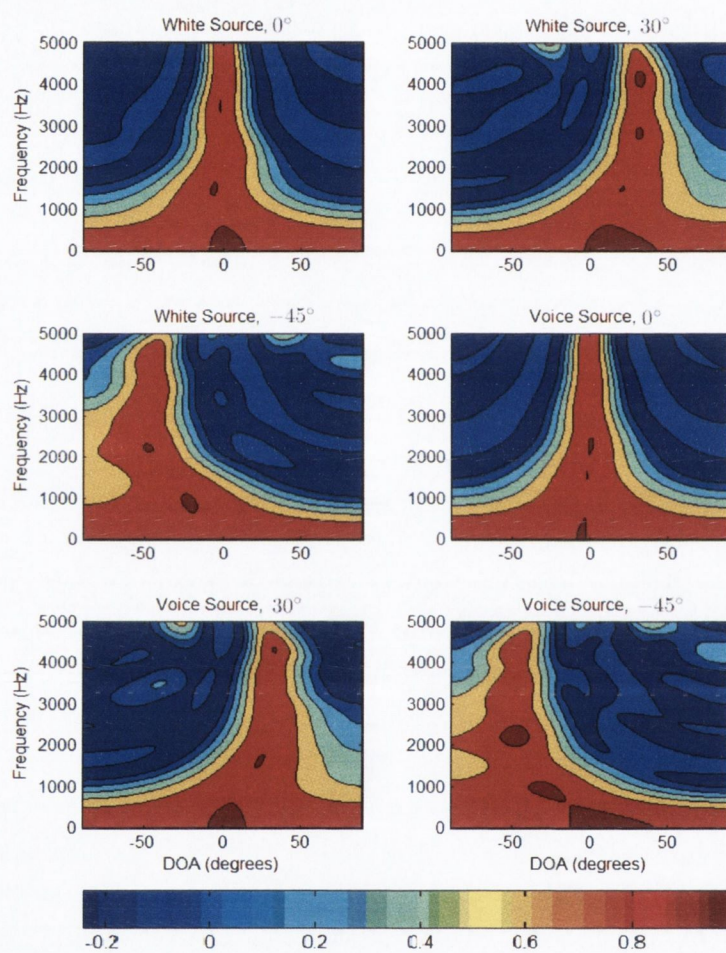


Figure 7.12: The array patterns corresponding to the self-steering beamformer, obtained from real-room recordings.

Leaky LMS	Real-valued Multiplications
Compute $e(n)$ L times per segment	
$e(n) = [x_0(n - \eta) - \mathbf{w}^T(n) \mathbf{x}(n - 1)]$	$(M - 1)L^2$
Update Filter coefficients L times per segment	
$\mathbf{w}(n + 1) = \alpha \mathbf{w}(n) + \mu e(n) \mathbf{x}(n - 1)$	$(M - 1)(2L^2 + L)$
Estimate the spectrum of $\mathbf{w}_m(n + 1)$ for each of $M - 1$ microphones	
$\widehat{W}_m(v) = FFT_L\{\mathbf{w}_m(n + 1)\} \quad m = 1 : M - 1$	$(M - 1) \left(\frac{L}{2} \log_2(L') - \frac{5L}{4} \right)$
Normalize (flatten) $\widehat{W}_m(v)$:	$2L(M - 1)$
Apply inverse DFT to obtain each of $M - 1$ steering filter weights.	
$\widehat{W}_m(v) = FFT_L^{-1} \left\{ \frac{\widehat{W}_m(v)}{ \widehat{W}_m(v) } \right\} \quad m = 1 : M - 1$	$(M - 1) [2L \log_2 L - 7L + 12]$
Total per segment :	$(M - 1) \left[3L^2 - \frac{21L}{4} + 12 + \frac{5L}{2} (\log_2 L) \right]$

Algorithm 1 Computational Complexity for the leaky-LMS-based method. $FFT_L\{*\}$ and $FFT_L^{-1}\{*\}$ denote the L -point fast Fourier transform and inverse fast Fourier transform respectively

the PHAT-GCC, we consider only those required to obtain the cross-correlation function ($\psi_{PHAT}(n)$ - see section 5.2.2) for each of $M - 1$ microphone pairs. Ancillary tasks, such as determining the maxima of the cross correlation functions, are ignored. Furthermore, we analyze computational complexity in terms of real-valued multiplications/divisions only. The comparatively trivial operations of subtraction and addition are ignored while a complex-valued multiplication/division is treated as four real-valued multiplications/divisions. Where a DFT or IDFT is required, we assume the use of the fast Fourier transform and inverse fast Fourier transform techniques described in [118].

Algorithm (1) details the calculations required to implement a leaky-LMS-based approach in which $\mathbf{w}(n)$ is updated every sample, while the coefficients of the steering filters are updated only once every L samples. Algorithm (2) outlines the computational requirements of a PHAT-GCC in which $\psi_{PHAT}(n)$ is updated for each of $M - 1$ microphone pairs once for every (non-overlapping) segment of data. It is assumed that each such segment spans the same duration in time as L consecutive samples as

PHAT-GCC	Real-valued Multiplications
Estimate the spectrum of $\mathbf{x}_m(n)$ for each of M microphones	Real-valued Multiplications
$X(v) = FFT_{L'}\{\mathbf{x}_m(n)\} \quad m = 0 : M - 1$	$M \left(\frac{L'}{2} \log_2(L') - \frac{5L'}{4} \right)$
Compute the PHAT-weighted cross spectrum for each of $M - 1$ microphone pairs	
$\hat{\Psi}_{PHAT,m}(v) = \frac{X_0(v)X_m^*(v)}{ X_0(v)X_m^*(v) } \quad m = 1 : M - 1$	$(M - 1)(4L' + 8)$
Calculate $\hat{\psi}_{PHAT}(n)$ for each of $M - 1$ microphone pairs	
$\hat{\psi}_{PHAT,m}(n) = FFT_{L'}^{-1}\{\hat{\Psi}_{PHAT,m}(v)\} \quad m = 1 : M - 1$	$(M - 1)(2L' \log_2 L' - 7L' + 12)$
Total per segment :	$(M - 1) \left[\frac{5L'}{2} \log_2 L' - \frac{17L'}{4} + 20 \right] + \frac{L'}{2} \log_2(L') - \frac{5L'}{4}$
 Algorithm 2 Computational Complexity for the PHAT-GCC time-delay estimation method. $FFT_L\{\ast\}$ and $FFT_L^{-1}\{\ast\}$ denote the L -point fast Fourier transform and inverse fast Fourier transform respectively	

used by the leaky-LMS-based method. However, to account for a possible difference in sampling rates, we denote the number of samples in each segment as L' . Comparing the total number of calculations per segment required of each approach, it is apparent that the LMS-based method is more efficient than the PHAT-GCC technique if the following holds,

$$3L^2 - \frac{21L}{4} + 12 + \frac{5L}{2} (\log_2 L) < \frac{5L'}{2} \log_2 L' - \frac{17L'}{4} + 20 + \left(\frac{1}{M-1} \right) \left(\frac{L'}{2} \log_2(L') - \frac{5L'}{4} \right) \quad (7.23)$$

As previously discussed, accurate time delay estimation (and hence steering) using the PHAT-GCC approach requires the use of very high sampling rates. As we have seen, however, the LMS-based approach allows us to achieve precise steering while sampling at comparatively low rates. Let us assume, therefore, that the PHAT-GCC requires data sampled at γ times that required by the LMS-based method. Replacing L' in (7.23) with γL and performing some simple algebraic manipulation yields an expression for the conditions under which the leaky-LMS-based method is less computationally complex than the PHAT-GCC approach.

$\log_2 L$	$M = 2$	$M = 6$	$M = 10$	$M = 14$
4	4	5	5	5
5	5	7	7	7
6	8	10	10	11

Table 7.1: Integer multiples of the Nyquist sampling rate for which the proposed leaky-LMS-based steering approach is less computationally complex than steering with time-delays determined using the PHAT-GCC method

$$\left(\frac{5\gamma L}{2} + \frac{\gamma L}{M-1}\right) (\log_2 \gamma + \log_2 L) - \frac{5L}{2} \log_2 L + 8 + \frac{21L(M-1) + 5\gamma L - 17\gamma L(M-1)}{4(M-1)} - 3L^2 > 0 \tag{7.24}$$

Table (7.1) shows the smallest integer value of γ for which (7.24) holds, for likely values of L and M .

7.5 Discussion

In this chapter we have proposed a method for beamformer steering based upon a multichannel leaky LMS filter. Furthermore, we have verified experimentally that the proposed approach can achieve precise steering even where the true inter-sensor time-delays are non-integer multiples of the sampling rate.

In section 7.4, we compared the computational complexity of the proposed technique with that of a conventional approach using PHAT-GCC time delay estimation. As was discussed in section 6.3.1, conventional steering techniques require the use of higher sampling rates when the array geometry is unknown and/or the source is potentially in the near-field.

For the array geometry used in the simulations in section 7.3.1 of this chapter, sampling at γ times the Nyquist rate allows us to steer towards $\gamma+1$ DOAs in the range $0^\circ : 90^\circ$ (see section 3.4.3, equation(3.31)). There are 7 sources and so we must sample in time at a minimum of 6 times the Nyquist rate if we are to be able to discriminate between them using conventional steering techniques. From inspection of table (7.1), however, it is apparent that the proposed method is a more computationally efficient approach if conventional steering techniques require us to sample at 5-or-more times the Nyquist rate.

Therefore, in many scenarios in which we are required to steer a beamformer (most notably, those in which the array geometry is unknown or the source in the near-field), leaky-LMS-based steering should be preferred as being more accurate and less computationally complex than conventional methods.

It is worth noting, once again, that blind source separation (BSS: see section 4.3.4) is a field of research concerned with methods for separating multiple simultaneous and statistically independent sources, using microphone arrays of unknown geometry. These techniques could be applied to scenarios in which there is a source or sources of interest in the presence of one or more local interferers. It should be remembered however that reverberation severely limits the effectiveness of these methods and this would render them unsuitable for the scenarios under investigation here. Nonetheless, recent work (notably [46] and [48]) has presented BSS techniques for use in modestly reverberant environments and future developments may yet see BSS techniques successfully applied in more typically-reverberant environments.

Chapter 8

Range Estimation

8.1 Introduction

In this chapter, we derive and demonstrate a novel method for estimating source-microphone ranges, which may be implemented using arrays of unknown geometry and is robust against error due to reverberation. We refer to this method, which we derive in section 8.3, as the “Range-Finder” algorithm. For comparison, we also derive a well-known range estimator that assumes an anechoic environment and a range-estimating variant of the steered-response-power method for source localization. In section 8.4, we analyze the effects of microphone geometry and relative source location upon the accuracy of the range estimation techniques. In section 8.5, we present the results of experiments, using real and simulated data, to demonstrate the efficacy of the proposed method. We discuss the potential uses of the Range-Finder algorithm and suggest future work in section 8.6.

8.2 Sound Propagation in Reverberant Environments

In a noiseless but reverberant environment the signal received at some microphone, m_0 , will consist of a direct-path component and multiple reflected components jointly referred to as reverberation. The output of the microphone may be modelled as the convolution of the source-microphone impulse response, $h_0(t)$, and the source signal, $s(t)$.

$$x_0(t) = \int_0^t s(p)h_0(t-p)dp \quad (8.1)$$

In the frequency domain

$$\begin{aligned} X_0(\omega) &= S(\omega) H_0(\omega) \\ &= S(\omega) (H_{dp_0}(\omega) + H_{mp_0}(\omega)) \end{aligned} \quad (8.2)$$

where $H_{dp_0}(\omega)$ is the component of $H_0(\omega)$ due to direct-path (non-reflected) propagation and $H_{mp_0}(\omega)$ is the reverberant component due to multipath reflections. The received signal power spectrum may be calculated as follows.

$$\begin{aligned} |X_0(\omega)|^2 &= |S(\omega)|^2 |H_0(\omega)|^2 \\ &= |S(\omega)|^2 (|H_{dp_0}(\omega)|^2 + |H_{mp_0}(\omega)|^2 + 2 \operatorname{Re}\{H_{dp_0}(\omega) H_{mp_0}^*(\omega)\}) \end{aligned} \quad (8.3)$$

where $\operatorname{Re}\{ \}$ denotes the real component and $*$ denotes the complex conjugate.

In air, for an omnidirectional source and receiver, the power of the direct-path component of sound, received at m_0 , is inversely proportional to the source-microphone distance, squared.

$$|H_{dp_0}(\omega)|^2 \propto \frac{1}{r_0^2} \quad (8.4)$$

where $r_0 = |\vec{s} - \vec{m}_0|$ and \vec{s} and \vec{m}_0 denote the Cartesian coordinates of the source and m_0 respectively. The direct-path component decays at a rate of 6dB per doubling of distance. This model does not address effects due to variations of air pressure or temperature, however, in a room environment it is reasonable to assume a homogenous medium. From (8.4), we may derive an expression for the power of the direct-path component of the sound received at some microphone m_a .

$$|H_{dp_a}(\omega)|^2 = |H_{dp_0}(\omega)|^2 \left[\left(\frac{r_0}{r_a} \right)^2 \right] \quad (8.5)$$

We define the following, noting that, for clarity, we omit the frequency index, ω , in the sequel.

$$F_{a,b} = \int |H_{mp_a}|^2 - |H_{mp_b}|^2 + 2 \operatorname{Re}\{H_{dp_a} H_{mp_a}^* - H_{dp_b} H_{mp_b}^*\} d\omega \quad (8.6)$$

where the a and b subscripts denote the impulse response components corresponding to the microphones m_a and m_b , respectively. Consider the cross-terms in (8.6). Direct path propagation applies a delay and scaling to a soundwave. Therefore, for any source-microphone impulse response, H_{dp} is a scaled exponential. Similarly, H_{mp} may be considered to be the sum of scaled exponentials corresponding to multiple reflected soundwaves. As such, $H_{dp} H_{mp}^*$ is also the sum of multiple scaled exponentials.

Therefore, invoking the central limit theorem, we shall assume $\int \text{Re}\{H_{dp_a}H_{mp_a}^*\}d\omega$ and $\int \text{Re}\{H_{dp_b}H_{mp_b}^*\}d\omega$ to be zero-mean normally-distributed random variables. Following from the results of our analysis in section 2.3.4, we also assume $\int |H_{mp_a}|^2 d\omega$ and $\int |H_{mp_b}|^2 d\omega$ to be random variables distributed about the same mean. Therefore, invoking the central limit theorem once again, we may consider $F_{a,b}$ to be a zero-mean normally-distributed random variable.

Note that, if H_{dp} and H_{mp} are non-zero at $\omega = 0$, $\int \text{Re}\{H_{dp}H_{mp}^*\}d\omega$ will exhibit a positive bias. We may ignore this however, as the frequency responses of real microphones will not have a non-zero component at $\omega = 0$.

8.3 Range Estimation

In this section we derive three range estimation algorithms: a well-known but naïve range estimator that assumes an anechoic environment, a modification of the well-known steered-response-power technique for source localization and a novel algorithm, which we present in this thesis and which we refer to as the “Range-Finder”.

8.3.1 A Naïve Range Estimator

When τ_a is the relative intersensor time-delay between m_a and m_0 .

$$r_a - r_0 = c\tau_a \quad (8.7)$$

where c is the speed of sound in air. Using any one of a variety of time-delay estimation techniques, we may obtain an estimate of the relative intersensor time-delay, $\tilde{\tau}_a$. In noiseless, anechoic environments the direct-path sound accounts for all acoustic energy received by the microphones and so, by substituting (8.7) into (8.5) and performing algebraic manipulation, we obtain a simple and well known estimator of r_0 .

$$\tilde{r}_0 = \frac{c\tilde{\tau}_a \sqrt{\frac{|H_a|^2}{|H_0|^2}}}{1 - \sqrt{\frac{|H_a|^2}{|H_0|^2}}} \quad (8.8)$$

Unfortunately, in non-ideal acoustic environments, the presence of interfering reverberation can severely distort this estimate, making the above range estimator unsuitable for practical environments. Where more than two microphones are available, the most accurate range estimate will be obtained by using only those two microphones closest to the source. These may be presumed to have the highest DRRs. The outputs of the remaining microphones will contain proportionally greater levels of reverberation and will, therefore lead to greater distortion in the range estimates.

8.3.2 The Steered-Response-Power Range Estimator

Steered-response-power (SRP) techniques are a classical method for source localization, whereby a beamformer is steered to a series of candidate locations with the source location estimate taken as that which maximizes the power of the output of the beamformer. Under conditions of diffuse reverberation, the optimal beamformer for such a task would be the superdirective beamformer (section 4.3.1). Unfortunately, calculating the filter weights for such a beamformer requires knowledge of the microphone array geometry, which we will not have. Therefore, we use the delay-and-sum beamformer (section 4.2). The resulting source location estimate may be expressed as follows

$$\vec{s} = \arg \max_{\vec{s}} \left\{ \int |\mathbf{D}^H(\vec{s})\mathbf{X}|^2 d\omega \right\} \quad (8.9)$$

Replacing the actual intersensor time delays in (3.59) with TDEs and replacing $|\vec{s} - \vec{m}_m|$ with $r_0 + c\tilde{\tau}_m$, we may define $\tilde{\mathbf{D}}(r_0)$ as shown below.

$$\tilde{\mathbf{D}}(r_0) = \left[\frac{a(r_0)}{r_0}, \frac{a(r_0)}{r_0 + c\tilde{\tau}_1} \exp\{-j\omega\tilde{\tau}_1\}, \dots, \frac{a(r_0)}{r_0 + c\tilde{\tau}_{M-1}} \exp\{-j\omega\tilde{\tau}_{M-1}\} \right]^T \quad (8.10)$$

where $a(r_0)$ is some scalar such that the norm of $\tilde{\mathbf{D}}(\omega, r_0)$ is unity. A range estimate may be obtained from

$$\tilde{r}_0 = \arg \max_{r_0} \left\{ \int |\tilde{\mathbf{D}}^H(r_0)\mathbf{X}|^2 d\omega \right\} \quad (8.11)$$

SRP source localization techniques are noted in the literature as being robust against the effects of reverberation (section 5.4.2). However, such assertions are based upon the results of experiments using closely spaced arrays and far-field sources - i.e. where $r_0 \approx r_0 + c\tau_m$. To the author's knowledge, no previous study has examined the use of SRP methods for determining source-microphone ranges.

8.3.3 The Range-Finder Algorithm

From (8.5) and (8.7)

$$(|H_{dp_a}|^2 - |H_{dp_b}|^2) = |H_{dp_o}|^2 \left[\left(\frac{r_0}{r_0 + c\tau_a} \right)^2 - \left(\frac{r_0}{r_0 + c\tau_b} \right)^2 \right] \quad (8.12)$$

The term in the square brackets is a function of r_0, τ_a and τ_b which we denote as $G_{a,b}(r_0, \tau_a, \tau_b)$

$$G_{a,b}(r_0, \tau_a, \tau_b) = \left(\frac{r_0}{r_0 + c\tau_a} \right)^2 - \left(\frac{r_0}{r_0 + c\tau_b} \right)^2 \quad (8.13)$$

Integrating (8.3) across the full bandwidth of the signal we obtain the total received signal power P_0 .

$$P_0 = \int |S|^2 (|H_{dp_0}|^2 + |H_{mp_0}|^2 + 2 \operatorname{Re}\{H_{dp_0} H_{mp_0}^*\}) d\omega \quad (8.14)$$

We define $\Lambda_{a,b}$ as being the difference between the total received signal power at m_a and m_b .

$$\Lambda_{a,b} = P_a - P_b \quad (8.15)$$

Let us assume, for the moment, that $|S|^2$ is a constant with respect to frequency (we shall return to this assumption later). Substituting (8.14) into (8.15) and performing algebraic manipulation yields

$$\Lambda_{a,b} = |S|^2 [kG_{a,b}(r_0, \tau_a, \tau_b) + F_{a,b}] \quad (8.16)$$

where $k = \int |H_{dp_0}|^2 d\omega$. From (8.16), we see that the difference between the signal power received at two microphones is proportional to the sum of a scaled, deterministic function, $G_{a,b}(r_0, \tau_a, \tau_b)$, and a zero-mean and normally distributed random variable, $F_{a,b}$. We define the following vectors, noting that we have omitted the arguments of the $G_{a,b}(r_0, \tau_a, \tau_b)$ terms for clarity.

$$\mathbf{G} = [G_{0,1}, G_{0,2}, \dots, G_{1,2}, G_{1,3}, \dots, G_{M-2, M-1}]^T \quad (8.17)$$

$$\mathbf{F} = [F_{0,1}, F_{0,2}, \dots, F_{1,2}, F_{1,3}, \dots, F_{M-2, M-1}]^T \quad (8.18)$$

$$\begin{aligned} \mathbf{\Lambda} &= [\Lambda_{0,1}, \Lambda_{0,2}, \dots, \Lambda_{1,2}, \Lambda_{1,3}, \dots, \Lambda_{M-2, M-1}]^T \\ &= |S|^2 [k\mathbf{G} + \mathbf{F}] \end{aligned} \quad (8.19)$$

Once again, using any of the many well-known techniques for delay-vector estimation, we may obtain the time-delay estimates $\tilde{\tau}_a$ and $\tilde{\tau}_b$. We then define $\tilde{G}_{a,b}(r_0)$ and the corresponding vector $\tilde{\mathbf{G}}(r_0)$ from

$$\tilde{G}_{a,b}(r_0) = G_{a,b}(r_0, \tilde{\tau}_a, \tilde{\tau}_b) \quad (8.20)$$

Following from the Cauchy-Schwartz inequality, the optimal range estimate, \tilde{r}_0 , is obtained by a matched-filtering of the power-difference-vector, $\mathbf{\Lambda}$, with $\frac{\tilde{\mathbf{G}}(r_0)}{|\tilde{\mathbf{G}}(r_0)|}$.

$$\tilde{r}_0 = \arg \max_{r_0} \left[\frac{1}{|\tilde{\mathbf{G}}(r_0)|} \tilde{\mathbf{G}}(r_0)^T \mathbf{\Lambda} \right] \quad (8.21)$$

Following from this estimate, we may easily obtain estimates of the distance between the source and the remaining microphones by inserting \tilde{r}_0 and the TDEs used to calculate $\tilde{\mathbf{G}}(r_0)$ into (8.7).

Previously, we assumed $|S(\omega)|^2$ to be a constant with respect to frequency. In many cases, including that of human speech, this is unrealistic. In reality, speech is both a lowpass and often harmonic signal. This poses particular problems. We have assumed $F_{a,b}$ to be a zero-mean, normal random variable. The analysis and experimental evidence underpinning this assumption are for broadband signals and we cannot reasonably expect it to hold for cases, such as speech, where the bulk of the energy is concentrated at low frequencies.

This problem was overcome as follows. The microphone outputs are split into individual, non-overlapping subbands. The bandwidth of these subbands are chosen such that they are narrow enough that $|S(\omega)|^2$ is roughly constant within the subband whilst also being wide enough that there is always a direct-path speech component present. $\mathbf{\Lambda}$ is then calculated for each subband. Each $\mathbf{\Lambda}$ is normalized and, from these, an average power-difference-vector, $\bar{\mathbf{\Lambda}}$, found across all the subbands. The range estimate is found, as in (8.21) by a matched-filtering of $\bar{\mathbf{\Lambda}}$ with $\tilde{\mathbf{G}}(r_0)/|\tilde{\mathbf{G}}(r_0)|$.

8.4 Estimate Distribution and Accuracy

Given multiple estimates for range, we might expect that, as the number of estimates increases, their mean will approach the true range. As we shall see in the following section, this is not necessarily the case. We shall also show how the accuracy of a range estimate is dependant upon the actual source-microphone ranges. We restrict our analysis to the situation where we have three microphones only - the minimum number required for the Range-Finder algorithm. We do this both for the sake of simplicity and to allow us to employ an alternative formulation of the Range-Finder algorithm. This alternative formulation more clearly illustrates how the distribution of range estimates is related to the distribution of the ratio of normal random variables, a well-understood, albeit non-trivial, distribution that has received extensive study in the literature.

8.4.1 An Alternative Formulation of the Range-Finder

The range estimate, \tilde{r}_0 , is that which maximizes the expression in (8.21). For two vectors with given norms, the dot product of the vectors is a maximum when they are proportional. Therefore, we may write $\tilde{\mathbf{G}}(\tilde{r}_0) \propto \mathbf{\Lambda}$. For the three-microphone case, this implies

$$[\tilde{G}_{0,1}(\tilde{r}_0), \tilde{G}_{0,2}(\tilde{r}_0)] \propto [\Lambda_{0,1}, \Lambda_{0,2}] \quad (8.22)$$

Using an equivalent expression, we define $Q_{0,1,2}$

$$\frac{\tilde{G}_{0,1}(\tilde{r}_0)}{\tilde{G}_{0,2}(\tilde{r}_0)} = \frac{\Lambda_{0,1}}{\Lambda_{0,2}} = Q_{0,1,2} \quad (8.23)$$

and, from this, we obtain an alternative expression for the Range-Finder.

$$\tilde{r}_0 = \arg \min_{r_0} \left[\left| Q_{0,1,2} - \frac{\tilde{G}_{0,1}(r_0)}{\tilde{G}_{0,2}(r_0)} \right| \right] \quad (8.24)$$

For 3 microphones there are, of course, 5 further permutations of Q ($Q_{0,2,1}$, $Q_{1,2,0}$ etc.). However all may be shown to yield identical range estimates and so we shall consider only $Q_{0,1,2}$. Furthermore, to simplify our analysis, we shall assume that $0 \leq \tau_1 \leq \tau_2$. We note that this relationship is for simplicity only and is not an absolute requirement. Rather, it is merely a result of the arbitrary way in which we assign labels to the microphones. Once again, omitting the arguments of the $G_{a,b}(r_0, \tau_a, \tau_b)$ terms for clarity.

$$Q_{0,1,2} = \frac{G_{0,1} + (F_{0,1})/k}{G_{0,2} + (F_{0,2})/k} \quad (8.25)$$

From (8.25), we see that $Q_{0,1,2}$ is the ratio of normally distributed and correlated random variables, with unknown variances and means of $G_{0,1}$ and $G_{0,2}$ respectively. Such a ratio is itself a Cauchy distributed random variable.

8.4.2 Cauchy Distribution

In [120] it is shown that, following a translation and a change of scale, $Q_{0,1,2}$ has the same distribution as the ratio of two uncorrelated normal random variables of unity variance, $\frac{N(\alpha,1)}{N(\beta,1)}$. The real constants α and β may be calculated as follows

$$\alpha = \pm \frac{\frac{G_{0,1}}{\sigma_{0,1}} - \frac{\rho G_{0,2}}{\sigma_{0,2}}}{\sqrt{1 - \rho^2}}, \quad \beta = \frac{G_{0,2}}{\sigma_{0,2}} \quad (8.26)$$

where $\sigma_{a,b}$ is the standard deviation of $(F_{a,b})/k$, ρ is the correlation between $\Lambda_{0,1}$ and $\Lambda_{0,2}$ (which may be shown to be 0.5) and the sign of α is chosen to be the same as

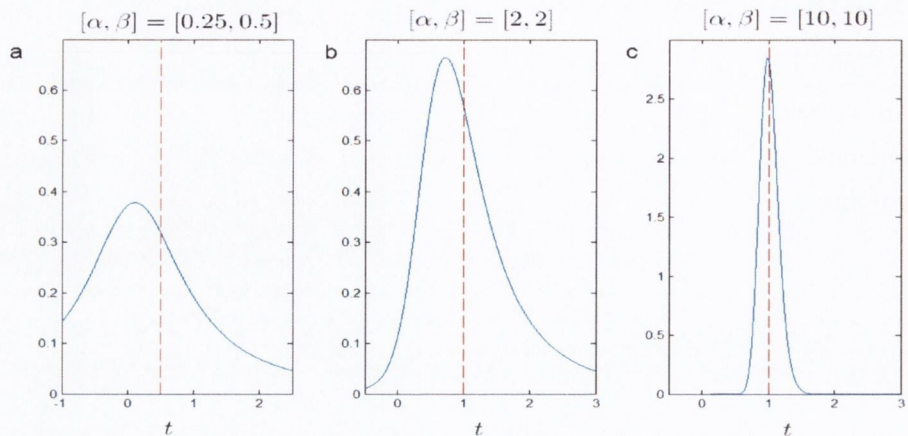


Figure 8.1: Portions of the PDFs of $\frac{N(\alpha, 1)}{N(\beta, 1)}$. Also shown is $\frac{\alpha}{\beta}$ (dashed line).

that of β . For the sake of simplicity and to avoid unwieldy equations, the following discussion shall be with reference to the simplified standard form $\frac{N(\alpha, 1)}{N(\beta, 1)}$. From [121], the probability density function (PDF), $p(t)$, of $\frac{N(\alpha, 1)}{N(\beta, 1)}$ may be given as shown below

$$p(t) = \frac{\exp\{-0.5(\alpha^2 + \beta^2)\}}{\pi(1 + t^2)} \left[1 + q \exp\{0.5q^2\} \int_0^q \exp\{-0.5x^2\} dx \right], \quad q = \frac{\beta + \alpha t}{\sqrt{1 + t^2}} \quad (8.27)$$

Figure (8.1) shows the PDFs for varying values of α and β . A very wide variety of distribution-shapes are possible and the ones shown are chosen for specific illustrative purposes. For a more complete selection of graphs see [121]. Shown also is $\frac{\alpha}{\beta}$ (dashed line). In figures (8.1a,b) the distribution is not symmetric about $\frac{\alpha}{\beta}$. This indicates that, contrary to what we might expect, the “mean” of $\frac{N(\alpha, 1)}{N(\beta, 1)}$ is biased away from $\frac{\alpha}{\beta}$. In fact, strictly speaking, the mean and variance of $\frac{N(\alpha, 1)}{N(\beta, 1)}$ do not exist. This is because $\frac{N(\alpha, 1)}{N(\beta, 1)}$ is undefined when the denominator equals zero.

In practice, we may calculate a pseudo-mean and pseudo-variance by considering only those estimates that fall within certain bounds. A natural bound would be that value of $Q_{0,1,2}$ corresponding to a range estimate of $0m$. In setting such bounds, however, we should be mindful that the consequent truncation of the PDF may introduce an additional bias in the pseudo-mean.

In general, when defined within sufficiently wide bounds, the pseudo-mean tends towards $\frac{\alpha}{\beta}$ for $|\alpha|, |\beta| \gg 1$, as occurs when $G_{0,b} \gg \sigma_{0,b}$. Furthermore, under these conditions, $Q_{0,1,2}$ tends to have quite a narrow distribution (see figure (8.1c)). Un-

fortunately, the converse is also the case. The problem is further compounded by the difficulty inherent in defining "sufficiently wide bounds" when $Q_{0,1,2}$ is widely distributed. Without knowing $\sigma_{0,1}$ or $\sigma_{0,2}$, we cannot calculate/estimate the bias that is present. We can, however, identify certain situations in which it is likely to be very large. Consider the case where $r_0 \gg c\tau_b$ - that is when the array is remote from the source. From inspection of (8.13), we see that, under these conditions, $G_{0,b} \rightarrow 0$ causing our estimates to be subject to what is likely to be a large bias.

8.4.3 The Effect of Array Geometry

The actual source-microphone ranges determine the values of r_0, τ_1 and τ_2 . We have seen how these parameters can effect the distribution of $Q_{0,1,2}$ and bias its pseudo-mean away from $\frac{G_{0,1}}{G_{0,2}}$. In this respect, therefore, the accuracy with which we may estimate range is determined by the array geometry. Array geometry also determines the extent to which a bias/error in $Q_{0,1,2}$ translates into an error in the corresponding range estimate. To investigate this second effect of array geometry, we examine how a fixed bias, ξ , translates into an error in the range estimate.

Consider an estimate, \tilde{r}_0 , of the true range, r_0 , and let us assume that this estimate contains some error, ϵ_0 .

$$\frac{\tilde{G}_{0,1}(\tilde{r}_0)}{\tilde{G}_{0,2}(\tilde{r}_0)} = Q_{0,1,2} = \frac{G_{0,1}}{G_{0,2}} + \xi \quad (8.28)$$

As an illustrative example, we plot $\frac{G_{0,1}}{G_{0,2}}$ against r_0 for $[c\tau_1, c\tau_2] = [1m, 5m]$ in figure (8.2). Outside of a small region around $r_0 = 0$, as r_0 increases the slope of the graph reduces and ϵ_0 becomes larger.

Figure (8.3), showing $\left| \frac{d}{dr_0} \left(\frac{G_{0,1}}{G_{0,2}} \right) \right|$ with respect to $\frac{c\tau_1}{r_0}$ and $\frac{c\tau_2}{r_0}$, provides a more complete description of how array geometry affects estimate accuracy. Note that the region where $\frac{c\tau_2}{r_0} < 1$ is not shown as in this region $\left| \frac{d}{dr_0} \left(\frac{G_{0,1}}{G_{0,2}} \right) \right| \rightarrow \infty$, obscuring the remaining detail in the graph. However, it is the region where $\frac{r_0 + c\tau_1}{r_0} \approx \frac{r_0 + c\tau_2}{r_0}$ that is of particular interest. Here, $\left| \frac{d}{dr_0} \left(\frac{G_{0,1}}{G_{0,2}} \right) \right|$ approaches zero leading to a very large ϵ_0 . In the extreme case, where $\tau_1 = \tau_2$, no range estimate may be found as $\frac{G_{0,1}}{G_{0,2}}$ will be unity for all values of r_0 . Similarly, no range estimate may be found if τ_1 or τ_2 equals zero, as $\frac{G_{0,1}}{G_{0,2}}$ will be zero or undefined respectively, for all values of r_0 .

The analysis in this section has been limited to the three microphone case. However, the results of our analysis have implications for implementations of the Range-Finder using any number of microphones. To obtain accurate range estimates, we require access to a minimum of three microphones for which no two are equidistant (or approximately equidistant) from the sound source. Furthermore, we will not achieve

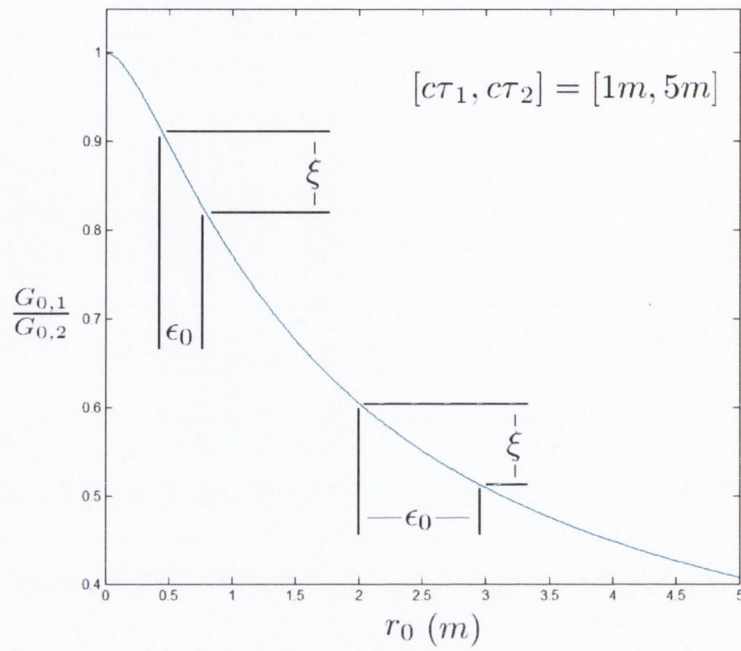


Figure 8.2: $\frac{G_{0,1}}{G_{0,2}}$ versus r_0 for $[c\tau_1, c\tau_2] = [1m, 5m]$. Range estimate error increases with r_0 .

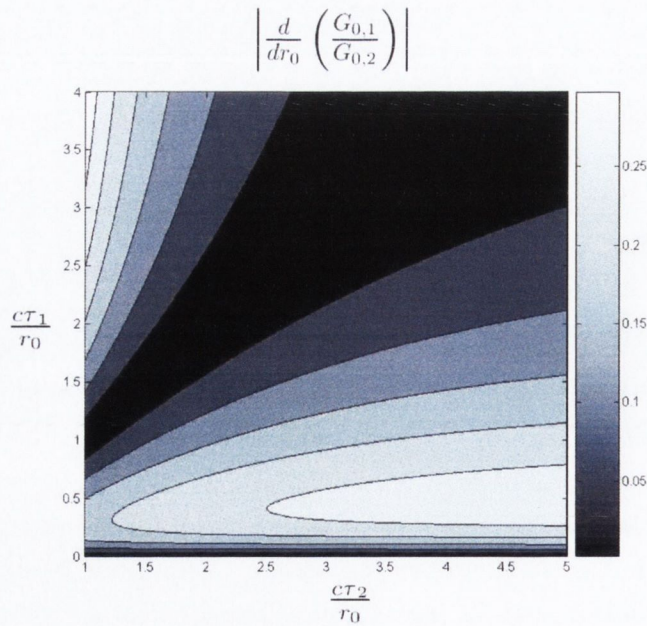


Figure 8.3: $\left| \frac{d}{dr_0} \left(\frac{G_{0,1}}{G_{0,2}} \right) \right|$ with respect to $\frac{c\tau_1}{r_0}$ and $\frac{c\tau_2}{r_0}$.

accurate range estimation when $r_0 \gg c\tau_1, c\tau_2$. Under such conditions we may expect $Q_{0,1,2}$ to exhibit a wide distribution and significant bias. This bias/error will then translate into a large error in the range estimate due to $\frac{r_0+c\tau_1}{r_0} \approx \frac{r_0+c\tau_2}{r_0}$.

We should not, therefore, apply the Range-Finder algorithm in what might be considered the classical microphone array scenario - that of closely-spaced microphones and a distant, farfield source. Rather, successful implementation would require microphones to be positioned in such a way that they are unlikely to be equidistant from the source and, ideally, we will have access to at least 3 microphones for which $r_0 \ll c\tau_1 \ll c\tau_2$. Similar requirements exist when implementing the SRP and Naive range estimators. From inspection of (8.11), it is apparent that the magnitude of the elements of $\tilde{\mathbf{D}}(r_0)$ tend to become equal as r_0 increases. Therefore, when $r_0 \gg c\tau_1, c\tau_2, \dots, c\tau_{M-1}$, $\tilde{\mathbf{D}}(\tilde{r}_0) \approx \tilde{\mathbf{D}}(r_0)$ over a large range of \tilde{r}_0 . Any error due to reverberation will, potentially lead to a large error in the range estimate. The Naive range estimator is also prone to errors when the two microphones used to determine a range estimate are approximately equidistant to the source. In such cases the output power of each microphone is approximately equal, causing the denominator of (8.8) to approach zero.

8.5 Simulations and Experiments

8.5.1 Simulations

A series of simulations were performed to examine the performance of the Range-Finder algorithm and compare it to that of the naïve and SRP-based range estimators, under varying reverberant conditions. Our simulated environment, figure (8.4), was a simple rectangular room of dimensions $[5.25m, 6.95m, 2.44m]$ and uniform surface absorption coefficient of 0.3. In this room we simulated three omnidirectional sources and six omnidirectional microphones, (see Table 8.1 for coordinates). The sampling frequency used was $10kHz$. The source-microphone impulse responses were generated using an acoustic modeling software package [117]. A raytracing algorithm was used to determine first $20ms$ of the impulse response after and including the arrival of the direct-path component. Statistical, random reverberant tails were used for the remaining reflections. Two “source signals” - a maximum-length-sequence (MLS) of $5.5s$ in duration and concatenated voice samples of approximately $13s$ total duration, both bandlimited to avoid aliasing - were convolved with each impulse response to obtain the simulated “recordings”. The TDEs were calculated geometrically, using the source and microphone coordinates and a known speed of sound.

The recordings were split into segments of 8192 samples and windowed using a

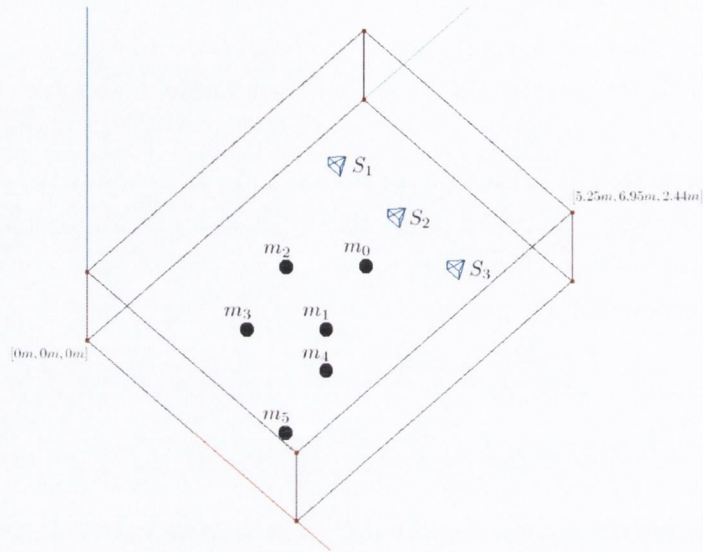


Figure 8.4: A diagram of the simulated room and setup. For precise coordinates of the microphones and loudspeakers, see Table 1.

(m)	m_0	m_1	m_2	m_3	m_4	m_5	S_1	S_2	S_3
x	3	3	2	2	4	4	1	2.5	4
y	4	3	3	2	2	1	5.5	5.5	5.5
z	2	1	2	1	2	1	1	1	1

Table 8.1: The coordinates of the microphone and source locations for the simulated room. Coordinates are in meters

Hamming window. The segment overlap was 50%. For each segment, the three range estimation techniques were then used to estimate the distance between the sources and each of the microphones. Negative range estimates and estimates greater than $5m$ were ignored. When applying the Range-Finder algorithm to speech recordings, the signals were separated into eight non-overlapping subbands with bandwidth $\frac{10}{16}kHz$ and $\bar{\Lambda}$ was determined as described in section (8.3.3). The estimates made using the naïve range estimator were found using the two microphones closest to the source so as to achieve the best possible results.

To investigate the effect of reverberation, the DRR -at- $1m$ of the simulated room was varied by applying an appropriate scaling to the direct-path components of the simulated impulse responses. Range estimates were then obtained as previously described.

In figure (8.5), the performance of the Range-Finder algorithm is compared to that of the naïve and SRP-based range estimators, for Source 2. The mean of the range estimates, \pm one standard deviation are shown with respect to the DRR -at- $1m$. The results shown relate to the estimates of r_0 only. Estimates of the remaining ranges (r_1 to r_5) are omitted because, as is apparent from (8.7), these will exhibit an identical bias and distribution to those corresponding to r_0 . Note that m_0 is the closest microphone to each source. The estimates of r_0 will, therefore, exhibit the greatest percentage error.

The Range-Finder and naïve range estimator behave as might be expected and return mean estimates that tend towards the correct range as the DRR -at- $1m$ increases. In both the voice and MLS cases, the Range-Finder algorithm outperforms the naïve range estimator for all values of the DRR -at- $1m$. The behavior of the SRP-based range estimator is somewhat more unusual, in that it returns estimates with a bias that remains almost constant with increasing DRR -at- $1m$. At low values of the DRR -at- $1m$, the SRP-based range estimator outperforms the Range-Finder.

Further comparisons of the performance of the Range-Finder and SRP-based range estimator are shown in figures (8.6) and (8.7), corresponding to Source 1 and Source 3 respectively. The results obtained using the naïve range estimator are omitted in each case. For the MLS recordings, the results returned by the naïve range estimator were broadly in line with those shown in figure (8.5), while, for the speech recordings, the range estimates obtained were greater than $5m$ and were, therefore, outside the cutoff boundary. The omission of the naïve estimator results allows us to more closely inspect the relative performance of the Range-Finder and SRP-based estimator.

The means of the results obtained using the voice recordings are slightly more accurate than those found using the MLS recordings, albeit with a significantly greater variance. For low values of the DRR -at- $1m$, range estimates obtained using the Range-

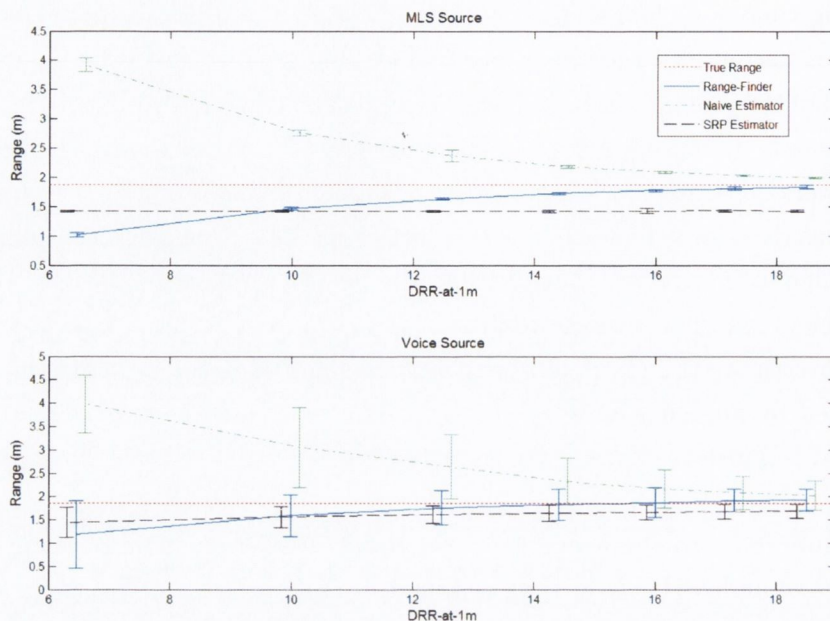


Figure 8.5: Range estimates \pm one standard deviation, obtained using the Range-Finder, Naïve and SRP-based methods, for source 2.

Finder are subject to a negative bias that reduces as the reverberation levels decrease. In the previous section we discussed the factors that may explain the presence of a bias in the range estimates. While, it is not necessarily the case that any such bias should be negative, from inspection of the PDFs in figure (8.1) we see that the density below the mean tends to be greater than that above. We may speculate, therefore, that any bias present would be negative, although the precise nature of such a bias is ultimately determined by the reverberation levels present and the array geometry and estimate bounds used. In all but one case (Source 1, voice recordings), the SRP-based range estimates exhibit a negative and almost constant bias. The reason for this bias is unclear but it is speculated that, as with the Range-Finder, it occurs as a result of the specific array geometry and specifications used.

With the exception of the results for Source 1 using the voice recordings, figures (8.6) and (8.7) repeat the pattern of figure (8.5), whereby the performance of the Range-Finder is initially inferior to but then “overtakes” that of the SRP-based approach. The results of our simulations suggest that the naive range estimator is the worst performing of those tested. The performance of the two remaining estimators may be said to be largely comparable with the Range-Finder algorithm and the SRP-based estimator tending to be superior at high and low values of the DRR -at-1m,

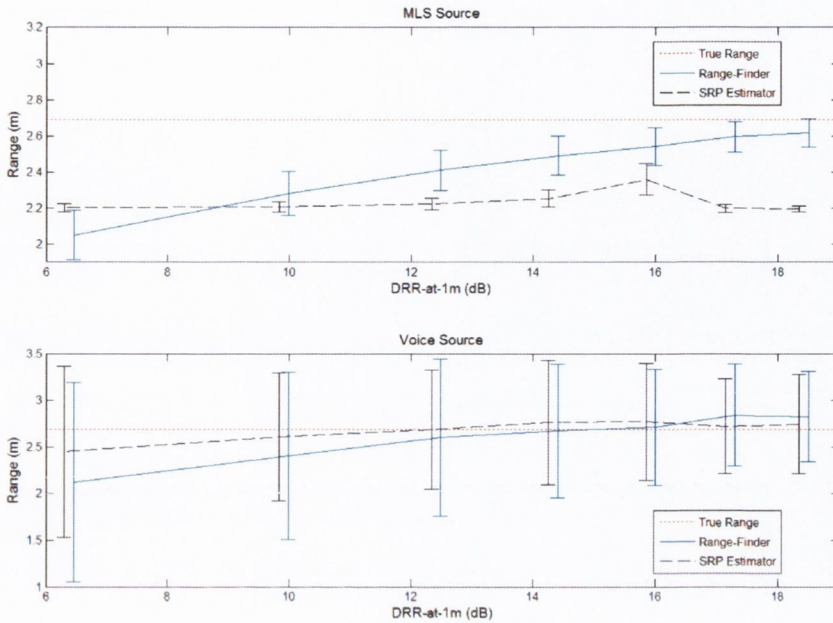


Figure 8.6: Range estimates \pm one standard deviation, obtained using the Range-Finder and SRP-based methods, for source 1.

respectively.

8.5.2 Experiments

A series of recordings were made to test the Range-Finder under real conditions. The recordings were made in an office, which was chosen for being a highly reverberant environment that would best test the performance of the range estimation algorithms. Six microphones were positioned at distances of between $0.8m$ and $3m$ from a loudspeaker, at intervals of roughly $0.5m$. The loudspeaker and microphones were arranged so as to be approximately colinear, so as to avoid errors due to the directionality of the source. Voice and MLS signals were produced by the loudspeaker. The microphone outputs were recorded before being bandlimited and downsampled to a sampling rate of $10kHz$. These recordings were then split into segments of 8192 samples and windowed using a Hamming window. The segment overlap was 50%. The TDEs were found using a PHAT-GCC [62] and range estimates were obtained for each segment. This procedure was repeated for each of three setups in which the loudspeaker and microphones were arranged colinearly along the length and each diagonal of the office respectively.

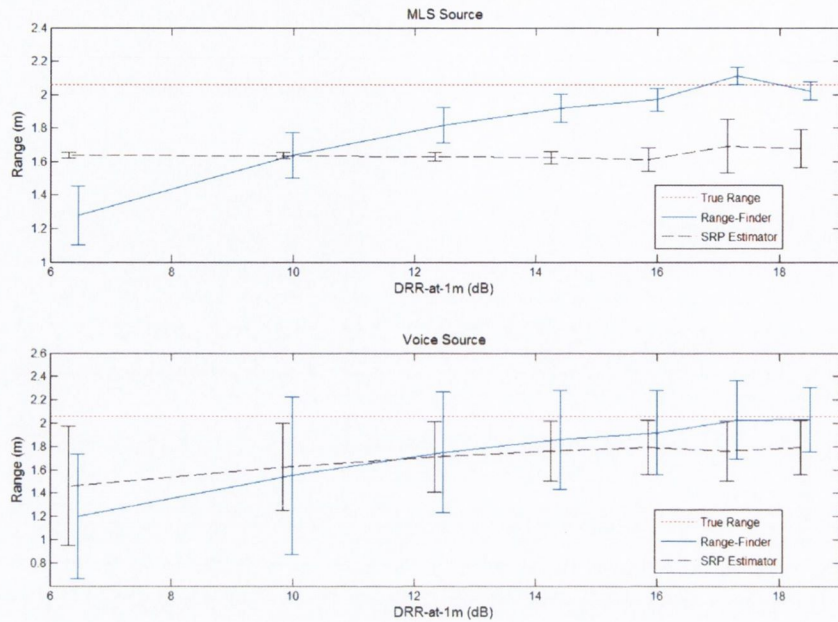


Figure 8.7: Range estimates \pm one standard deviation, obtained using the Range-Finder and SRP-based methods, for source 3.

The results, for the MLS and voice recordings, are shown in figures (8.8) and (8.9) respectively. For the MLS recordings, the Range-Finder and SRP-based range estimator demonstrate comparable performance and both are clearly superior to the Naive range estimator. For the voice recordings, the Range-Finder provides the best range estimates while, once again, the Naive range estimator yields the poorest results. Also, we observe no noticeable trend with respect to the bias in the mean of the estimates of the Range-Finder and the SRP-based estimator.

8.6 Discussion

We have proposed a novel method for estimating source-microphone ranges - the “Range-Finder” - that is robust against the effects of reverberation and which requires no information, regarding microphone locations, in order to return a range estimate. We have discussed the factors affecting the distribution and accuracy of the range estimates obtained by our method. We have presented the results of experiments, using real and simulated data, which demonstrate the efficacy of the Range-Finder algorithm and compare its performance with that of “naive” and SRP-based based range estimators.

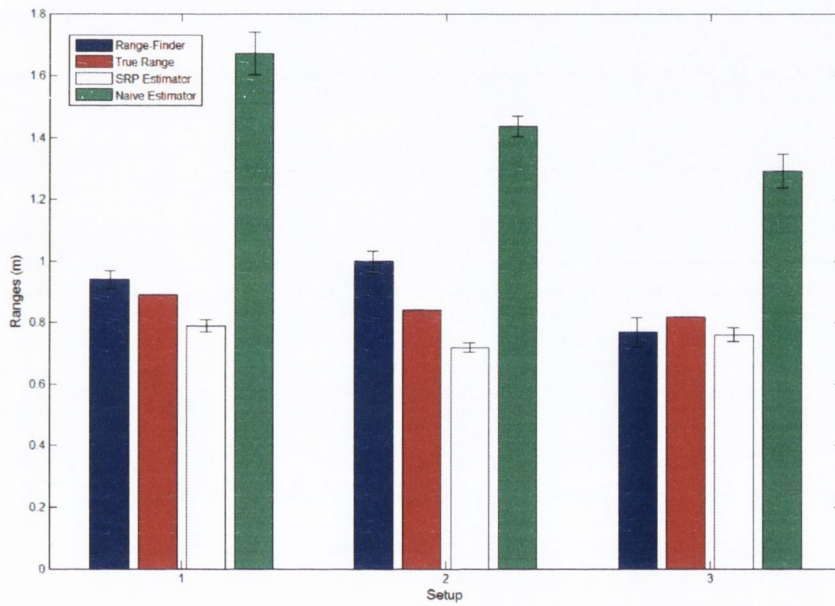


Figure 8.8: Range estimates \pm one standard deviation, obtained using the Range-Finder, Naïve and SRP-based methods, using real recordings of a maximum-length sequence.

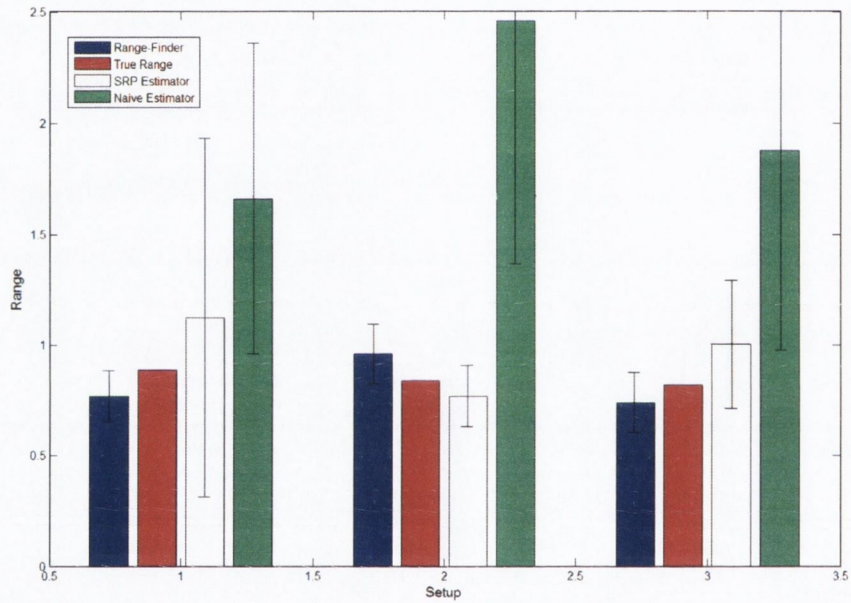


Figure 8.9: Range estimates \pm one standard deviation, obtained using the Range-Finder, Naïve and SRP-based methods, using real recordings of concatenated speech samples.

Of the three range estimators examined, the naive estimator consistently yielded the poorest results. The performances of the Range-Finder algorithm and SRP-based estimator were largely comparable and neither method was observed to be consistently superior to the other. Either of these approaches would, therefore, be suitable for range estimation using microphone arrays of unknown geometry.

While none of the range estimation methods tested require knowledge of the microphone locations, our analysis in section 8.4 revealed that the accuracy of any range estimates so obtained is, nonetheless, effected by the relative positioning of the microphones and the sound source. In particular, it was found that we can expect the range estimates to be inaccurate if $r_0 \gg c\tau_1, c\tau_2, \dots, c\tau_{M-1}$. Rather, successful implementation of the Range-Finder requires that the microphones be positioned such that there is a sufficient "spread" in the distances from the source to each microphone. This then precludes the application of the Range-Finder method to the classical scenario of closely spaced microphones and a farfield source. However, when the source is in the near-field or interior of an array of distributed microphones, it is also likely that we will have access to at least three microphones for which $r_0 \ll c\tau_1 \ll c\tau_2$. We may, therefore, expect accurate range estimates in the scenarios under investigation in this thesis.

It has been previously suggested that range estimation methods might be compared to TDE-based source localization techniques when the array geometry is known. However, the relative positions of the microphones and sound source have a significant bearing upon the accuracy or otherwise of both range-estimation techniques and TDE-based source localization algorithms. Consider, for example, a scenario in which the microphones and source are colinear. In such cases, the relative intersensor time delays are identical for all values of r_0 (assuming that the source is not in the interior of the array). As a result, TDE-based source localization algorithms could not return a unique estimate of r_0 . Where, the source and microphones are nearly colinear, we can expect TDE-based methods to exhibit large estimation errors due to errors in the TDEs. Conversely, the range-estimation techniques will fail in scenarios in which the microphones are all equidistant from the source. It is apparent, therefore, that any experimental comparison such as that suggested would yield results that are specific to the array geometry and source locations used and could not, therefore, be considered valid in general.

So far, we have assumed an omnidirectional source. In doing so, we have ignored a very pressing practical problem. In reality, sources of interest are likely to be directional and the received sound intensity will depend not only upon the microphone's distance from the source but also its relative azimuth and elevation. If the azimuth-elevation-dependant gain were known for each microphone it could easily be included in

our formulation of the Range-Finder. However, we are unlikely to have such information, or, indeed, to know the orientation of the source relative to the microphones. A further complicating factor is that source-directionality is frequency-dependant, with sources typically becoming increasingly directional with frequency.

We should, however, be careful not to overstate the difficulties that directionality presents. Some studies would suggest that directivity would not be a significant factor at frequencies below $4kHz$ and within an azimuth of $\pm 30^\circ$ relative to the direction in which a talker is facing [122]. If we could assume that the microphones were within some angular boundaries relative to the source then we may apply the Range-Finder with confidence. Yet, in the absence of comprehensive data regarding azimuth-elevation-dependant gain for the source of interest, it is hard to see how we might specify and justify the required angular boundaries. We therefore require such data and are limited in application when it is not available.

We note that not all microphones need be within the specified boundaries - only a minimum of 3 need be - and the remaining ranges may be found from the TDEs. Future work will focus on determining the directionality of typical sources and on methods for automatically determining which, if any, of the microphones we should use in the presence of a directional source.

While the Range-Finder may not be reliably employed in all situations, it remains effective in precisely those situations where most source-localisation strategies fail. Consider, once again, the case of colinear or nearly colinear microphones and sound source. While TDE-based source-localization techniques fail, the Range-Finder remains effective. Furthermore, source directionality is no-longer a problem. This suggests a role for the Range-Finder as an auxiliary source localization algorithm.

Chapter 9

Conclusion

In this final chapter we summarize our main findings and contributions, present suggestions for future work and conclude the thesis.

9.1 Summary

In the introduction to this thesis we outlined the potential advantages represented by speech capture using arrays of spatially distributed, remote microphones. In particular, we chose to focus upon the scenario of classroom-based videoconferencing and discussed the associated benefits and technical challenges. The remainder of the thesis was then outlined.

In chapter 2, we discussed the characteristics and phenomena associated with sound propagation in enclosed “room” environments. We introduced the concept of the source-room-microphone system as being linear and time-invariant and hence wholly characterized by an impulse/frequency response. Following from an investigation of such impulse responses it was found that the reverberant component of the received signal energy at any microphone could be modelled as a random variable distributed about a range-independent mean. This important result would provide part of the basis for the derivation of the novel range estimation algorithm proposed in chapter 8 and led us to propose a new metric for quantifying the degree of reverberation present in a room – the DRR-at-1m. A review of the literature confirmed that noise and reverberation degrade both the quality and intelligibility of recorded speech – thus verifying the need for techniques to mitigate against this degradation.

In chapter 3, we outlined the theoretical underpinnings of array processing. We addressed the issues of steering, spatial aliasing and wavenumber smoothing. We focused, in particular, upon filter-and-sum processing and discussed the impact of sensor characteristics and array geometry upon the wavenumber-frequency response

of such systems. Finally we defined the notation used throughout the remainder of the thesis and described the anechoic and reverberant signal models typically used throughout the literature.

The state-of-the-art for multi-microphone speech enhancement was reviewed in chapter 4, while previously-published time-delay estimation and source localization techniques were reviewed in chapter 5. Throughout these, the relationships between the various techniques were explored and highlighted. In chapter 6, we discussed practical implementations of the methods reviewed in the previous two chapters and showed how these were based, explicitly or implicitly, upon one or both of two simplifying assumptions/requirements – those of far-field sources and known array geometry. We went on to discuss how these assumptions/requirements – while simplifying to a certain degree – could, in fact, represent a significant disadvantage in some scenarios. An alternative approach, which treated the array geometry as unknown and the source location as being potentially in the near-field, was proposed. The technical challenges posed by such an approach provided the motivation for the two novel algorithms proposed in this thesis – leaky-LMS-based steering and the “Range-Finder” technique.

In chapter 7 we derived and tested a leaky-LMS-based method for steering arrays of unknown geometry. As discussed in (section 5.2.3), LMS time-delay estimation has been previously proposed in various forms and by several authors. These methods would yield inter-sensor time-delay estimates which could be used to steer an array. However, in contrast with these, our approach requires no intermediate estimation of time-delays. Furthermore, we verify the efficacy of our method under reverberant conditions. To our knowledge, the performance of LMS-based time delay estimation techniques in reverberant environments has not been previously investigated. We show our approach to achieve correct steering in simulated and reverberant environments. In addition, we identified those conditions under which our method is less computationally complex than steering using the well-known PHAT-GCC approach.

In chapter 8, we derived and demonstrated a novel method for source-microphone range estimation in reverberant environments for arrays of unknown geometry. Our approach – the “Range-Finder” algorithm – was compared with a well-known but naïve range estimator as well as a range estimation technique based upon a modest variation of the classical steered-response-power (SRP) method for source localization. We analyzed and discussed the impact of the relative locations of the microphones and source on the accuracy of our approach and tested it under real and simulated conditions. The Range-Finder was found to significantly outperform the naïve range estimator under all tested conditions while offering comparable (if not slightly improved) performance to that achieved by the SRP-based range estimation method.

9.2 Future work

9.2.1 Leaky-LMS-Based Steering

Future work should investigate the impact of the update parameters with respect to the speed of convergence of the filters and the steering accuracy achieved by the converged weightvector. The object of any such investigation should be to determine a range of values for α and μ that are appropriate for given reverberation levels and/or received signal energy etc.

In this thesis, we have applied leaky-LMS-based steering to non-adaptive beamformers. Future work should investigate the application of such steering to adaptive beamformers. In particular, it would be desirable to determine whether or not the precise steering achieved by the leaky-LMS method is capable of reducing the degree of target-signal-cancellation, due to missteering, that is observed in implementations of generalized sidelobe cancellers.

We have analyzed the performance of the proposed method for steering under reverberant but noise free conditions. Future work should investigate the performance of this method using a speech-noise discriminator in the presence of temporally sparse noise as well as that achieved using spectral subtraction as proposed in [35], [36] and [37] for stationary noise.

9.2.2 Range Estimation.

As previously discussed, practical implementation of the Range-Finder and other range estimation approaches will require greater understanding of source-directionality. Future work should, therefore, investigate source-directionality with particular attention paid to the case where the sound source is a human talker. Investigations of head source directionality are available in the literature [122],[123]. These, however, used dummy (mannequin) heads and investigate source-gain variations with respect to changes in azimuth only. Further studies, using real people or verifiably accurate models, are required. These should investigate head directionality with respect to both azimuth and elevation and should be large enough to provide information regarding the variation of head directionality within the population.

Additional future work can then investigate methods by which directionality-information may be incorporated to obtain improves range estimates. We note that source directionality information could be exploited to obtain accurate estimates of \mathbf{H}_{dp} , which could then be used in the implementation of the speech enhancement and source localization methods reviewed in chapters 4 and 5 respectively. Such possibilities should also be explored.

Further comparative investigation of the Range-Finder and SRP-based range estimation should be undertaken to determine the conditions (if any exist) under which one method might be expected to consistently outperform the other. This would provide a basis for discriminating between the scenarios in which either method should be preferred to the other.

9.3 Conclusion

More than two decades have passed since the publication of the landmark array-processing papers by Griffiths and Jim, [21], and Knapp and Carter, [62]. In that time and despite considerable research efforts, microphone array processing techniques have failed to evolve to towards any significant commercial application. This failure may be attributed, at least in part, to a predominant focus in the literature upon applications featuring fixed, far-field arrays of known geometry.

Whereas the often-stated goal of microphone array processing is to provide unobtrusive sound capture and location information for unencumbered talkers, fixed arrays are cumbersome and lack the flexibility and portability that would allow them to achieve these aims in practice. In addition, placing microphones far from target sound sources is counter-productive as both the quality and intelligibility of speech is known to degrade with increasing source-microphone distance/range.

Applications using easily-deployed, ad-hoc arrays promise useful adaptability. Despite this, such applications have received little attention in the literature. In this thesis, we have proposed and demonstrated methods for steering and range estimation which may be implemented using arrays of unknown geometry in reverberant environments. It is hoped that, in the future, the development of these and similar methods will lead to practical and useful technologies.

Appendix A

The Multichannel Wiener Filter

We seek to optimally approximate some reference signal, $K(\omega)$, by applying a weightvector, $\mathbf{W}(\omega)$, to an observation vector $\mathbf{Y}(\omega)$. The approximation error, $\varepsilon(\omega)$ is given by

$$\varepsilon(\omega) = K(\omega) - \mathbf{W}^H(\omega)\mathbf{Y}(\omega) \quad (\text{A.1})$$

Multiplying each side by its conjugate and applying the expectation operator,

$$E\{|\varepsilon(\omega)|^2\} = E\{|K(\omega)|^2\} - \mathbf{W}^H(\omega)\mathbf{R}_{\mathbf{Y}\mathbf{Y}}(\omega)\mathbf{W}(\omega) - E\{\mathbf{W}^H(\omega)\mathbf{Y}(\omega)K^*(\omega) + K(\omega)\mathbf{Y}^H(\omega)\mathbf{W}(\omega)\} \quad (\text{A.2})$$

$E\{|\varepsilon(\omega)|^2\}$ is a real-valued and analytic function with respect to $\mathbf{W}(\omega)$ and $\mathbf{W}^H(\omega)$. This being the case, it is shown in [16] that we may determine the stationary points of $E\{|\varepsilon(\omega)|^2\}$ by treating $\mathbf{W}(\omega)$ and $\mathbf{W}^H(\omega)$ as independent variables, differentiating with respect to $\mathbf{W}(\omega)$ or $\mathbf{W}^H(\omega)$ and setting the resulting derivative equal to zero. Therefore, differentiating with respect to $\mathbf{W}^H(\omega)$, $E\{|\varepsilon(\omega)|^2\}$ is minimized if

$$\mathbf{R}_{\mathbf{Y}\mathbf{Y}}(\omega)\mathbf{W}(\omega) - E\{\mathbf{Y}(\omega)K^*(\omega)\} = 0 \quad (\text{A.3})$$

Therefore, it is apparent that the optimal weightvector (in the sense that it minimizes the expected squared difference between the filter output and the reference signal) is given by

$$\mathbf{W}_{opt}(\omega) = \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1}(\omega)E\{\mathbf{Y}(\omega)K^*(\omega)\} \quad (\text{A.4})$$

In the special case where the reference signal is the noise-free output of some microphone in the array (let us arbitrarily assign the reference microphone as being m_0) we get

$$\mathbf{W}_{opt}(\omega) = \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1}(\omega)E\{\mathbf{Y}(\omega)X_0^*(\omega)\} \quad (\text{A.5})$$

Assuming that the target signal and noise are uncorrelated with each other, we may write

$$\mathbf{W}_{opt}(\omega) = (\mathbf{R}_{\mathbf{X}\mathbf{X}}(\omega) + \mathbf{R}_{\mathbf{N}\mathbf{N}}(\omega))^{-1} \mathbf{R}_{\mathbf{X}\mathbf{X}}(\omega)\mathbf{v}_0 \quad (\text{A.6})$$

where $\mathbf{v}_0 = [1, 0, \dots, 0]_{1 \times M}^T$

This solution may be decomposed as follows. Let us assume that $R_{NN}(\omega)_{i,i} = R_{NN}(\omega)_{j,j} = P_N(\omega) \forall i, j$, then we may define the signal coherence matrix, $\mathbf{\Gamma}_{NN}(\omega)$.

$$\mathbf{\Gamma}_{NN}(\omega) = P_N^{-1}(\omega) \mathbf{R}_{NN}(\omega) \quad (\text{A.7})$$

From this (and omitting the frequency index, ω , for clarity)

$$\mathbf{W}_{opt} = (|S|^2 \mathbf{H}\mathbf{H}^H + P_n \mathbf{\Gamma}_{NN})^{-1} (|S|^2 \mathbf{H}\mathbf{H}^H \mathbf{v}_0) \quad (\text{A.8})$$

By applying the matrix inversion lemma we obtain the following

$$(|S|^2 \mathbf{H}\mathbf{H}^H + P_n \mathbf{\Gamma}_{NN})^{-1} = \frac{1}{P_n} \left[\mathbf{\Gamma}_{NN}^{-1} - \frac{\mathbf{\Gamma}_{NN}^{-1} \mathbf{H}\mathbf{H}^H \mathbf{\Gamma}_{NN}^{-1}}{\mathbf{H}^H \mathbf{\Gamma}_{NN}^{-1} \mathbf{H} + \frac{P_n}{|S|^2}} \right] \quad (\text{A.9})$$

Inserting this result into our previous expression for \mathbf{W}_{opt} yields

$$\mathbf{W}_{opt} = \frac{|S|^2}{P_n} \left[\mathbf{\Gamma}_{NN}^{-1} - \frac{\mathbf{\Gamma}_{NN}^{-1} \mathbf{H}\mathbf{H}^H \mathbf{\Gamma}_{NN}^{-1}}{\mathbf{H}^H \mathbf{\Gamma}_{NN}^{-1} \mathbf{H} + \frac{P_n}{|S|^2}} \right] \mathbf{H}\mathbf{H}^H \mathbf{v}_0 \quad (\text{A.10})$$

and following some simple algebraic manipulation we obtain

$$\mathbf{W}_{opt} = \underbrace{\left[(\mathbf{H}^H \mathbf{\Gamma}_{NN}^{-1} \mathbf{H})^{-1} (\mathbf{\Gamma}_{NN}^{-1} \mathbf{H}) H_0^* \right]}_{\mathbf{W}_{p1}} \underbrace{\left[\frac{|S|^2 \mathbf{H}^H \mathbf{H}}{|S|^2 \mathbf{H}^H \mathbf{H} + P_N} \right]}_{\mathbf{W}_{p2}} \quad (\text{A.11})$$

Appendix B

Constrained Optimization

The following constrained optimization problem can be solved using the method of Lagrange Multipliers

$$\mathbf{W}_{opt} = \arg \min_{\mathbf{W}} \{ \mathbf{W}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{W} \} \text{ subject to } \mathbf{W}^H \mathbf{C} = \mathbf{c} \quad (\text{B.1})$$

The Lagrangian of this expression is given by

$$\mathbf{W}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{W} + \lambda (\mathbf{W}^H \mathbf{C} - \mathbf{c}) \quad (\text{B.2})$$

where λ is the Lagrange multiplier. Finding the gradient of the Lagrangian with respect to \mathbf{W}^H and setting the result equal to zero yields

$$2\mathbf{R}_{\mathbf{Y}\mathbf{Y}} \mathbf{W} + \lambda \mathbf{C} = \mathbf{0} \quad (\text{B.3})$$

from which we get

$$\mathbf{W}_{opt} = -\mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1} \mathbf{C} \lambda / 2 \quad (\text{B.4})$$

This result must satisfy our constraint. Replacing $\mathbf{W}^H \mathbf{C} = \mathbf{c}$ with the equivalent $\mathbf{C}^H \mathbf{W} = \mathbf{c}^H$ we obtain

$$-\mathbf{C}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1} \mathbf{C} \lambda / 2 = \mathbf{c}^H \quad (\text{B.5})$$

from which we get

$$\lambda = \frac{-2\mathbf{c}^H}{\mathbf{C}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1} \mathbf{C}} \quad (\text{B.6})$$

Reinserting this result into (B.4) yields the final solution

$$\mathbf{W}_{opt} = \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1} \mathbf{C} (\mathbf{C}^H \mathbf{R}_{\mathbf{Y}\mathbf{Y}}^{-1} \mathbf{C})^{-1} \mathbf{c}^H \quad (\text{B.7})$$

Bibliography

- [1] Carl F. Eyring, "Reverberation time in dead rooms," *J. Acoust. Soc. Am.* 1, 168 (1930)
- [2] K.S. Sum and J. Pan, "On the steady state and transient decay methods for estimation of reverberation time," *Journal of the Acoustical Society of America*, vol 112, issue 6, pp. 2583-2588
- [3] Heather A. Knecht, Peggy B. Nelson, Gail M. Whitelaw, and Lawrence L. Feth, "Background noise levels and reverberation times in unoccupied classrooms," *American Journal of Audiology*, Vol.11 65-71, December 2002
- [4] American National Standards Institute, (2002). Acoustical performance criteria designa requirements and guidelines for schools. ANSI s12.60.
- [5] M. M. Sondhi, D. R. Morgan and J. L. Hall, "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Sig. Proc. Letters*, Vol 2, No 8, pp.146-151, August 1995
- [6] International Organization for Standardization (ISO) 226, 2003
- [7] J. S. Bradley, "Speech intelligibility studies in classrooms," *J. Acoust. Soc. Am.* 80, 846, 1986
- [8] H. Haas, "The influence of a single echo on the audibility of speech," *Journal of the Audio Engineering Society* 20, 145-159, 1972
- [9] J. S. Bradley, R. D. Reich, and S. G. Norcross, "On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility," *J. Acoust. Soc. Am.* 106, 1820, 1999
- [10] Sylvio R. Bistafa and John S. Bradley, "Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics," *J. Acoust. Soc. Am.* 107, 861, 2000
- [11] Murray Hodgson and Eva-Marie Nosal, "Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms," *J. Acoust. Soc. Am.* 111, 931, 2002
- [12] Anna K. Nabelek and Pauline K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages," *J. Acoust. Soc. Am.* 71, 1242, 1982

- [13] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.* 7, 724-739, 1974
- [14] A. J. Duquesnoy, "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," *J. Acoust. Soc. Am.* 74, 739, 1983
- [15] D. A. Berkley, "Normal listeners in typical rooms," *Acoustical Factors Affecting Hearing Aid Performance*, G.A Studebaker & I. Hochberg Eds., University Park Press, 1980
- [16] Don H. Johnson, Dan E. Dudgeon, *Array Signal Processing; Concepts and Techniques*, Prentice Hall, 1992
- [17] Harry L. Van Trees, *Optimum Array Processing; Detection, Estimation, and Modulation Theory, Part IV*, Wiley-Interscience
- [18] H. R. Abutaleb, H. Sheikhzadeh, R. L. Brennan and G. H. Freeman, "A hybrid subband adaptive system for speech enhancement in diffuse noise fields," *Signal Processing Letters, IEEE* , vol.11, no.1pp. 44- 47, Jan 2004
- [19] I. A. McCowan and H. Bourslard, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on* , vol.11, no.6pp. 709- 716, Nov 2003
- [20] O.L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE* , vol.60, no.8pp. 926- 935, Aug 1972
- [21] L. Griffiths and C Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol.30, no.1pp. 27- 34, Jan 1982
- [22] L. Griffiths and K. Buckley, "Quiescent pattern control in linearly constrained adaptive arrays," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol.35, no.7pp. 917- 926, Jul 1987
- [23] Feng Qian and B. D. Van Veen, "Partially adaptive beamformer design subject to worst-case performance constraints," *Signal Processing, IEEE Transactions on* [see also *Acoustics, Speech, and Signal Processing, IEEE Transactions on*] , vol.42, no.5pp.1218-1221, May 1994
- [24] K. Duvall, *Signal Cancellation in Adaptive Arrays; Phenomena and a Remedy*, PhD Thesis, Dept. Elec. Eng., Stanford Univ., Sept 1983
- [25] I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beaming," *Antennas and Propagation, IEEE Transactions on* , vol.40, no.9pp.1093-1096, Sep 1992

- [26] S. Nordebo, I. Claesson and S. Nordholm, "Adaptive beamforming: spatial filter designed blocking matrix," *Oceanic Engineering, IEEE Journal of* , vol.19, no.4pp.583-590, Oct 1994
- [27] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol.2, pp.925-928, 7-10 May 1996
- [28] O. Hoshuyama, A. Sugiyama and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *Signal Processing, IEEE Transactions on*, vol.47, no.10pp.2677-2684, Oct 1999
- [29] O. Hoshuyama, A. Sugiyama and A. Hirano, "A robust adaptive microphone array with improved spatial selectivity and its evaluation in a real environment," *Acoustics, Speech, and Signal Processing, IEEE International Conference on* , vol.1, pp.367-370, 21-24 Apr 1997
- [30] O. Hoshuyama and A. Sugiyama, "Robust Adaptive Beamforming," *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Press, 2001
- [31] W. H. Neo and B. Farhang-Boroujeny, "Robust microphone arrays using sub-band adaptive filters," *Vision, Image and Signal Processing, IEE Proceedings-* , vol.149, no.1pp.17-25, Feb 2002
- [32] Jianfeng Chen, L. Shue and Senjin Liu, "A fixed blocking matrix for robust microphone array beamforming," *Signal Processing and Its Applications, Seventh International Symposium on* , vol.2, pp. 407- 410, 1-4 July 2003
- [33] Zhu Liang Yu, Qiyue Zou and Meng Hwa Er, "A robust wideband array beamformer using fan filter," *Circuits and Systems, Proceedings of the 2003 International Symposium on*, vol.4, pp. IV-101- IV-104, 25-28 May 2003
- [34] Qiyue Zou, Zhu Liang Yu and Zhiping Lin, "Norm and coefficient constraints for robust adaptive beamforming," *Circuits and Systems, Proceedings of the 2003 International Symposium on* , vol.4, pp. IV-349- IV-352, 25-28 May 2003
- [35] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multi-microphone speech enhancement," *Signal Processing, IEEE Transactions on* [see also *Acoustics, Speech, and Signal Processing, IEEE Transactions on*] , vol.50, no.9pp. 2230- 2244, Sep 2002
- [36] A. Spriet, M. Moonen, and J. Wouters, "A multi-channel subband generalized singular value decomposition approach to speech enhancement," *Eur. Trans. Telecommun.*, Special Issue on Acoustic Echo and Noise Control, no. 2, pp.149-158, Mar-Apr 2002
- [37] A. Spriet, M. Moonen and Wouters, J. 2004. "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.* 84, 12, Dec 2004

- [38] M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with microphone array," *Vehicular Technology, IEEE Transactions on*, vol.48, no.5pp.1518-1526, Sep 1999
- [39] S. Nordholm, I. Claesson and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *Speech and Audio Processing, IEEE Transactions on*, vol.7, no.3pp.241-252, May 1999
- [40] B. Widrow, P.E. Mantey, L.J. Griffiths and B.B Goode, "Adaptive antenna systems," *Proceedings of the IEEE*, vol.55, no.12pp. 2143- 2159, Dec 1967
- [41] A. Spriet, M. Moonen and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *Speech and Audio Processing, IEEE Transactions on*, vol.13, no.4pp. 487- 503, July 2005
- [42] S. Haykin, Ed., *Unsupervised Adaptive filtering*, John Wiley and Sons, New York, 2000
- [43] A. Cichocki, and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley and Sons, New York, 2002
- [44] S. Araki, R. Mukai, S. Makino, T. Nishikawa and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *Speech and Audio Processing, IEEE Trans. On*, vol 11, no.2, March 2003
- [45] K. Torkolla, "Blind separation for audio signals –Are we there yet?," *Proc. Workshop on Independent Component Analysis and Blind Signal Separation*, Aussois, France, Jan 11-15, 1999
- [46] S. Makino, H. Sadawa, R. Mukai and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Trans. Fundamentals*, vol E88-A, no. 7, July 2005
- [47] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP Jnl. On Applied Signal Processing*, vol 11, pp.1157-1166, 2003
- [48] R. Mukai, H. Sawada, S. Araki and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP Jnl. Applied Signal Processing*, Volume 2006, pp. 1-13, 2006
- [49] A.V. Oppenheim, R.W. Schaffer, *Digital Signal Processing*, Prentice-Hall, 1975
- [50] G. M. Bogert, M. J. Healy and J.W. Tukey, "The quefrequency analysis of times series for echoes," *Proceedings of a Symposium on Time Series Analysis*, ed. M. Rosenblatt. New York: John Wiley, 1962

- [51] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," *Acoustics, Speech, and Signal Processing, IEEE International Conference on* , vol.1, pp. I-557- I-560, 2002
- [52] E. Jan, P. Svaizer and J. L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," *Circuits and Systems, IEEE International Symposium on* , vol.2, pp.1460-1463 vol.2, 30 Apr-3 May 1995
- [53] J. L. Flanagan, A. C. Surendran and E. E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.* 13, 1-2, pp.207-222, Oct 2003
- [54] N. Suditu, J. van de Laar and P. C. W. Sommen, "Evaluation of dereverberation capabilities of broadband beamformers," *12th Annual Workshop on Circuits, Systems and Signal Processing*, 29-30 November 2001
- [55] J. G. Ryan and R. A. Goubran, "Near-field beamforming for microphone arrays," *Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on* , vol.1, pp.363-366
- [56] J. Gonzalez-Rodriguez, J. L. Sanchez-Bote and J. Ortega-Garcia, "Speech dereverberation and noise reduction with a combined microphone array approach," *Acoustics, Speech, and Signal Processing, IEEE International Conference on* , vol.2, pp.II1037-II1040
- [57] Q. G. Liu, B. Champagne and P. Kabal, "Room speech dereverberation via minimum-phase and all-pass component processing of multi-microphone signals," *Communications, Computers, and Signal Processing, IEEE Pacific Rim Conference on* , pp.571-574, 17-19, 1993
- [58] S. M. Griebel and M. S. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," *Acoustics, Speech, and Signal Processing, 2001 IEEE International Conference on*, Volume 1, Page(s):201 - 204
- [59] B. W. Gillespie, H. S. Malvar and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Acoustics, Speech, and Signal Processing. Proceedings. 2001 IEEE International Conference on* , vol.6, pp.3701-3704
- [60] S. M. Griebel, *A Microphone Array System for Speech Source Localization, Denoising, and Dereverberation*, PhD Thesis, Harvard University, 2002
- [61] N. Gaubitch, P.A. Naylor and D.B. Ward. "On the use of linear prediction for dereverberation of speech," *Proceedings of the International Workshop on Acoustic Echo/Noise Control* , pages 99-102, September 2003.
- [62] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech, and Signal Processing, IEEE Transactions on* , vol.24, no.4pp. 320- 327, Aug 1976

- [63] S. Bedard, B. Champagne and A. Stephenne, "Effects of room reverberation on time-delay estimation performance," *Acoustics, Speech, and Signal Processing, 1994 IEEE International Conference on*, vol.ii, pp.II/261-II/264 vol.2, 19-22 Apr 1994
- [64] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," *Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on*, vol.1, no.pp.375-378 vol.1, 21-24 Apr 1997
- [65] Jingdong Chen, Jacob Benesty and Yiteng (Arden) Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 26503, 19 pages, 2006. doi:10.1155/ASP/2006/26503
- [66] <http://iwaenc05.ele.tue.nl/proceedings/papers/S03-10.pdf>
- [67] Michael S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.* 105, 2914, 1999
- [68] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol.4, no.pp. iv/985- iv/988 Vol. 4, 18-23 March 2005
- [69] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multisource TDOA estimation with a microphone array," *International workshop on acoustic echo and noise control, IWAEN2003, Sept 2003, Kyoto Japan*
- [70] D. Bechler and K. Kroschel, "Three different criteria for time-delay estimates," *Eusipco 2004, September 6-10, Vienna Austria*
- [71] P. Feintuch, N. Bershad and F. Reed, "Time delay estimation using the LMS adaptive filter—Static behavior," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol.29, no.3pp. 561- 571, Jun 1981
- [72] P. Feintuch, N. Bershad and F. Reed, "Time delay estimation using the LMS adaptive filter—Dynamic behavior," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol.29, no.3pp. 571- 576, Jun 1981
- [73] D. Youn, N. Ahmed and G. Carter, "On using the LMS algorithm for time delay estimation," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol.30, no.5pp. 798- 801, Oct 1982
- [74] Y. T. Chan, J. B. Plant and J. M. F. Riley, "Modeling of time delay and its application to estimation of nonstationary delays," *IEEE Trans. Acoust., Speech and Signal Proc.* Vol. ASSP-29, no. 3 Pt. 2, pp. 577-581. 1981

- [75] Y Chan, J Riley and J Plant, "A parameter estimation approach to time-delay estimation and signal detection," *IEEE Transactions on Acoustics Speech and Signal Processing*, 1980
- [76] P. C. Ching and Y. T Chan, "Adaptive time delay estimation with constraints," *Acoustics, Speech, and Signal Processing, IEEE Transactions on* , vol.36, no.4pp.599-602, Apr 1988
- [77] Shiunn-Jang Chern and Shyh-Neng Lin, "An adaptive time delay estimation with direct computation formula," *J. Acoust. Soc. Am.* 96, 811, 1994
- [78] S. Lin and S. Chern, "A new adaptive constrained LMS time delay estimation algorithm," *Signal Process.* 71, 1, 29-44, Nov 1998
- [79] Jacob Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.* 107, 384, 2000
- [80] Y. Huang, J. Benesty and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system," *Acoustics, Speech, and Signal Processing, 1999 IEEE International Conference on* , vol.2, pp.937-940, 15-19 Mar 1999
- [81] Yiteng Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *Signal Processing, IEEE Transactions on*, vol.51, no.1pp. 11- 24, Jan. 2003
- [82] Yiteng Huang and J. Benesty, "Adaptive blind channel identification: multi-channel least mean square and Newton algorithms," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol.2, pp.1637-1640, 2002
- [83] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Process.* 82, 8, 1127-1138, Aug 2002
- [84] D. Gisch and J.M. Ribando, "Apollonius' problems: a study of their solutions and connections," *American Journal of Undergraduate Research*, Vol 3, pp 15-26, 2004
- [85] Eric W. Weisstein, "Apollonius' Problem." From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/ApolloniusProblem.html>
- [86] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *Speech and Audio Processing, IEEE Transactions on* , vol.5, no.3pp.288-292, May 1997
- [87] M. Lallart and F. Boland, " Source localization using piezoelectric ultrasonic transducers," *ISSC 2006*, June 28-30, Dublin
- [88] M.S. Brandstein, J.E. Adcock and H.F. Silverman, "A closed-form method for finding source locations from microphone-array time-decay estimates,"

Acoustics, Speech, and Signal Processing, 1995 International Conference on , vol.5, pp.3019-3022 vol.5, 9-12 May 1995

- [89] M.S. Brandstein, J.E. Adcock and H.F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *Speech and Audio Processing, IEEE Transactions on* , vol.5, no.1pp.45-50, Jan 1997
- [90] Y. Huang, J. Benesty and G. W. Elko, "An efficient linear-correction least-squares approach to source localization," *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp.67-70, 2001
- [91] Y. Huang, J. Benesty and G. W. Elko, "Passive acoustic source localization for video camera steering," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol.2, pp.II909-II912, 2000
- [92] Y. Huang, J. Benesty, G. W. Elko and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *Speech and Audio Processing, IEEE Transactions on* , vol.9, no.8pp.943-956, Nov 2001
- [93] Kung Yao, R.E. Hudson, C.W. Reed, Daching Chen and F. Lorenzelli, "Blind beamforming on a randomly distributed sensor array system," *Selected Areas in Communications, IEEE Journal on* , vol.16, no.8pp.1555-1567, Oct. 1998
- [94] J. C. Chen, R. E. Hudson and Kung Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *Signal Processing, IEEE Transactions on*, vol.50, no.8pp.1843-1854, Aug 2002
- [95] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *Acoustics, Speech, and Signal Processing, IEEE Transactions on* , vol.31, no.5pp. 1210- 1217, Oct 1983
- [96] M. Viberg and A. L. Swindlehurst, "A Bayesian approach to auto-calibration for parametric array signal processing," *Signal Processing, IEEE Transactions on*, vol.42, no.12pp.3495-3507, Dec 1994
- [97] J.H. DiBiase, H.F. Silverman and M.S. Brandstein, "Robust Localization in Reverberant Rooms," *Microphone Arrays; Signal Processing Techniques and Applications*, M. Brandstein, D. Ward (Eds.), Springer Press
- [98] Kung Yao, J. C. Chen and R. E. Hudson, "Maximum-likelihood acoustic source localization: experimental results," *Acoustics, Speech, and Signal Processing, IEEE International Conference on* , vol.3, pp. III-2949- III-2952, 2002
- [99] Jingdong Chen, J. Benesty and Yiteng Huang, "Robust time delay estimation exploiting spatial correlation," *Acoustics, Speech, and Signal Processing, IEEE International Conference on* , vol.5, no.pp. V- 481-4 vol.5, 6-10 April 2003
- [100] S. T. Birchfield and D. K. Gillmor, "Acoustic source direction by hemisphere sampling," *Acoustics, Speech, and Signal Processing, IEEE International Conference on* , vol.5, pp.3053-3056, 2001

- [101] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian acoustic localization," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Vol. 2, pp. II/1793-II/1796. 2002
- [102] Paolo Minero, "State of the art on localization and beamforming of an acoustic source," minero@eecs.berkeley.edu, June 21, 2004
- [103] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," Speech and Audio Processing, IEEE Transactions on, vol.12, no.5pp. 499- 508, Sept. 2004
- [104] R. Duraiswami, D. Zotkin and L. S. Davis, "Active speech source localization by a dual coarse-to-fine search," Acoustics, Speech, and Signal Processing, IEEE International Conference on , vol.5, pp.3309-3312, 2001
- [105] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp.71-74, 2001
- [106] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," EURASIP Journal on Applied Signal Processing, 2003
- [107] B. Mungamuru and P. Aarabi, "Enhanced sound localization," Systems, Man and Cybernetics, Part B, IEEE Transactions on , vol.34, no.3pp. 1526- 1540, June 2004
- [108] J. C. Chen, Kung Yao and R. E. Hudson, "Source localization and beamforming," Signal Processing Magazine, IEEE , vol.19, no.2pp.30-39, Mar 2002
- [109] T. Pham and B. Sadler, "Aeroacoustic wideband array processing for detection and tracking of ground vehicles," J. Acoust. Soc. Am. 98, 2969, 1995
- [110] H. Wang and M. Kaveh, "Estimation of angles-of-arrival for wideband sources," Acoustics, Speech, and Signal Processing, IEEE International Conference on, vol.9, pp. 279- 282, Mar 1984
- [111] B. Friedlander and A.J Weiss, "Direction finding using spatial smoothing with interpolated arrays," Aerospace and Electronic Systems, IEEE Transactions on , vol.28, no.2pp.574-587, Apr 1992
- [112] Tie-Jun, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , vol.33, no.4pp. 806- 811, Aug 1985
- [113] Michael S. Brandstein, John E. Adcock, and Harvey F. Silverman, "Microphone-array localization error estimation with application to sensor placement," J. Acoust. Soc. Am. 99, 3807, 1996

- [114] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, 1997
- [115] S Forchhammer, A Fosgerau, PSK Hansen, R Sharp, E. Todricia and A. Zsigri, "Video conferencing for a virtual seminar room," Proc. 4th International Conference on Digital Signal processing and its applications, Moscow, 2002
- [116] Qiong Liu, D. Kimber, J. Foote and Chunyuan Liao , "Multichannel video/audio acquisition for immersive conferencing," Multimedia and Expo, International Conference on, Vol.3, 6-9 July 2003, Pages: III- 45-8 vol.3, 2003
- [117] EASE; Enhanced Acoustic Simulator for Engineers, version 4.0, <http://www.renkus-heinz.com/ease/>
- [118] H. Sorensen, D. Jones, M. Heideman and C. Burrus, "Real-valued fast Fourier transform algorithms," Acoustics, Speech, and Signal Processing, IEEE Transactions on, Vol.35, Iss.6, pp.849- 863, Jun 1987
- [119] M. J. Link and K. M. Buckley, "Prewhitening for intelligibility gain in hearing aid arrays," J. Acoust. Soc. Amer. 93 (4), 2139-2140, 1993
- [120] G. Marsaglia, "Ratios of Normal Variables," Journal of Statistical Software, Vol. 16, No. 4. May 2006
- [121] G. Marsaglia, "Ratios of normal variables and ratios of sums of variables," Journal of the American Statistical Association, Vol. 60, No. 309, pp. 193-204, Mar 1965
- [122] J. Huopaniemi, K. Kettunen and J. Rahkonen, "Measurement and modeling techniques for directional sound radiation from the mouth," Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on, pp.183-186, 1999
- [123] James L. Flanagan, "Analog measurements of sound radiation from the mouth," J. Acoust. Soc. Am. 32, 1613, 1960