Robust Acoustic Beamforming in the Presence of Channel Propagation Uncertainties

by
Richard Stanton

A Thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy of Imperial College London
and the Diploma of Imperial College London

Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London
2016

Declaration of Originality

I hereby certify that this thesis is the outcome of the research conducted by myself under supervision from Mike Brookes in the Department of Electrical and Electronic Engineering at Imperial College London. Any work that has been previously published and included in this thesis has been fully acknowledged in accordance with the standard referencing practices of this discipline.

Copyright

Abstract

Beamforming is a popular multichannel signal processing technique used in conjunction with microphone arrays to spatially filter a sound field. Conventional optimal beamformers assume that the propagation channels between each source and microphone pair are a deterministic function of the source and microphone geometry. However in real acoustic environments, there are several mechanisms that give rise to unpredictable variations in the phase and amplitudes of the propagation channels. In the presence of these uncertainties the performance of beamformers degrade. Robust beamformers are designed to reduce this performance degradation. However, robust beamformers rely on tuning parameters that are not closely related to the array geometry.

By modeling the uncertainty in the acoustic channels explicitly we can derive more accurate expressions for the source-microphone channel variability. As such we are able to derive beamformers that are well suited to the application of acoustics in realistic environments. Through experiments we validate the acoustic channel models and through simulations we show the performance gains of the associated robust beamformer.

Furthermore, by modeling the speech short time Fourier transform coefficients we are able to design a beamformer framework in the power domain. By utilising spectral subtraction we are able to see performance benefits over ideal conventional beamformers. Including the channel uncertainties models into the weights design improves robustness.

We design a dereverberation algorithm that operates in the short time Fourier transform domain that performs, with knowledge of the impulse response, as well as competing time domain methods.

Acknowledgements

I would primarily like to thank my supervisor, Mr Mike Brookes, who has had unwavering patience throughout the PhD. Our regular meetings have offered great guidance and support.

Many thanks also go to the Signal Processing group for making the process much more enjoyable. Special thanks go to John and Sithan for the many fruitful technical conversations and supportive friendship over the last 8 years.

Finally I would like to thank Tricia, for the huge amount of support and friendship, especially during the final stages.

Contents

List of Tables

List of Figures

Notation

For the majority of the thesis the following mathematical notation is used:

$x \in [a, b]$	$\{x \vert a \leq x \text{ and } x \leq b\}$;
x	A scalar quantity
$\mathbf{x}$	A vector quantity
$\mathbf{X}$	A matrix quantity
x_{ab}	The element at row a and column b of matrix $\mathbf{X}$
$\mathbf{X}^T$	The transpose of the matrix $\mathbf{X}$
$\mathbf{X}^*$	The complex conjugate of matrix $\mathbf{X}$
$\mathbf{X}^H$	The Hermitian transpose of matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	The inverse of the matrix $\mathbf{X}$
$\langle x \rangle$	The expectation of x
$\mathcal{N}$	Normal distribution
$\mathbb{R}$	The set of real numbers
$\mathbf{X} \in \mathbb{R}^{a \times b}$	The matrix $\mathbf{X}$ has dimensions $a \times b$ is from the set of real numbers
$\mathbb{I}$	The set of imaginary numbers
$\mathbb{C}$	The set of complex numbers
j	The imaginary constant
$\|\mathbf{x}\|$	The Euclidean norm of the vector $\mathbf{x}$
$\|x\|$	The absolute value of x
$\check{x}[l, k]$	The STFT of x, with frame index l and frequency index k
$f * g$	The convolution of f and g
$\hat{x}$	Estimate of x

$\bar{x}$	The average/nominal value of x
$\tilde{x}$	The uncertainty in x

Symbols

The following symbols are used for all chapters, except Chapter 4:

α_ω	The complex kurtosis at frequency ω
α_i	Head directivity pattern Fourier coefficient
c	Speed of sound
$d_{\epsilon a}$	Propagation coefficient from source a to microphone ϵ
$\mathbf{d}_a$	Propagation coefficients from source a to all microphones
$\mathbf{D}$	Propagation coefficients from all sources to all microphones
$\bar{d}_{\epsilon a}$, $\bar{\mathbf{D}}$	Nominal propagation coefficient
$\delta_{\phi_{\epsilon,a}}$	Azimuth rotation angle of the channel from the source a to the microphone ϵ
$\mathbf{F}$	The combined power domain beamformer weights
$g(\mathbf{x})$	Inverse propagation speed uncertainty
$\mathbf{g}$	Power domain weights
$\bar{h}_{\epsilon a}$	Nominal propagation channel amplitude gain
$\tilde{h}_{\epsilon a}$, $\tilde{\mathbf{H}}$	Uncertainty in propagation channel amplitude
κ^2	Channel speed variance
$\bar{\mathbf{m}}_\epsilon$, $\tilde{\mathbf{m}}_\epsilon$	Position vector of microphone ϵ
M	Number of microphones in the array
N_ω	Number of frequency bins
ω_k	Frequency at bin k
$\bar{\mathbf{p}}_a$, $\tilde{\mathbf{p}}_a$	Position vector of source a
P	Number of sources
ϕ_a	Head rotation distribution of source a
σ^2	Channel speed spatial variance
σ_m^2, σ_p^2	Position uncertainty variance for sources and microphones
ς	MSE optimal scaling parameter

s, $\mathbf{s}$, $\tilde{s}$	Source STFT value, source vector, desired source
$\bar{t}_{\epsilon a}$	Nominal propagation time from source a to microphone ϵ
$\tilde{t}_{\epsilon a}$, $\tilde{\mathbf{T}}$	Uncertainty in propagation time
u	Power domain beamformer output
$\mathbf{v}$	Noise STFT coefficients per microphone
v_κ	Diffuse noise
v_η	Sensor noise
$\mathbf{V}$	Noise and interference covariance
$\mathbf{w}$, $\mathbf{W}$	Linear beamformer weights
$\mathbf{x}$	Array data in the STFT domain
y	Linear beamformer output
z	Power domain beamformer output in the time domain

Chapter 1

Introduction

1.1 Research aims and motivation

The acquisition and reproduction of speech is damaged by acoustic noise and reverberation which affect both the perceptual quality and intelligibility of the signal. This thesis addresses two aspects of this problem: dereverberation and optimal beamforming that is robust to propagation delay variations and source movement.

In addition to the direct sound received from a target source, a microphone will also receive sound that has reflected off walls and other surfaces within a room. Reverberation is the name given to the perceptual effect of these reflections and is damaging to both the quality and intelligibility of a speech signal. Applications such as automatic speech recognition are severely hindered by its presence. The further the target source is from the microphone the greater the effects of reverberation, because the energy received from the source decreases with the microphone distance, whereas the reverberant energy remains approximately constant.

Single microphones are typically omnidirectional or have weak directionality and hence do not take advantage of spatial separation of the target and interfering sources. By using an array of microphones, it is possible to combine their outputs to make

a highly directional sensor. This is known as beamforming. Beamforming is a popular multichannel signal processing technique used in conjunction with microphone arrays to spatially filter a sound field. Beamformers exploit the spatial diversity of the acoustic sources in the environment in order to suppress interference and reverberation whilst amplifying the sound from a desired source location. This spatial diversity manifests itself as consistent phase and amplitude differences at the microphones that are dependent on source position. Conventional beamformers require that these differences are accurately known. The differences are either specified *a priori* or are found adaptively. However in reality, there are uncertainties in the phase and magnitude responses of the source-microphone channels and these uncertainties become larger with increasing microphone separation. In these circumstances the performance of conventional beamformers degrades.

Robust approaches to beamforming reduce this degradation. The robust beamformer designs rely on tuning parameters whose choice is not always well defined or based on the intended application of acoustic channels. By modeling the uncertainty in the acoustic channels we can derive more accurate expressions for the source-microphone channels. As such we are able to derive beamformers that are well suited to the application of acoustics in realistic environments.

Furthermore, beamformers are typically utilised in the time or the short time Fourier transform (STFT) domain. As such, when the source-microphone channels are exactly known, the beamformer performance is always limited by an upper bound. By studying the behaviour of speech signals in the power domain and the power-squared domain, we can build beamformers in higher domains to exploit this knowledge. When designing beamformers in the power domain we are not limited by the same upper bounds as the time and STFT domains, as we are able to use spectral subtraction to further reduce the effects of interference and noise.

1.2 Thesis overview

Chapter 3 introduces a detailed analysis of speech signals and the corresponding STFT coefficients. We find accurate statistical models for the STFT coefficients and show a better fit than traditional methods. We also find functions for the complex kurtosis from the second order STFT coefficient statistics and the fourth order statistics, which are needed for the beamformers introduced in later chapters.

Chapter 4 details a novel algorithm for dereverberation in the STFT domain. The approach utilises near by frames to remove reverberation. It overcomes the shortfalls of the time domain algorithms. With knowledge of the impulse response it can successfully remove any amount of reverberation.

In Chapter 5 we derive models for realistic acoustic propagation channels, which extend the traditional models to incorporate a series of uncertainties. Both time and amplitude uncertainties are formulated. We validate the models with experimental data. We also consider the impact of human head directivity patterns and random directions on the propagation channels.

In Chapter 6 we present an overview of beamformers from the previous literature. This covers both conventional beamforming and robust beamforming.

Chapter 7 details the derivation of a robust STFT domain beamformer. We incorporate the channel propagation coefficients from Chapter 5 to create a robust beamformer. By considering random array geometries we show that the performance of the robust beamformer does not degrade in the presence of channel uncertainties like other traditional beamformers. We test the beamformer in the presence of time based channel uncertainties before extending to include the head rotation model as well.

We design a novel power domain beamformer framework in Chapter 8. By extending the traditional STFT domain beamformer framework to the power domain we can observe better results. By combining multiple beamformers in the power domain, we are able to utilise spectral subtraction to further improve the traditional beamformer performance. We utilise the uncertain propagation coefficients to ensure robustness to time and amplitude uncertainties.

Finally in Chapter 9 we summarise the thesis with concluding remarks and address areas in which the work can be extended.

1.3 Thesis original contributions

The following contributions are believed to be original:

- Sections 3.3 and 3.2: Accurate modeling of speech STFT coefficients and forming the complex kurtosis of speech STFT coefficients.
- Section 4.2: A STFT domain dereverberation method.
- Sections 5.3 and 5.5: Realistic acoustic propagation channel coefficients which incorporate a time and amplitude uncertainties model.
- Chapter 7: The propagation channel uncertainties robust STFT domain beamformer.
- Section 8.2: The power domain beamformer framework.
- Section 8.7: The channel uncertainties robust power domain beamformer.

1.3.1 Publications

1. Richard Stanton and Mike Brookes, "Speech Dereverberation in the STFT Domain," to be submitted, Nov. 2013.
2. Richard Stanton and Mike Brookes, "Path Uncertainty Robust Beamforming," EUSIPCO, Sept. 2014.
3. Richard Stanton, Nikolay D Gaubitch, Patrick Naylor and Mike Brookes, "A Differentiable Approximation to Speech Intelligibility Index with Applications to Listening Enhancement," Proc. Audio Eng. Soc. (AES) Conf. on Audio Forensics, 2014. - **Awarded best paper prize.**

4. Richard Stanton and Mike Brookes, "Robust Beamforming in the Power Domain," to be submitted, IEEE Trans. Signal Process., 2016.

5. Richard Stanton and Mike Brookes, "Head Rotation Robust Beamforming," to be submitted, IEEE Trans. Signal Process., 2016.

Chapter 2

Underpinning Signal Processing Techniques

In this chapter we will briefly present the relevant signal processing techniques that are used later in the thesis.

2.1 Speech processing in the STFT domain

The short time Fourier transform (STFT) is an invertible transform from the time domain to the time-frequency domain. The STFT domain is especially useful for the real-time processing speech signals for a number of reasons. Speech is periodic in nature, which means frequency analysis is ideal for use in enhancement applications. As the STFT is based on the Fourier transform, each frequency band can be processed separately. Speech is non-stationary and the frames used in the transform ensure stationarity is achieved. In the STFT domain speech signals are normally sparse and time delays are converted into frequency-dependent phase shifts. Most of the processing in this thesis will be performed in this domain.

2.1.1 Short time Fourier transform

The STFT is summarised in Figure 2.1.

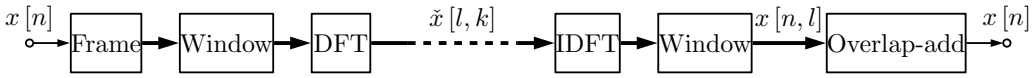

Figure 2.1: The forwards and inverse STFT process.

The time domain signal, x, is split into a sequence of frames using an analysis window, w, of length QR. The discrete Fourier transform (DFT) of each frame is taken. It is defined as follows:

$$\check{x}[l,k] \triangleq \sum_{n=0}^{QR-1} x[n+lR]\, w[n] \exp\left(-j2\pi \frac{kn}{QR}\right), \tag{2.1}$$

where l is a discrete-time frame index, k is a discrete-frequency index, Q is the overlap factor, R is the frame increment. From this point forward, signals denoted with both time and frequency indices are STFT domain signals.

The STFT is used to analyse signals, which are non-stationary, in the frequency domain. The choice of analysis window has a large impact of the STFT. It needs to be short enough that the windowed signal can be assumed quasi-stationary. A longer analysis window increases the number of frequency bins, but decreases in the number of frames to process. There is a tradeoff with time resolution against frequency resolution. The complexity remains the same. This choice can be informed by the statistics of the signals we are processing. Real time applications may prefer shorter windows, which reduces the latency of the processing system. The frames are overlapped to improve the time resolution whilst maintaining the frequency resolution.

Speech is inherently non-stationary, therefore speech processing algorithms are frequently applied in the STFT domain, where frame lengths of around 10 ms to 250 ms are often used. The STFT of a speech segment is shown in the spectrogram in Fig. 2.2. It confirms that the speech energy is sparse in the STFT domain, [2].

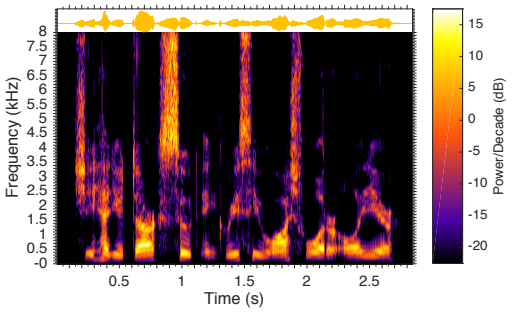

Figure 2.2: A spectrogram of a speech segment.

When considering the convolution of time domain signals, f and g, a common used identity is the convolution theorem, which states convolution in time is equivalent to multiplication of the Fourier transforms:

$$\mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\},$$

where $\mathcal{F}$ denotes the Fourier transform. In the STFT domain the convolution theorem is an approximation, and is only accurate when the length of f or g is short compared to the STFT frame length.

To reconstruct the time domain signal we first use the inverse DFT (IDFT) to reform time frames:

$$x\left[l,n\right] \quad = \quad \frac{1}{QR}\sum_{k=0}^{QR-1}\check{x}\left[l,k\right]\exp\left(j2\pi\frac{kn}{QR}\right). \tag{2.2}$$

These are then overlap-added [3] to form the time signal:

$$x[n] = \sum_{l=l_n-Q+1}^{l_n} x[l, n-lR]w[n-lR], \tag{2.3}$$

where $l_n = \left\lfloor \frac{n}{R} \right\rfloor$, in which $\lfloor x \rfloor$ denotes the smaller integer not exceeding x. Perfect reconstruction of the original time domain signal is obtained when the window used for analysis and synthesis satisfies, [4]:

$$\sum_{q=0}^{Q-1} w^2[qR + n] = 1 \qquad \forall n \in [0,\, R-1]\,.$$

This is satisfied for the case $Q = 2$ for a square-root Hamming window while for the case $Q = 4$ is satisfied with a Hamming window.

2.2 Speech evaluation metrics

The effect of speech enhancement can be broadly measured in two categories: the speech intelligibility and the speech quality. Herein speech quality is defined as the acceptability of the speech signal. Intelligibility is defined as the proportion of words that can be understood. This literature review will concentrate on speech quality.

Speech quality can be measured either subjectively or objectively. Subjective methods require the use of human listeners. Such tests include the Mean Opinion Score (MOS) [5], in which trained listeners score the signal quality from 1-5 and the results are averaged over all listeners. Alternatively, preference scoring is used, where listeners select the best signal between two or more signals. Subjective tests are commonly designed using ITU-T recommendation P.835 [6]. Subjective methods can be time intensive and non-repeatable, as such they are not commonly used to evaluate early stage research algorithms, although they are still used [7] in practice.

Objective methods have been developed to tackle these issues. They can be divided into intrusive and non-intrusive methods, where intrusive methods require the clean speech signal and non-intrusive uses just the degraded signal. In the following sections various objective speech quality metrics will be described.

In the following sections we denote the processed signal, $\hat{s}$, as an estimate for the original clean speech before degradation.

2.2.1 SNR

Signal-to-noise ratio (SNR) is a commonly used intrusive measure to evaluate the quality of signals degraded by additive noise. It is defined as the ratio of source power to noise power and is commonly expressed in decibels (dB). In the STFT domain the powers are summed over all time frames:

$$SNR\left(k\right) \quad = \quad 10\log_{10}\left(\frac{\sum_{l}\left|\check{s}\left[l,k\right]\right|^{2}}{\sum_{l}\left|\check{v}\left[l,k\right]\right|^{2}}\right).$$

It has been extended to the segmental-SNR [8], which averages the SNR computed across frames typically of 15 to 20 ms. This improves performance in the presence of non-stationary noise. If the noise is taken from the difference of the processed signal and the original speech, $\check{v}\left[l,k\right] = \check{\hat{s}}\left[l,k\right] - \check{s}\left[l,k\right]$, the processed and original signals must be both time and amplitude aligned. The SNR is computationally inexpensive but does not always correlate well with MOS [9].

2.2.2 MSE

Mean squared error (MSE) can be defined in many domains as the squared difference between the original and processed signals. In the time domain:

$$MSE \quad = \quad \frac{1}{N}\sum_{n=1}^{N}\left|\hat{s}\left[n\right] - s\left[n\right]\right|^{2},$$

where N is the time length of the signals. In the STFT domain we can define it in each frequency bin as follows:

$$MSE\left(k\right) \quad = \quad \frac{1}{L}\sum_{l=1}^{L}\left|\check{\hat{s}}\left(l,k\right) - \check{s}\left(l,k\right)\right|^{2}. \tag{2.4}$$

Similarly to the SNR, both the original and degraded speech need to be aligned in time and amplitude. Due to its squared based utility function, the MSE is sensitive to outliers. MSE is often used as a cost function in optimisation algorithms due to its mathematical convenience.

2.2.3 BSD

Bark Spectral Distortion (BSD) [10] is based on the psychoacoustic idea that speech quality relates to speech loudness. Across all voiced regions of the speech, it maps loudness vectors from the original speech to the degraded signal. The average Euclidean distance between the loudness vectors is taken as the BSD. By processing the signal spectra it is more robust to time alignment errors in the two signals. It was extended to the modified BSD (mBSD) which considers noise masking thresholds. It has been shown to correlate well with subjective speech quality [11].

2.2.4 PESQ

Perceptual Evaluation of Speech Quality (PESQ) is defined in ITU-T P.862 [12]. It is a perceptually motivated measure of speech quality and varies between -0.5 and 4.5. The scores can be mapped to the subjective MOS values [13]. PESQ has been shown to be highly correlated with subjective quality scores than measures such as SNR [14, 9].

PESQ has been extended to the Perceptual Objective Listening Quality Analysis (POLQA) in ITU-T P.863 [15], which enables higher bandwidths.

2.2.5 STOI

A short-time objective intelligibility (STOI) metric [16, 17] was derived as a method to estimate intelligibility in situations where noisy speech is processed using a time-frequency varying gain function. Correlation based intelligibility methods which rely

on long-term averages can be dominated by a few regions of high amplitude. Methods based on very short time frames (20-30 ms) reduce the usefulness of low temporal modulations which are important for speech intelligibility. Therefore rather than using long-term average statistics, STOI uses short time frames of 386 ms. The correlation between temporal envelopes of the clean and degraded speech are computed over all the frames, from which the STOI is calculated. It was evaluated on three listening tests and showed high correlation with intelligibility. STOI can be applied to non-linear changes in the speech signal, many of the previous methods are confined to linear filtering.

2.2.6 Metric time alignment

A sample shift was applied to one of the input signals to measure the effect of time alignment issues on the various metrics mentioned above, with a sampling frequency of 16 kHz. In this case the estimated clean signal is the clean signal shifted by a value t:

$$\hat{s}\left[n\right] \;=\; s\left[n-t\right].$$

The change in metrics against sample shift is shown in Fig. (2.3). BSD, PESQ and STOI do not experience any change when the time alignment is not correct. However the STFT based MSE, (2.4), is not robust to time alignment issues, there is a large increase in MSE for all non-zero sample shifts. Therefore when using the MSE we need to ensure the signals are coorectly time aligned.

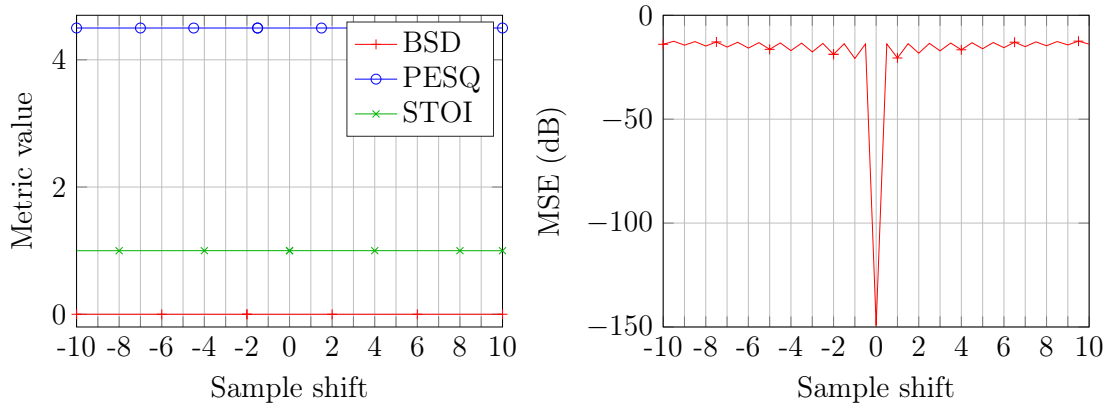

Figure 2.3: The value of several metrics against a sample shift in one of the input signals.

2.2.7 Metric amplitude alignment

An amplitude scaling was applied to the reference signal to investigate the effect of amplitude mis-alignment on the various metrics:

$$\hat{s}\left[n\right] \quad = \quad \alpha s\left[n\right].$$

The change in metrics against the amplitude scaling factor is shown in Fig. (2.4). Whilst BSD is robust to time alignment errors, it is not robust to amplitude mis-alignments. Both PESQ and STOI are robust to both alignments. Similarly to the time alignment, MSE is not robust to amplitude mis-alignment. Based on these findings PESQ and STOI seem the most reliable metrics to evaluate the results of the work of this thesis.

2.2.8 A-weighting

Metrics such as SNR and MSE can be computed across each frequency bin separately. For an equivalent metric over wideband signals we can apply a weighted average over all frequencies. The weight for each frequency is taken from the A-weighting curve as

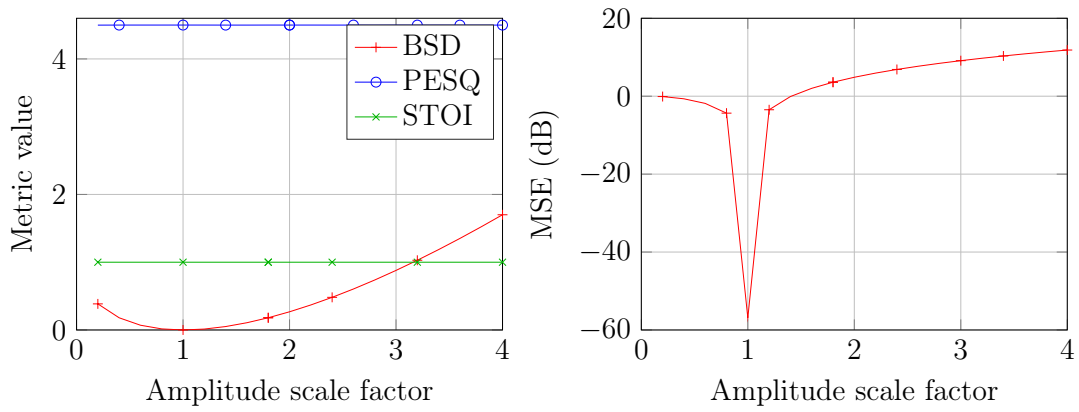

Figure 2.4: The value of several metrics against an amplitude change in one of the input signals.

defined within [18] and is shown in Fig. 2.5. It is perceptually defined to be reflect the relative sensitivity of the human ear which increases in mid-band frequencies around 2 kHz.

2.3 Databases

Speech enhancement algorithms are typically evaluated using metrics, including those outlined above, applied to a large number speech segments. Running algorithms over a large set of speech segments gives a better representation of the algorithm's performance over unseen speech and real world performance.

2.3.1 TIMIT

The TIMIT dataset [19] is a collection of speech files. It comprises 630 English talkers over eight different American dialects, of both sexes. Each talker reads 10 segments, which are phonetically balanced. The dataset is divided into a training and a test set.

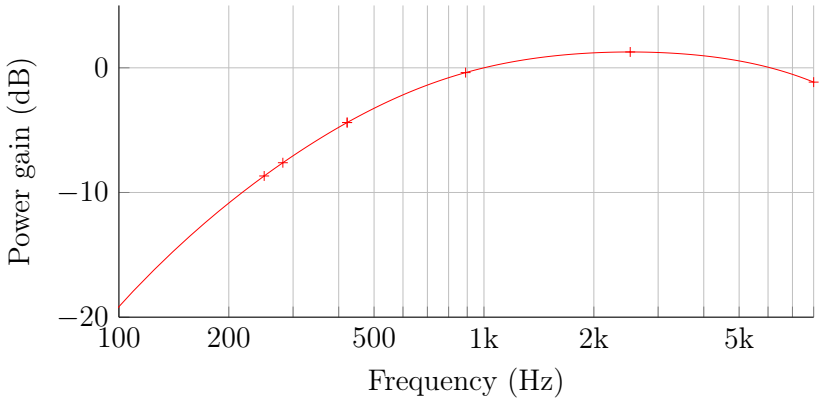

Figure 2.5: The power gain of the A-Weighting curve.

The dataset is typically used to evaluate speech enhancement systems. They are also used in intelligibility tests; the segments are unpredictable which means language models and context do not offer assistance in recognition.

For the remainder of the thesis, except Chapter 4, the signals are processed in the STFT domain, as such the $\check{x}$ notation is dropped.

Chapter 3

Speech STFT Coefficient Properties

Most of the algorithms discussed in later chapters of this thesis process speech signals in the STFT domain as defined in Sec. 2.1.1. In this chapter we present empirical models for three statistical properties of speech: average power spectrum, STFT coefficient distribution and STFT coefficient kurtosis.

3.1 Average power spectrum

In Chapter 7 we design beamformers that require the expected source power, $\langle |s_\omega|^2 \rangle$, at all frequencies, ω. If we assume the sources are human talkers, we can utilise models of human speech in order to design the beamformers.

The long-term average speech spectrum (LTASS) [20] is commonly used for the expected speech power of human sources. It is measured across many English talkers through third octave filter banks in which the output is smoothed over time. It is similar across different languages. The LTASS is shown in Fig. 3.1.

The TIMIT dataset [19] is a collection of speech files as discussed in Section 2.3.1. Throughout the remainder of this work, the TIMIT core test set will be used. It consists of 24 talkers and 240 sentences. LTASS will be used for the expected speech

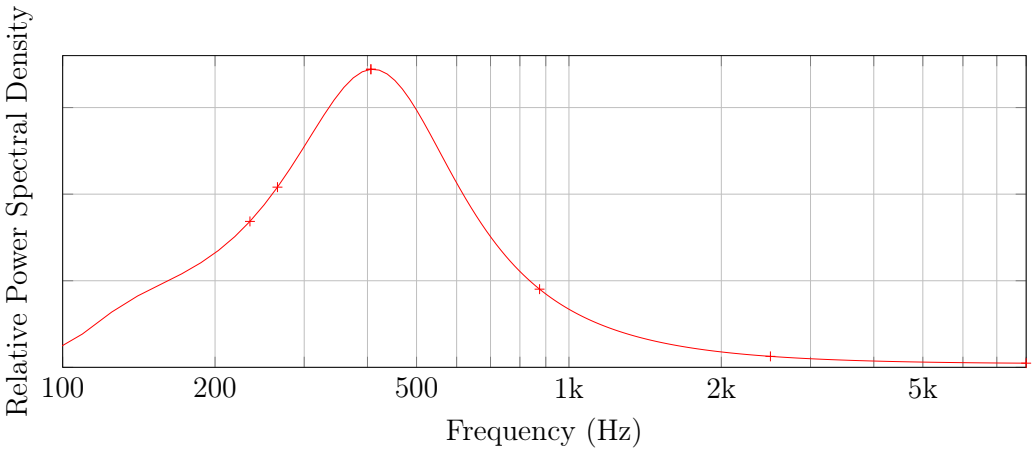

Figure 3.1: The average speech power against frequency as denoted by LTASS. The speech power is relative to a sound pressure of 20 μPa at 1m from the lips.

power to design the beamformers, whereas the signals which are processed in each algorithm will be taken from TIMIT. In order to ensure the LTASS statistics match the TIMIT speech segments, the LTASS power spectrum was scaled to minimise the MSE with the average TIMIT segment power spectrum.

To calculate the average power from the TIMIT segments, they were first normalised in order to have an active speech level of 0 dB according to ITU-T P.56 [21]. Then they were transformed into the STFT domain using 64 ms frames and a square root Hamming window. The power per frequency bin was averaged over all frames. The respective powers are shown in Fig. 3.2. The peak LTASS power occurs at a lower frequency compared with TIMIT, but they share a similar shape.

3.2 STFT coefficient distribution

The STFT coefficients of speech can be modeled as random variables which follow a probability density function (PDF). The accuracy of the PDF is important for various speech modeling approaches [22]. It is commonly assumed that the real and imaginary parts of the STFT coefficients follow independent Gaussian distributions

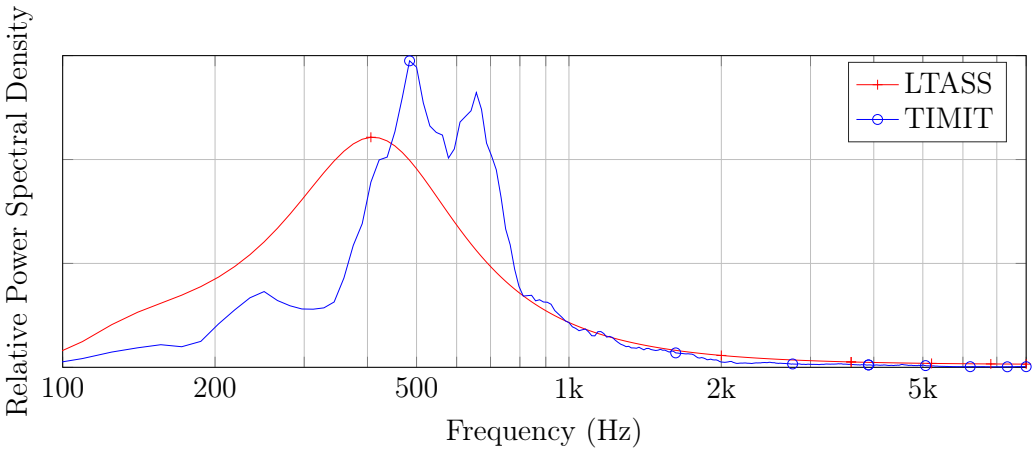

Figure 3.2: The average speech power against frequency as denoted by LTASS and the TIMIT speech database. The speech power is relative to a sound pressure of 20 μPa at 1m from the lips.

[23, 24, 25, 26, 27]. This approach is usually validated through the use of the central limit theorem. When the STFT frame is very long, the coefficients converge to a Gaussian distribution. However in practice frame sizes are much too small for this approximation to be valid. In this Section we investigate different distributions for the modeling of speech STFT coefficients.

Over the 40,000 TIMIT STFT frames, we plot a histogram of the real part of the coefficients corresponding to 400 Hz, this is shown in Fig. 3.3. The y-axis is on a log scale. To avoid noise mis-shaping the histogram, a frequency dependent voice activity detector (VAD) was used to filter out frames where speech energy is not present [21]. The commonly used Gaussian distribution has been fitted to the data, through the use of moment matching. The equivalent histogram when processing white Gaussian noise is shown in Fig. 3.4. It is clear that from Fig. 3.4 that the STFT of Gaussian noise is modeled well by a Gaussian distribution. However, the shape of the histogram in Fig. 3.3 shows that the speech STFT coefficients do not fit a Gaussian model. Applying the Shapiro-Wilk normality test to the speech STFT coefficients, shows that the probability of a Gaussian distribution generating the histogram data has $p < 10^{-7}$. The histogram shown in Fig. 3.3 is much closer to a Laplace distribution,

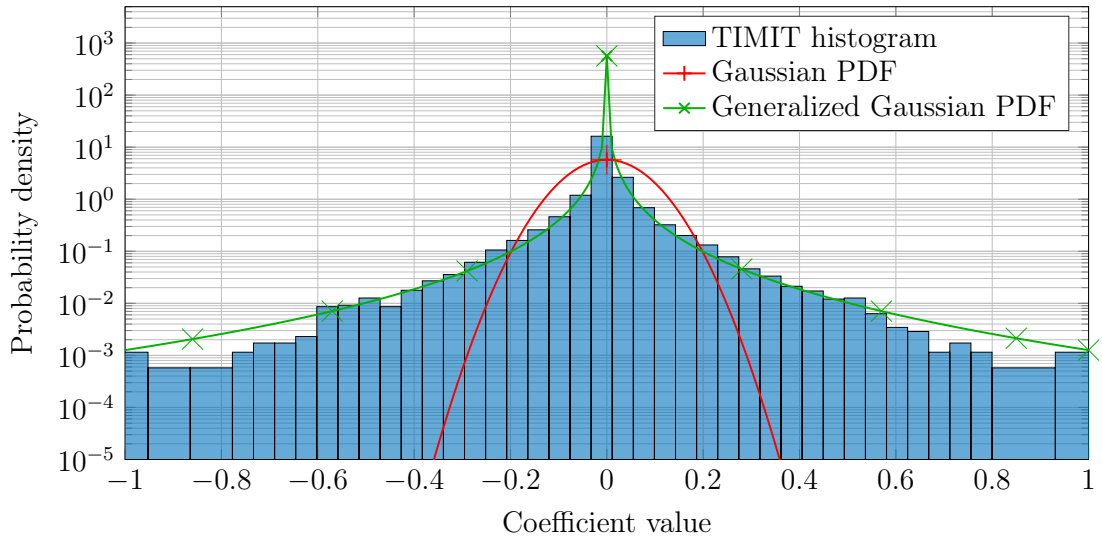

Figure 3.3: Histogram of the real part of the STFT coefficients of speech at 400 Hz across all TIMIT frames. Gaussian ($\sigma = 0.005$) and generalized Gaussian ($\alpha = 4e-5$, $\beta = 0.25$) distributions have been fitted to the histogram with Jensen-Shannon divergences of 0.13 and 0.006 respectively.

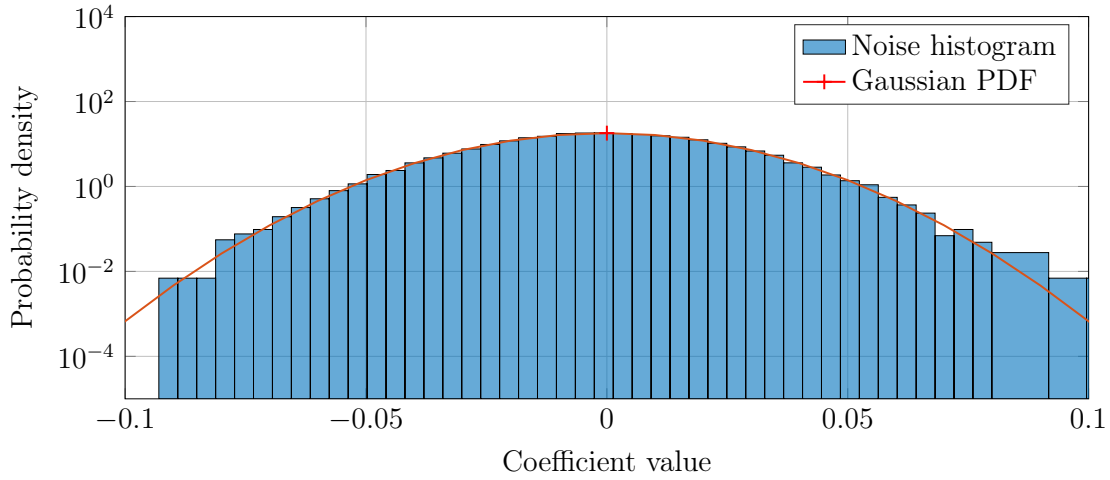

Figure 3.4: Histogram of the real part of the STFT coefficients of Gaussian noise at 400 Hz across all noise frames, with a fitted Gaussian distribution overlaid. Jensen-Shannon divergence of $2 \cdot 10^{-4}$.

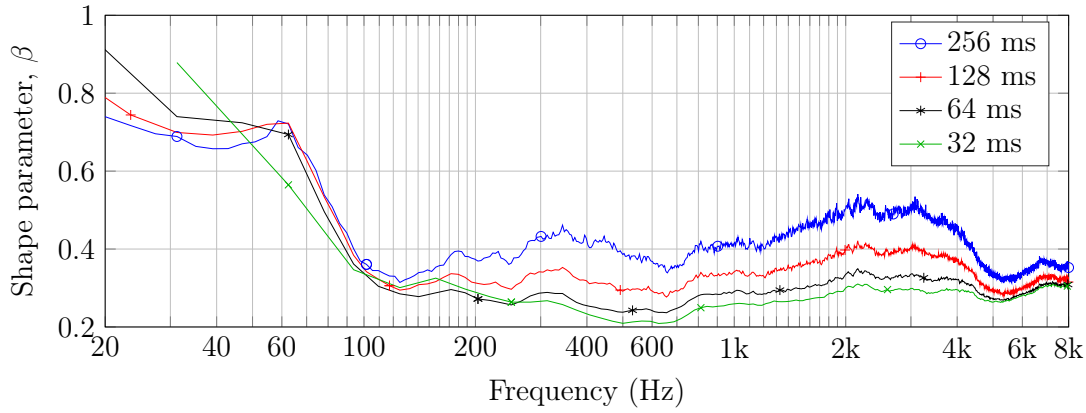

Figure 3.5: The Generalized Gaussian shape parameter against frequency for different analysis window length.

which has been proposed by [22]. Both the Laplace and the Gaussian distribution are a subset of the Generalized Gaussian distribution, whose PDF is given as:

$$p\left(x\right) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-(|x-\mu|/\alpha)^{\beta}\right),$$

where β is a shaping parameter. The shaping parameter determines the kurtosis of the distribution, $\beta = 2$ corresponds to a Gaussian and $\beta = 1$ corresponds to a Laplace distribution. The log axis of the histograms shown previously accentuates the exponential term of the PDF, $-(|x-\mu|/\alpha)^{\beta}$. The shape parameter determines the concavity of the histogram. For example, a linear line, $\beta = 1$, in the log-histogram will correspond to a Laplace distribution.

We can optimally fit a Generalized Gaussian distribution to the STFT coefficients at each frequency, through an iterative Newton Raphson approach to maximum likelihood estimation [28, 29, 30]. For each frequency we fit a Generalized Gaussian to the TIMIT STFT coefficients. One example, for 400 Hz, is shown in Fig. 3.3, which shows a close fit to the TIMIT histogram. The resulting shape parameter, β, for all frequencies and different analysis window lengths is shown in Fig. 3.5. The shape

parameter varies with frequency. Lower frequencies are closer to a Laplace distribution, whereas anything higher than 100 Hz are more leptokurtic. The results show that the Gaussian distribution, with $\beta = 2$, is a poor fit for speech STFT coefficients. The median Jensen-Shannon divergence across all frequencies for a Gaussian model is 0.1, where as the median for the Generalized Gaussian model is 0.01. As the analysis window length shortens, the distribution becomes more leptokurtic. Fewer time samples are contributing to the STFT coefficient, this means that the effect of the central limit theorem is reduced and the distribution is further from Gaussian. If a single distribution were applied across all frequency bins, the shape parameter would be between 0.3 and 0.4.

3.3 STFT coefficient complex kurtosis

In Chapter 8 we require the expected 4th order moments of the speech source signals, $\left\langle |s_\omega|^4 \right\rangle$. Unlike the second order moments there is not a standardised spectrum for 4th order moments. We describe the relationship between the 2nd and 4th order moments with the complex kurtosis, a dimensionless quantity defined by

$$\alpha_\omega = \frac{\left\langle |s_\omega|^4 \right\rangle}{\left\langle |s_\omega|^2 \right\rangle^2}.$$

The frequency index, ω, has been dropped from the remaining section for brevity. The complex kurtosis is computed independently for each frequency bin. Whilst the above function is similar in nature to kurtosis, $\left(\frac{\left\langle |x|^4 \right\rangle}{\left\langle |x|^2 \right\rangle^2} \right)$, the speech STFT coefficients are complex

$$s = \mathbb{R}(s) + j\mathbb{I}(s) = x + jy,$$

where x and y are commonly assumed to be identically, independently distributed and have the same variance. This is confirmed in Fig. 3.6, where the covariance of the real and imaginary parts, $\langle \mathbb{R}(s)\, \mathbb{I}(s) \rangle$, is small compared to the variance of the

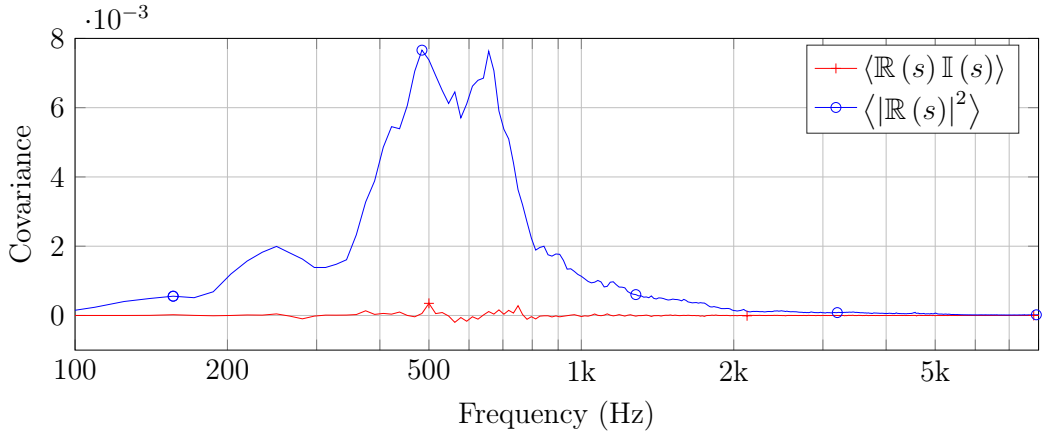

Figure 3.6: The covariance of the real and imaginary parts of speech DFT coefficients compared to the variance of the real part. The vertical axis represents power spectral density in arbitrary units.

real part, $\langle |\mathbb{R}(s)|^2 \rangle$. As such we need to derive the complex kurtosis in terms of the moments of the underlying real and imaginary distributions. We can expand the speech moments in terms of x and y:

$$\begin{aligned} \langle |s|^4 \rangle &= \langle |x|^4 \rangle + 2 \langle |x|^2 \rangle \langle |y|^2 \rangle + \langle |y|^4 \rangle \\ \langle |s|^2 \rangle &= \langle |x|^2 \rangle + \langle |y|^2 \rangle . \end{aligned}$$

The complex kurtosis can be expanded as:

$$\alpha_\omega = \frac{\langle |s|^4 \rangle}{\langle |s|^2 \rangle^2} = \frac{\langle |x|^4 \rangle + 2 \langle |x|^2 \rangle \langle |y|^2 \rangle + \langle |y|^4 \rangle}{\langle |x|^2 \rangle^2 + 2 \langle |x|^2 \rangle \langle |y|^2 \rangle + \langle |y|^2 \rangle^2}.$$

Thus to compute the complex kurtosis we require the 2nd and 4th moments of the real and imaginary distributions.

If we assume the underlying distributions are Gaussian, the complex kurtosis returns a constant:

$$\begin{aligned} \left\langle |x|^4 \right\rangle &= 3\sigma^4 \\ \left\langle |x|^2 \right\rangle &= \sigma^2 \\ \alpha_\omega &= 2, \end{aligned}$$

where σ is the standard deviation of the distribution.

If we assume an underlying Laplace distribution, α_ω increases:

$$\begin{aligned} \left\langle |x|^4 \right\rangle &= 24b^4 \\ \left\langle |x|^2 \right\rangle &= 2b^2 \\ \alpha_\omega &= \frac{7}{2}, \end{aligned}$$

where b is the scale of the Laplace distribution.

However, using a Generalized Gaussian distribution offers more flexibility. The complex kurtosis is found to be a function of the distribution shape parameter, the moments are taken from [31, 32]:

$$\begin{aligned} \left\langle |x|^4 \right\rangle &= \frac{\sigma^4 \Gamma\left(\frac{5}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} \\ \left\langle |x|^2 \right\rangle &= \frac{\sigma^2 \Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} \\ \alpha_\omega &= \frac{\Gamma\left(\frac{1}{\beta}\right)\Gamma\left(\frac{5}{\beta}\right) + \Gamma\left(\frac{3}{\beta}\right)^2}{2\Gamma\left(\frac{3}{\beta}\right)^2}, \end{aligned} \tag{3.1}$$

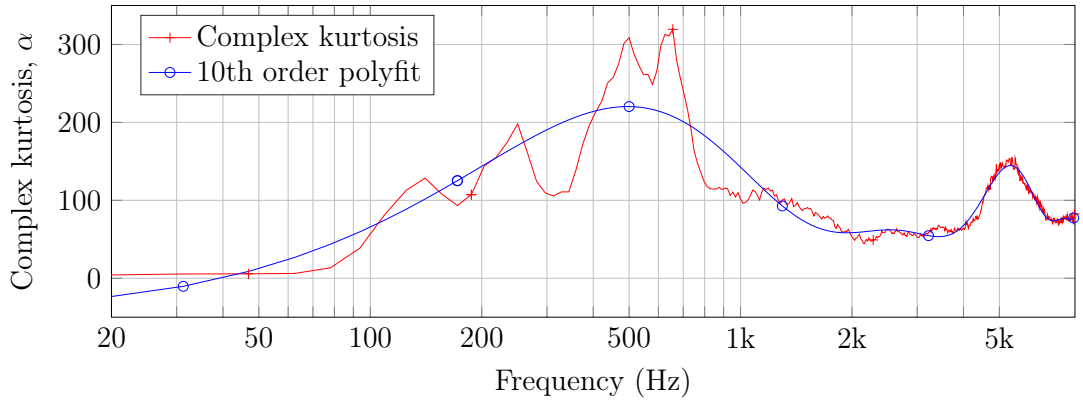

Figure 3.7: The complex kurtosis taken from (3.1), with a 10th order polynomial fit.

where $\Gamma(\cdot)$ is the gamma function. The above formulation confirms the kurtosis values for the Laplace and Gaussian when substituting for $\beta = 1$ or $\beta = 2$ respectively. It also shows that the fourth order moments can be defined as a function of the second order moments.

3.3.1 TIMIT analysis

Using the shape parameters estimated from TIMIT in Sec. 3.2 we can find the corresponding complex kurtosis from (3.1). To smooth the function a 10th order polynomial was fitted to the complex kurtosis and is shown along with the complex kurtosis in Fig. 3.7. The polynomial is used to generate the complex kurtosis at any frequency.

Alternatively the complex kurtosis can be found from the TIMIT average powers and the average squared powers:

$$\left\langle |s_\omega|^4 \right\rangle = \alpha_\omega \left\langle |s_\omega|^2 \right\rangle^2 . \tag{3.2}$$

For each normalised TIMIT segment we use a simple VAD based on [21] to remove silences from the beginning and end. We take the STFT and compute the power and the squared power at each frequency. The expectation of the complex kurtosis, α_ω,

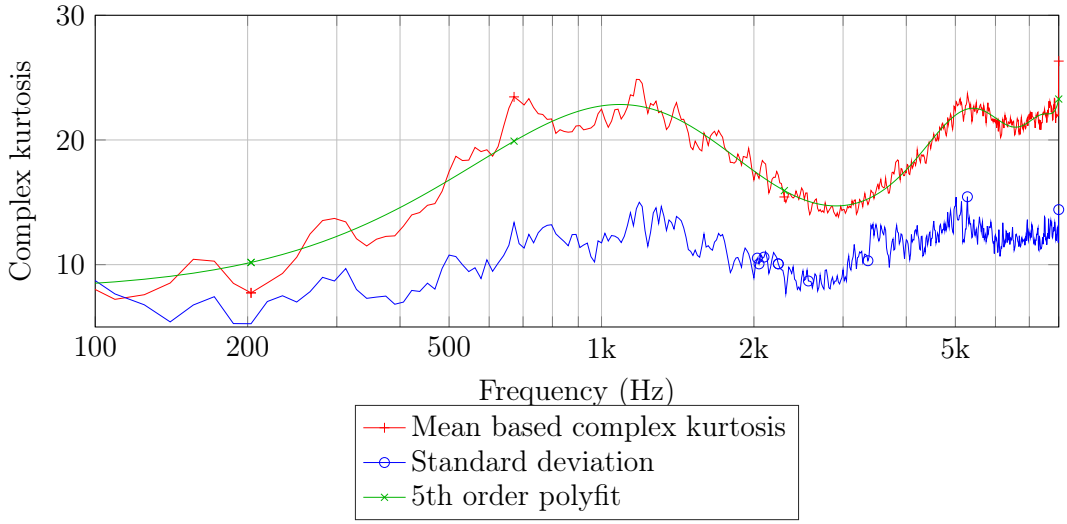

Figure 3.8: The complex kurtosis based on the mean of TIMIT segments, with the standard deviation and a 5th order polynomial fit.

is found for each segment by averaging over all the frames of that segment. Over all segments the average complex kurtosis is found. The complex kurtosis is smoothed using a 5th order polynomial, and is shown in Fig. 3.8. The result follows a similar shape to the complex kurtosis derived from the Generalized Gaussian shape parameter shown in Fig. 3.7, but reduced in value.

If we take the expected power, $\langle |s_\omega|^2 \rangle$, from LTASS, we can apply the complex kurtosis to estimate the effective speech power squared spectrum using the two different complex kurtosis functions. The resulting power squared spectra are compared with the average speech power squared from the TIMIT dataset as shown in Fig. 3.9. The Generalized Gaussian shape based complex kurtosis over estimates the power squared compared with TIMIT, whereas the mean based complex kurtosis underestimates. The generalized Gaussian fit from Fig. 3.3 estimates well up to 3σ; however, at 4σ it over estimates the larger coefficients. This will cause the complex kurtosis to over estimate. However, later in Chapter 8 we will show that this estimation error is not critical to the performance of the beamformers with which it is used with.

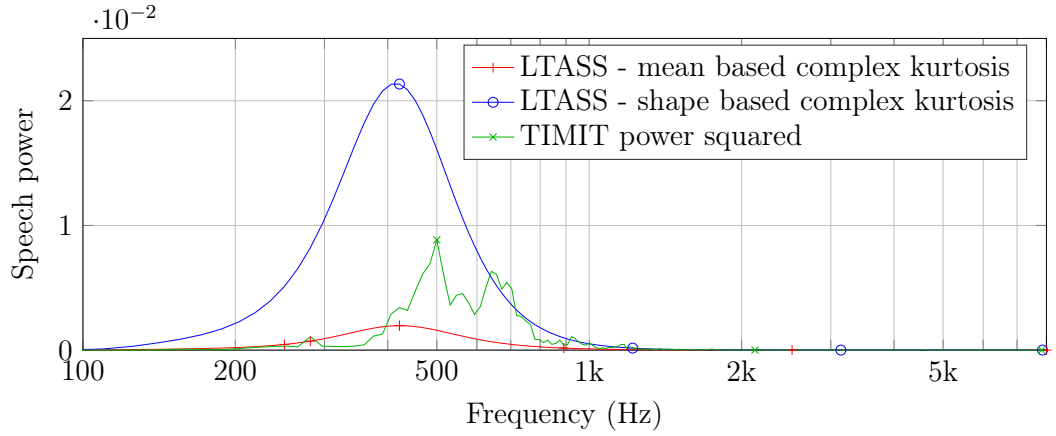

Figure 3.9: The complex kurtosis applied to LTASS compared with the mean speech power squared of the TIMIT dataset.

3.4 Conclusions

In this chapter we have shown that the Generalised Gaussian distribution is better fit for the STFT coefficient of speech signals than that of the Gaussian distribution. At each frequency the parameters of the Generalised Gaussian distribution was trained on the TIMIT dataset. We use the moments of the distribution to form a complex kurtosis from the expected speech power and the expected speech power squared and compare this to an equivalent complex kurtosis formed from the expected TIMIT powers.

The complex kurtosis can be used to construct speech algorithms that are based on the fourth order statistics, such as the power domain beamformer designed in Chapter 8.

Chapter 4

Dereverberation in the STFT Domain

Reverberation is damaging to both the quality and the intelligibility of a speech signal. In this chapter, we propose a novel single-channel method of dereverberation based on a linear filter in the Short Time Fourier Transform domain. Each enhanced frame is constructed from a linear sum of nearby frames based on the channel impulse response. The results show that the method can resolve any reverberant signal with knowledge of the impulse response to a non-reverberant signal.

In this chapter $h\,[n]$ refers to the impulse response of a channel, in subsequent chapters h refers to the channel amplitude uncertainties.

4.1 Introduction

Reverberation occurs from multi-path propagation of an acoustic signal, $s[n]$, through a channel with impulse response $h[n]$ to a microphone. The further the target source is from the microphone the greater the effects of reverberation, because the direct path energy received by the microphone falls with the square of the distance while the reverberant energy remains approximately constant. A typical sampled anechoic

and reverberant impulse response are shown in Fig. 4.1, the sampling frequency is 16 kHz.

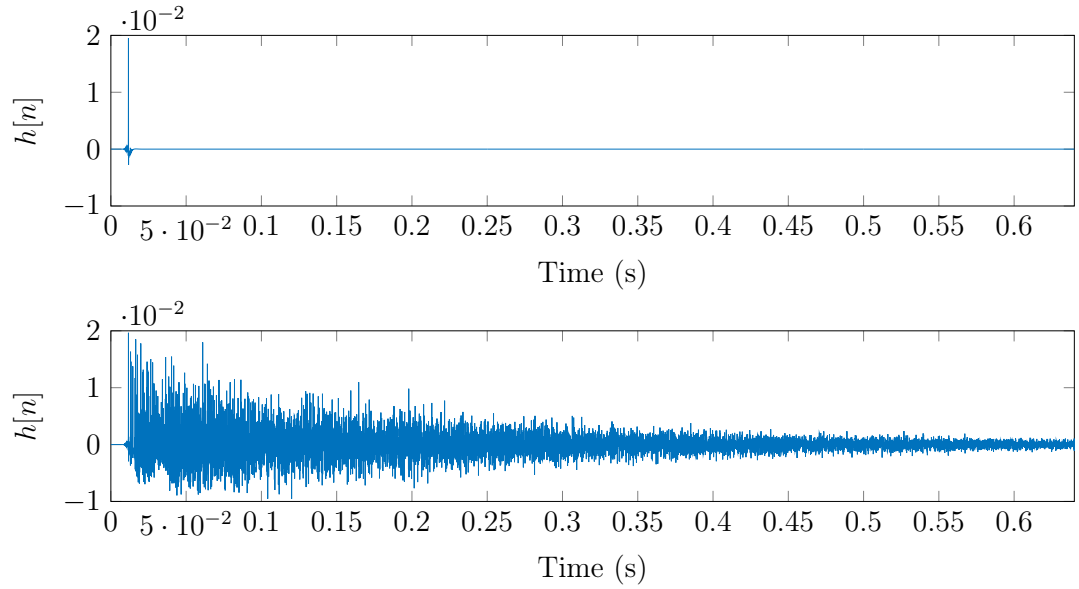

Figure 4.1: A typical anechoic impulse response (top) and a reverberant impulse response (bottom), generated using [1].

Reverberation causes speech to sound distant and spectrally distorted and, especially when combined with acoustic noise, reduces intelligibility [33]. It is noted that early reflections, which occur within 50 ms of the direct path, can benefit intelligibility, whilst later reflections damage intelligibility [34]. Automatic speech recognition is severely hindered by reverberation [35, 36]. Therefore channel inversion methods are of high importance in spatial filtering fields.

Dereverberation methods can be divided into those that require knowledge of the impulse response and those that do not. Within the approaches that do not require the true impulse response, spectral subtraction has been used to estimate the power spectrum of the late reverberation and subtract this from the current spectrum to leave the direct path, [37]; this approach was extended in [38] to introduce the frequency dependence of the reverberation time. The approach in [39] uses multi-step forward linear prediction to estimate the late reverberation tail before using spectral

subtraction to remove it. The approach from [40] estimates the fundamental frequency and the harmonic structure of the target speech in order to find a dereverberation operator.

Other methods of dereverberation exist which utilise knowledge of the system impulse response, $h[n]$. In recent years techniques have been developed to blindly estimate the impulse response, as described in [41, 42]. Least squares has previously been used to create an inverse filter from knowledge of the impulse response, [43]. The method is implemented in the time domain, as such it cannot always produce a minimum phase filter, which means perfect dereverberation is not always possible. This approach was extended into the multichannel domain with the Multiple-input/output INverse Theorem (MINT), [44]. It uses multiple transmission channels to form an inverse for the desired source. As long as there are no common zeros present in all the channels, it is capable of finding exact inverse filters. MINT has been extended to improve robustness through regularisation [45]. Perceptually motivated extensions focused on reducing only the late reverberation tails which hinder intelligibility [46]. Similarly channel shortening has been used to reduce the effects of late reverberation [47, 48, 49].

Furthermore beamformers have been applied to dereverberation to attenuate reflections [50, 51]. The extensions of the time domain inverse filter [43] are typically multichannel approaches. In this contribution we will concentrate on single channel methods that use knowledge of the impulse response.

Speech signals are commonly processed in the STFT domain due to its non-stationarity nature and its sparsity in the STFT domain. We wish to create a single channel dereverberation algorithm, which utilises knowledge of the impulse response, $h[n]$, transformed into the STFT domain, in order to take advantage of the benefits of the STFT domain. However simply creating an inverse filter in the STFT domain is not straightforward, as the STFT process is time-variant. We present a single-channel method of dereverberation based on a linear filter which combines nearby frames which uses a novel method to account for the time varying nature of the STFT domain. The frames are linearly combined using coefficients computed through a least squares based method on the impulse response.

4.2 STFT-domain dereverberation

In discrete time, the observed reverberant signal, $y[n]$, at the microphone is the convolution of the source signal, $s[n]$, and the channel impulse response, $h[n]$:

$$y[n] = \sum_{m=0}^{M-1} h[m]s[n-m], \tag{4.1}$$

where M is the length of the impulse response, h. Exploiting knowledge of the channel impulse response, we propose a new method to reduce the effects of reverberation on $y[n]$, to form an estimate, $\hat{s}[n]$, of the original signal.

The reverberant signal, $y[n]$, is transformed into the STFT domain, $\check{y}[l,k]$, using (2.1), as shown in Sec. 2.1.1:

$$\check{y}[l,k] = \text{STFT}\left(y[n]\right).$$

The enhanced signal is formed through a linear sum of nearby frames of the reverberant signal:

$$\check{\hat{s}}[l,k] = \sum_{r=-A}^{B} g_k[r]\check{y}[l-r,k], \tag{4.2}$$

where A is the number of future frames and B is the number of past frames to be used in the enhancement and $g_k[r]$ are the linear weights for the k-th frequency bin. The resulting frames are then transferred back into time frames using the inverse STFT from (2.2) and (2.3):

$$\hat{s}[n] = \text{ISTFT}\left(\check{\hat{s}}[l,k]\right).$$

Perfect reconstruction of the reverberant signal, $\hat{s}[n] = y[n]$, is obtained with the coefficients $g_k[r] = \delta[r]$, where δ is the Kronecker delta function. In order to process

the reverberant signals in the STFT domain we assume that the length of the impulse response is short compared to the STFT frame length.

4.3 Optimal coefficients

Assuming that $h[n]$ is known, our goal is to determine the filter coefficients $\mathbf{g}_k = \begin{bmatrix} g_k[-A] & \dots & g_k[B] \end{bmatrix}^T$ so that we have perfect reconstruction of the original clean speech: $\hat{s}[n] \approx s[n]$.

Consider the response of (4.2) when the input signal is an impulse at sample λ:

$$s^{(\lambda)}[n] = \delta[n - \lambda], \quad \lambda \in [0, R-1],$$

where R is the frame increment. Only a frame increment is used as when using overlapping frames the relationship is repeated.

When processing in the STFT domain, the earliest output frame that is affected by the impulse occurs at $l_{min} = 1 - Q - A$, whereas the latest frame affected is $l_{max} = 1 + B + \left\lfloor \frac{M+\lambda-2}{R} \right\rfloor$. Applying the process from (4.2) we can find a relationship between the channel STFT of the impulse response, $H^{(\lambda)}[l,k]$, and the desired impulse response $\tilde{H}^{(\lambda)}[l,k]$. In this case the desired impulse response is the STFT of the direct path impulse response, when there are no reflections present.

We determine $\mathbf{g}_k$ to minimise the difference between the two. So for each frequency bin, k, we have an overdetermined set of equations:

$$\begin{aligned} \hat{H}^{(\lambda)}[l,k;\mathbf{g_k}] =& \sum_{r=A}^{B} g_k[r] H^{(\lambda)}[l-r,k] \\ \approx& \tilde{H}^{(\lambda)}[l,k], \end{aligned} \tag{4.3}$$

for each $\lambda \in [0 : R-1]$ and $l \in [l_{min} : l_{max}]$. This gives us $(2 + A + B + Q)\, R + M - 1$ equations, with $A + B + 1$ unknowns. This process is shown in Fig. 4.2. We combine

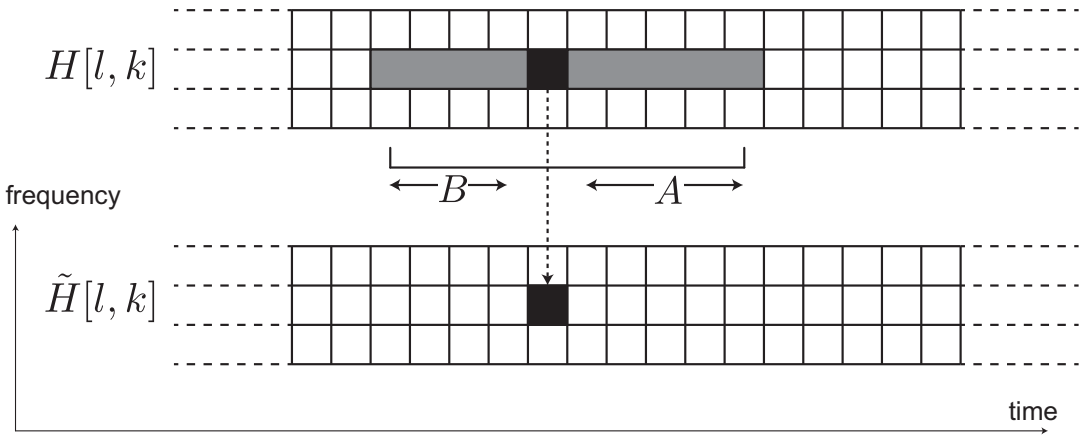

Figure 4.2: The above plots show the STFT of both $H[l,k]$ and $\tilde{H}[l,k]$. For each frequency bin the filter linearly combines future and past frames of $H[l,k]$ to best match $\tilde{H}[l,k]$.

B past frames with A future frames to best approximate the current frame from the desired impulse response.

The optimal $\mathbf{g}_k$ is found by minimising the mean squared error of the desired impulse response, $\tilde{H}_\lambda[l,k]$, and the estimated impulse response, $\hat{H}^{(\lambda)}[l,k;\mathbf{g}_k]$.

$$\mathbf{g}_k = \arg\min_{\mathbf{g}_k} \sum_{\lambda=0}^{R-1} \sum_{l=l_{min}}^{l_{max}} \left(\hat{H}^{(\lambda)}[l,k;\mathbf{g_k}] - \tilde{H}^{(\lambda)}[l,k] \right)^2. \tag{4.4}$$

We solve the above equation using linear least squares, [52]. Alternatively, the above summations can be summarised as:

$$\mathbf{g}_k = \arg\min_{\mathbf{g}_k} \left(\mathbf{g}_k^T \mathbf{H}_k - \tilde{\mathbf{H}}_k \right)^2,$$

where the least squares solution is:

$$\mathbf{g}_k = \left(\mathbf{H}_k^H \mathbf{H}_k \right)^{-1} \mathbf{H}_k^H \tilde{\mathbf{H}}_k.$$

The overall impulse response of the computed channel is time-variant, as the position of each sample, λ, produces a different optimal impulse response. We can determine an average channel response as the inverse STFT of:

$$\hat{H}\left[l,k\right] = \frac{1}{R}\sum_{\lambda=0}^{R-1}\hat{H}^{(\lambda)}[l,k;\mathbf{g}_k]\exp\left(j2\pi k\frac{\lambda}{QR}\right), \tag{4.5}$$

where a phase shift is applied to correspond with the sample position within the frame.

4.3.1 Time domain error bound

The weights are formulated in the STFT domain, in the following section we will detail a corresponding bound on the error in the time domain. The above minimisation problem minimises the reverberation present in the enhanced signal.

The target response of the dereveration process to an impulse at time λ is $\tilde{h}^{(\lambda)}[n] = \delta[n-\lambda]$ and the actual response is $\hat{h}^{(\lambda)}[n]$. The error in the time domain impulse response is $h_e^{(\lambda)}[n] = \tilde{h}^{(\lambda)}[n] - \hat{h}^{(\lambda)}[n]$. In the STFT domain, $H_e^{(\lambda)}[l,k]$ is the DFT of $h_e^{(\lambda)}[l, qR+r]$. Using Parseval's theorem the energy of the STFT domain error is equivalent to the framed time domain error:

$$\sum_{q=0}^{Q-1}\sum_{r=0}^{R-1}\left(h_e^{(\lambda)}[l,qR+r]\right)^2 = \frac{1}{QR}\sum_{k=0}^{QR-1}\left|H_e^{(\lambda)}[l,k]\right|^2. \tag{4.6}$$

The total energy over all frames of the time domain impulse response error is calculated as follows:

$$\sum_{l}\sum_{r=0}^{R-1}\left(h_e^{(\lambda)}[lR+r]\right)^2 \;=\; \sum_{l}\sum_{r=0}^{R-1}\left(\sum_{q=0}^{Q-1}w[qR+r]h_e^{(\lambda)}[l-q,qR+r]\right)^2$$

using the Cauchy-Schwartz inequality:

$$\begin{aligned}&\textstyle\sum_l \sum_{r=0}^{R-1} \left(\sum_{q=0}^{Q-1} w[qR+r] h_e^{(\lambda)}[l-q, qR+r]\right)^2 \\ \leq &\textstyle\sum_l \sum_{r=0}^{R-1} \left(\sum_{q=0}^{Q-1} w^2[qR+r] \sum_{q=0}^{Q-1} \left(h_e^{(\lambda)}[l-q, qR+r]\right)^2\right),\end{aligned}$$

from the previous definition of the window, $\sum_{q=0}^{Q-1} w^2[qR+r] = 1$, we can remove the window contribution:

$$\begin{aligned}\sum_l \sum_{r=0}^{R-1} \left(\sum_{q=0}^{Q-1} w^2[qR+r] \sum_{q=0}^{Q-1} \left(h_e^{(\lambda)}[l-q, qR+r]\right)^2\right) &= \sum_l \sum_{r=0}^{R-1} \sum_{q=0}^{Q-1} \left(h_e^{(\lambda)}[l-q, qR+r]\right)^2 \\ &= \frac{1}{QR} \sum_l \sum_{k=0}^{QR-1} \left|H_e^{(\lambda)}[l,k]\right|^2.\end{aligned}$$

Therefore a bound on the time domain error exists which is limited by the STFT domain error:

$$\sum_l \sum_{r=0}^{R-1} \left(h_e^{(\lambda)}[lR+r]\right)^2 \leq \frac{1}{QR} \sum_l \sum_{k=0}^{QR-1} \left|H_e^{(\lambda)}[l,k]\right|^2.$$

The above is shown for a single λ. The bound is extended to all $\lambda \in [0, R-1]$ in a similar fashion and averaging both sides over λ, where the $R-1$ limit is used to avoid summing every sample Q times.

4.4 Evaluation

To evaluate the reduction in reverberation, we use two metrics: the Direct-to-Reverberant Ratio (DRR) [53] and the Signal-to-Reverberation Ratio (SRR) [8]. To evaluate the perceptual quality of the enhanced signals Perceptual Evaluation Of

Speech Quality (PESQ), [54], is used. The DRR [dB] is defined as follows:

$$\text{DRR} = \frac{10}{R}\sum_{\lambda=0}^{R-1}\log_{10}\left\{\frac{E_d\left(h_\lambda\right)}{\left(\sum_n h_\lambda^2[n]\right)-E_d\left(h_\lambda\right)}\right\}, \tag{4.7}$$

where E_d is the direct path energy and h_λ is the inverse STFT of the impulse response constructed at a shift of λ. The direct path in the impulse response may occur in between samples, therefore the path energy will be spread across the nearby samples with a sinc function. Thus the direct path energy is computed using a convolution with a sinc function with a varying offset until a maximum is found:

$$E_d\left(h_\lambda\right) = \max_\sigma \sum_{n=-\eta}^{\eta}\left(\frac{\sin\left(\pi\left(n+\sigma\right)\right)}{\pi\left(n+\sigma\right)}h_\lambda[n+n_d]\right)^2,$$

where n_d is the nearest index of the direct path in the impulse response, $\eta = 8$ is the number of sidelobes of the sinc function to use in the summation and $\sigma \in [-1:1]$ is the offset that finds the maximum power.

The SRR [dB] is defined on a frame by frame basis and then averaged across the whole signal:

$$\text{SRR}_{\text{seg}} = \frac{10}{M}\sum_{k=0}^{M-1}\log_{10}\left\{\frac{\sum_{n=kR}^{kR+QR-1} s_d[n]^2}{\sum_{n=kR}^{kR+QR-1}\left(s_d[n]-\hat{s}[n]\right)^2}\right\}, \tag{4.8}$$

where M is the total number of frames, $s_d[n]$ represents the original direct path signal and $\hat{s}[n]$ is the enhanced signal. It gives a measure of the reverberation energy in relation to the useful direct path energy. It is a similar measure to the DRR but is calculated from the clean and reverberant speech signals rather than directly from the impulse response.

The optimal coefficients from Section 4.3 were calculated for a Room Impulse Response (RIR) and the corresponding channel response from (4.5) was found. A

total of 600 RIRs were used to test the system. These correspond to a single source and microphone in 40 different rooms and 15 different position combinations in each. The impulse responses were generated using the Room Impulse Response Generator from [1], which is based on the image method [55]. In all cases we considered four times overlapping frame, $Q = 4$, a frame increment of 4 ms at 16kHz sampling frequency, $R = 64$, and 9 frames either side of the currently processed frame: $A = 9$, $B = 9$.

As both the SRR and PESQ are calculated from clean and reverberant speech signals rather than from the impulse response, speech samples were taken from the TIMIT core test set [56]. Each speech sample was convolved with the impulse response, $h[n]$, before undergoing enhancement as described in (4.2). The clean and processed signals, $s[n]$ and $\hat{s}[n]$, were then used with the SRR and PESQ metrics to gauge any improvement.

The performance of the proposed algorithm has been compared to the time domain inverse filter as proposed by Widrow, [43]. The method designs a time-domain inverse filter, $g[n]$, through least squares to best invert the system response, $h[n]$, [44]:

$$\begin{bmatrix} \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} h\left[0\right] & & & 0 \\ \vdots & h\left[0\right] & & \\ h\left[N_h-1\right] & \vdots & \ddots & \\ & h\left[N_h-1\right] & & h\left[0\right] \\ & & \ddots & \vdots \\ 0 & & & h\left[N_h-1\right] \end{bmatrix} \times \begin{bmatrix} g\left[0\right] & g\left[1\right] & \ldots & g\left[M-1\right] \end{bmatrix}^T,$$

where $N_h = 1024$ in our case and the position of the 1 on the left hand side reflects the position of the direct path in the impulse response. In the presence of no impulse response errors it represents an ideal solution to channel equalisation in the time domain. We aim to match the performance whilst operating in the STFT domain.

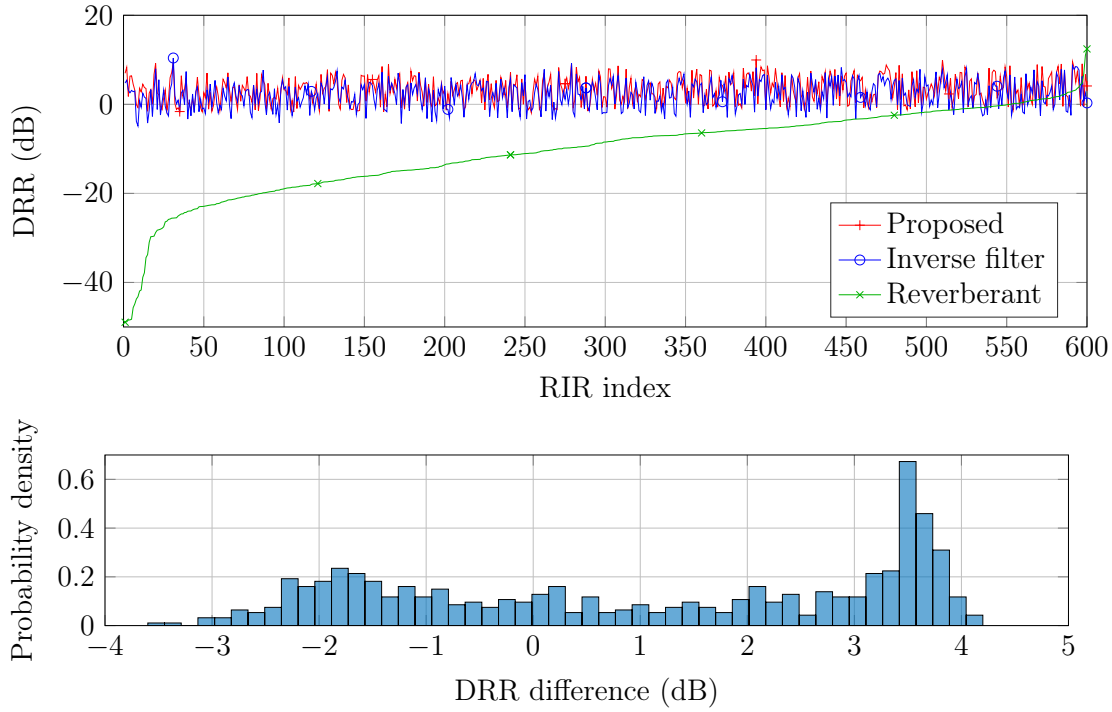

Figure 4.3: The DRR after the algorithm for 600 RIRs. A histogram of the differences in DRR between the proposed method and the inverse filter is shown in the bottom plot.

4.4.1 Results

The DRR was computed for both $h[n]$ and $\hat{h}[n]$ across all 600 RIRs. The results comparing the DRR before and after the algorithm are shown in Fig. 4.3 in which the tested rooms have been ordered by the reverberant DRR. The DRR improved for all the impulse responses tested except those where the original DRR exceeds 0 dB. It produced a mean average improvement of 1.0 dB over the inverse filter method. The resulting performance is largely independent of the amount of reverberation in the initial signal and lies close to 6 dB, giving an improvement of up to 34 dB. Thus the algorithm, given perfect knowledge of the impulse response, is able to reduce reverberation to the same level regardless of how reverberant the original channel is.

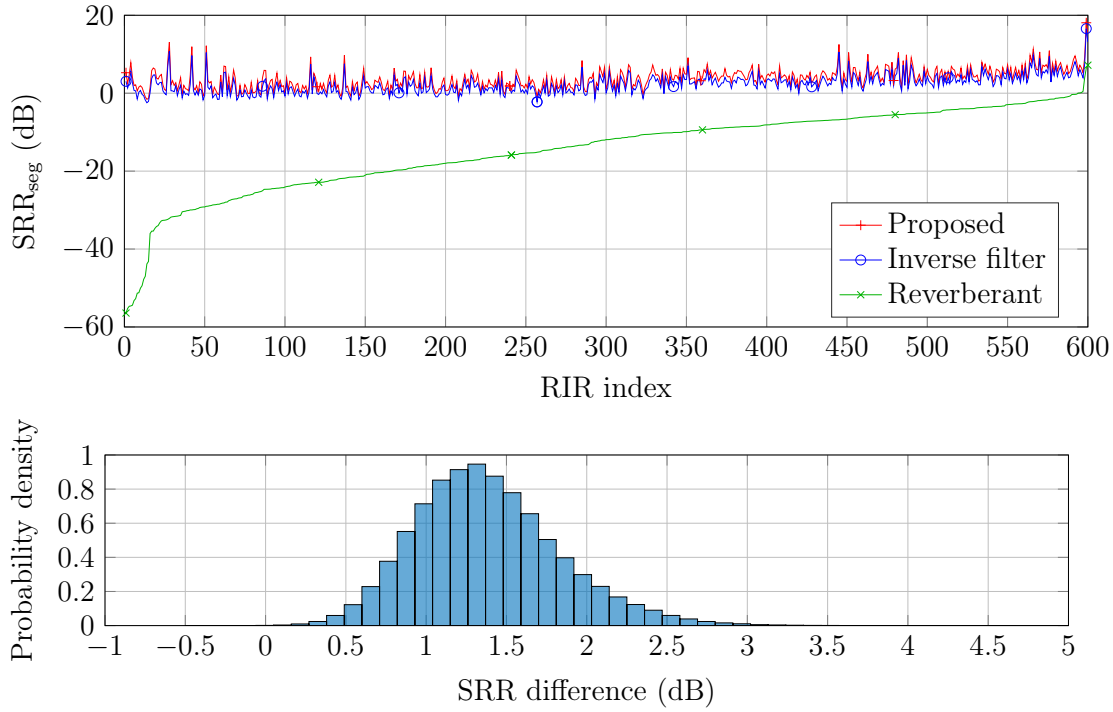

Figure 4.4: The SRR for each RIR after enhancement (top). A histogram of the differences in SRR between the proposed method and the inverse filter (bottom).

The averaged SRR for each RIR and all 240 utterances is shown in Fig. 4.4. It follows a similar pattern to the DRR. The speech signals after enhancement show a much improved SRR compared to the reverberant signals, and a mean average improvement of 1.4 dB over the inverse filter method. The enhanced signals lie around 0 dB. When the original SRR surpassed 0 dB, the algorithm was unable to make any further improvements, and caused slight degradation to these non-reverberant signals.

The averaged PESQ results are shown in Fig. 4.5. The enhancement provides moderate gains in perceptual quality which indicates that the algorithm does not introduce significant distortion or artifacts, there is a small gain of 0.08 PESQ over the inverse filter method. It should be noted that PESQ is designed to assess the perceptual impact of non-linear effects such as reverberation.

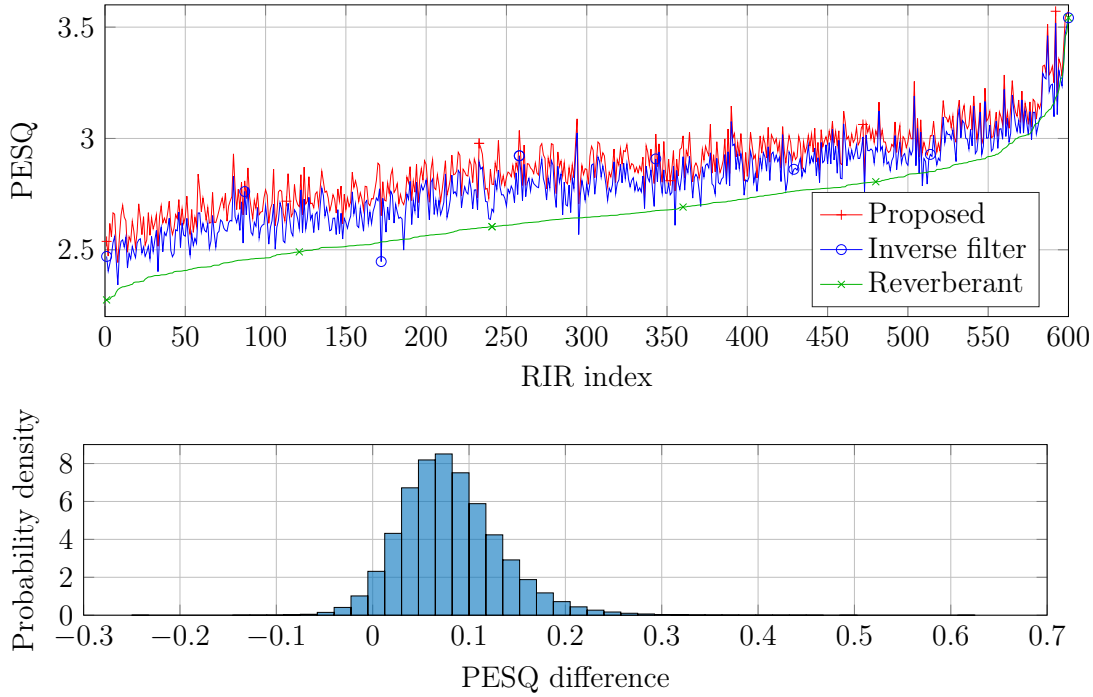

Figure 4.5: PESQ is shown for 600 different RIRs before and after enhancement (top). A histogram of the differences in DRR between the proposed method and the inverse filter (bottom).

4.5 Pre-echo reduction

The dereverberation algorithm was applied to an impulse response from a typical room, the resulting channel response is shown in Fig. 4.6. It is clear from the figure that the amount of reverberation is greatly reduced. However the enhanced impulse response, $\hat{h}[n]$, has introduced a small amount of pre-echo. Pre-echo is reverberation that occurs before the direct path, in some cases it can cause more irritation to a listener than the post-echo, [57]. Using weighted least squares we can change the optimisation problem to place more emphasis on the impulse response before the direct path, which should ensure the pre-echo effect is limited. The weighted coefficients are given as:

$$\hat{\mathbf{g}}_k = \left(\mathbf{H}_k^H \mathbf{W} \mathbf{H}_k\right)^{-1} \mathbf{H}_k^H \mathbf{W} \tilde{\mathbf{H}}_k, \tag{4.9}$$

where $\mathbf{W} \in \mathbb{R}^{R(2N-1)\times R(2N-1)}$ is a diagonal matrix of the weightings. Each diagonal element of $\mathbf{W}$ relates to the corresponding frame of $\mathbf{D}_k$. For example, we initialise $\mathbf{W}$ so that the first N values are 1, followed by $N-1$ coefficients exponentially decaying to 0. This method will emphasise the removal of energy in all impulse response taps up to the direct path, taps after the direct path will gain the energy that was removed from earlier taps. The weightings matrix, $\mathbf{W}$, could also be designed to factor in that early reflections are beneficial to intelligibility by incorporating low weights after the direct path.

The weighted least squares was applied to the impulse response from Fig. 4.6, and the resulting suppression is shown in Fig. 4.7. There is more energy in the post-echo than before however the pre-echo has decreased. The procedure has the effect of shifting the errors to the unweighted frames, in this case the post-echo; whilst the pre-echo power has decreased. This still gives a clear reduction in reverberation with respect to the original impulse response, without the introduction of significant pre-echo.

As the effect of pre-echo and post-echo have different perceived effects on a human listener, [57], energy based metrics such as DRR and SRR are not useful in evaluating

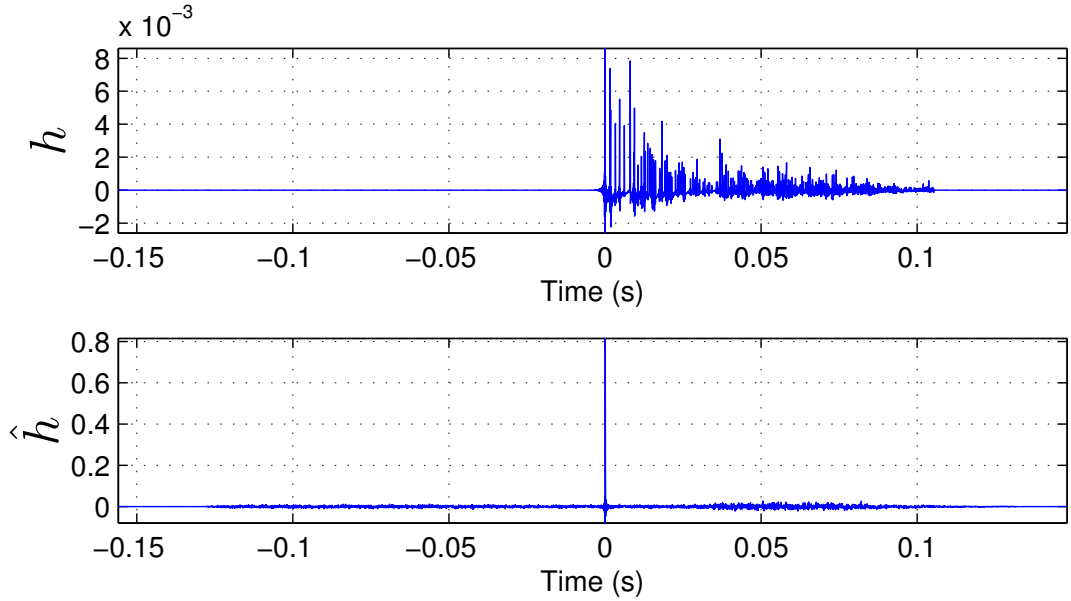

Figure 4.6: Top: Example of the effective channel response from an impulse response of a room measuring $4\,\mathrm{m} \times 6\,\mathrm{m} \times 8\,\mathrm{m}$. Bottom: The resulting impulse is close to the desired impulse, $\delta[n]$, with a small amount of distortion both before and after the peak.

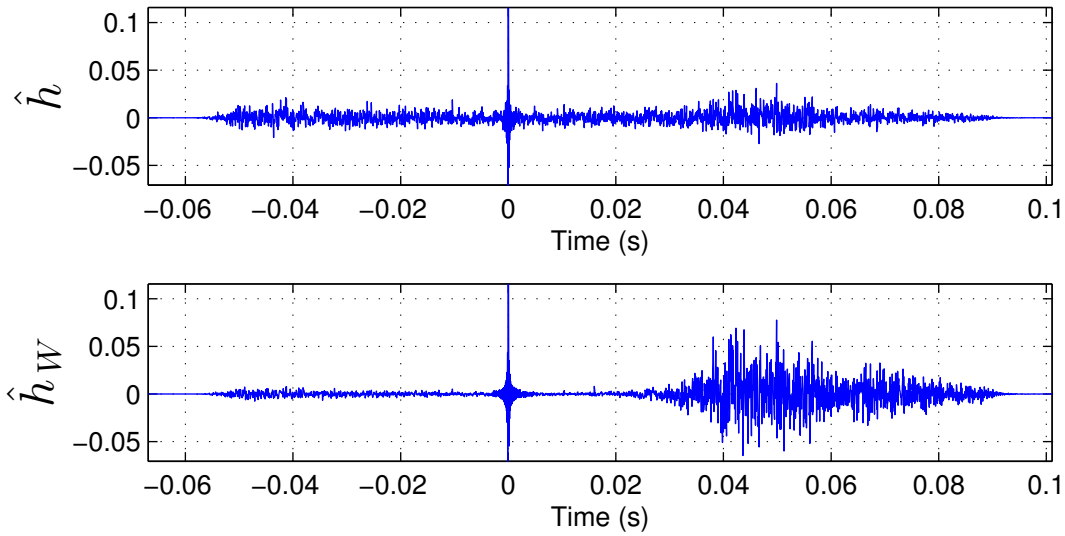

Figure 4.7: Top: Example of the standard least squares result. Bottom: Example of the pre-echo reduction technique using weighted least squares, pre-echo has been reduced at the cost of post-echo. The y-axis is zoomed for more detail.

the reduction in pre-echo. The output signals were briefly evaluated by informal listening tests, which confirmed the reduction in perceived reverberation. Furthermore, the weights could be chosen to optimise a perceptual metric such as PESQ.

4.6 Conclusions

In this chapter we have described a novel approach to dereverberation using a linear filter in the STFT domain. Using knowledge of the channel impulse response we can find an optimal combination of frames to reduce the effects of reverberation. It can overcome the time-variance of the STFT by considering all the possible impulse positions within a single frame. The algorithm gives clear performance gains in dereverberation on average. Both the DRR and the SRR show that, regardless of the amount of initial reverberation present, the enhanced signal has a similar low level of reverberation present, whilst not introducing distortion.

We have shown that the proposed STFT domain algorithm performs as well as, or slightly better than, the time domain inverse filter; allowing us to apply dereverberation in the STFT domain with the benefits discussed in Sec. 2.1.1.

Chapter 5

Acoustic Propagation Modeling

In this thesis, the signals from multiple microphones are combined in order to improve the SNR of the desired acoustic source. In this chapter we discuss the factors that affect the acoustic signals received by the microphones and develop models that account for their variability.

5.1 Sound propagation

In this section we will describe the current models of the way in which sound propagates through a medium.

In the following sections we consider a 3D space, where the cartesian position of point i in space is denoted by the vector $\bar{\mathbf{p}}_i \in \mathbb{R}^{3\times 1}$ with units *metres*. The time taken for sound to travel through a medium in a straight line, between a source at $\bar{\mathbf{p}}_a$ and a microphone at $\bar{\mathbf{m}}_\epsilon$, the propagation path time, is given by:

$$\bar{t}_{\epsilon a} \triangleq \frac{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}{c}, \tag{5.1}$$

where $\|\cdot\|$ is the Euclidean distance and c is the propagation speed of the medium. The propagation speed in air is proportional to the square root of absolute temperature and is approximately 343 m/s at room temperature.

Assuming lossless propagation, the acoustic energy of a sound source is inversely proportional to the square of the distance from source to microphone. For a source at $\bar{\mathbf{p}}_a$ and a microphone at $\bar{\mathbf{m}}_\epsilon$, the sound pressure, and hence the amplitude of the microphone signal, is proportional to:

$$\bar{h}_{\epsilon a} \triangleq \frac{1}{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}. \tag{5.2}$$

In the STFT domain, the propagation time results in a phase shift that is different in each frequency bin. We therefore define the complex propagation coefficient in the STFT domain as the combination of the channel amplitude and phase contributions:

$$\begin{aligned} \bar{d}_{\epsilon a}(k) &\triangleq \bar{h}_{\epsilon a} \exp\left(-j\omega_k \bar{t}_{\epsilon a}\right), & (5.3)\\ &= \frac{1}{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|} \exp\left(-j\omega_k \frac{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}{c}\right), & (5.4) \end{aligned}$$

where ω_k in frequency bin k corresponds to the frequency of the bin.

For a single source received by M microphones, we can form an $M \times 1$ column vector, known as the "steering vector", by concatenating the propagation coefficients:

$$\bar{\mathbf{d}}(k) = \begin{bmatrix} \bar{d}_{1,1}(k) \\ \vdots \\ \bar{d}_{M,1}(k) \end{bmatrix}. \tag{5.5}$$

Where there are P sources we concatenate the steering vectors into a single $M \times P$ matrix:

$$\bar{\mathbf{D}}(k) = \begin{bmatrix} \bar{d}_{1,1}(k) & \dots & \bar{d}_{1,P}(k) \\ \vdots & & \\ \bar{d}_{M,1}(k) & & \bar{d}_{M,P}(k) \end{bmatrix}.$$

As each frequency bin is processed independently, we omit the frequency index for clarity.

5.2 Proposed uncertainties model

In the field of microphone array processing the propagation channels between each source and each microphone, $\bar{\mathbf{D}}$, are normally assumed to be a deterministic function of the source and microphone geometry. However, in real acoustic environments, there are several mechanisms that can give rise to unpredictable correlated variations in the phases and amplitudes of the $\bar{d}_{\epsilon a}$.

For example, position calibration errors can occur, where either the sources or microphones are not precisely located [58, 59]. Algorithms can be designed with the assumption that incident sources are modeled as plane waves, but are then presented with near field sources degrading the performance [60, 61]. Source spreading and surface reflections from multipath scenarios such as reverberant rooms can distort the estimation of steering vectors [62, 63, 64, 65]. Temperature and other factors can affect the speed of propagation through a medium, which causes an uncertainty in the phase at the microphones, [66, 67]. Delay variations occur from diffraction around objects in the propagation path.

The uncertainties introduced by the above phenomena all have a similar effect of changing the propagation channel from each source to each microphone. Thus the steering vectors used to design the spatial algorithms, such as beamformers, do not match the real environment. These differences will degrade the performance of a non-robust beamformer [68]; this is the case even with small perturbations [69, 70], and particularly in setups where the performance can rapidly fall when the steering vector

is incorrectly assumed. By modeling these deviations, we can construct beamformers that are robust to these random phase changes.

For example, in the case of position calibration errors, the errors cause a change in propagation distance and phase differences, in a similar way to steering vector mismatches. By identifying which element positions are less well defined than others we are able to utilise the most reliable microphones to avoid large phase uncertainties.

Channel uncertainty errors, from changes in propagation speed cause correlated errors terms across different microphones. Modeling channel uncertainties is especially important when using distributed beamformers with widely separated microphones. It is a common assumption that the channel propagation speed between a source and two different microphones is identical. However when the arrays are far apart this no longer applies [71]. Because the channels may be widely separated in space, the variations can become large enough to cause phase differences that degrade the performance of the beamformer.

These considerations apply to any situation where there is uncertainty in wave propagation speed. For example, in the field of medical imaging, ultrasound scans can be adversely affected by speed-of-sound errors [72] and in sonar systems the propagation speed may vary greatly [73, 74].

Uncertainties in microphone positions or channel propagation speeds result in phase uncertainties whose magnitude is proportional to frequency. Thus the higher the frequency, the larger the deviations and the less reliable is the corresponding microphone. In this chapter we model these correlated and uncorrelated variations. Using these uncertainties we can design a more robust beamformer that utilises the most reliable microphones at each frequency.

In the following section we detail our framework for modeling the uncertainties in the propagation coefficients and validate the models with experimental data. The effects of uncertainties in the propagation channel can be divided into two broad sections which are discussed separately: time uncertainties and amplitude uncertainties.

5.3 Time uncertainties

The following section will explain the notion of uncertainties that are caused by changes to the propagation path time. When either the propagation distance or the propagation speed vary, the propagation path time changes.

We can write the total propagation time from source a to microphone ϵ as $\bar{t}_{\epsilon a} + \tilde{t}_{\epsilon a}$ where $\bar{t}_{\epsilon a}$ is the mean propagation delay and $\tilde{t}_{\epsilon a}$ is a zero-mean random variable which we assume to follow a Gaussian distribution. We model the contributions as a random time difference which translates into a phase shift:

$$d_{\epsilon a}(k) \;=\; \bar{d}_{\epsilon a}(k) \exp\left(j\omega_k \tilde{t}_{\epsilon a}\right), \tag{5.6}$$

where $\overline{d}$ represents the conventional deterministic propagation model shown in (5.4), $\tilde{t}_{\epsilon a}$ represents the variations in the propagation path time between a source at $\bar{\mathbf{p}}_a$ and a microphone at $\bar{\mathbf{m}}_\epsilon$. In matrix form this can be expressed as:

$$\mathbf{D} \;=\; \overline{\mathbf{D}} \odot \exp\left(j\omega_k \tilde{\mathbf{T}}\right), \tag{5.7}$$

where $\odot$ denotes element-by-element multiplication.

As will be seen later in Chapter 8, when designing a robust beamformer we require the covariance matrix of the propagation coefficients, $\left\langle \mathbf{D}\mathbf{D}^H \right\rangle$, where $\langle\cdot\rangle$ denotes expectation over all $\tilde{\mathbf{T}}$ and $(\cdot)^H$ is the Hermitian transpose. Below we expand $\left\langle \mathbf{D}\mathbf{D}^H \right\rangle$, for the case of a single source, $\mathbf{D} \in \mathbb{C}^{M\times 1}$, to show how it depends on the time uncertainties covariance matrix, $\left\langle \tilde{\mathbf{T}}\tilde{\mathbf{T}}^H \right\rangle$:

$$\left\langle \mathbf{D}\mathbf{D}^H \right\rangle \;=\; \left\langle \left(\overline{\mathbf{D}} \odot \exp(j\omega_k \tilde{\mathbf{T}})\right) \left(\overline{\mathbf{D}} \odot \exp(j\omega_k \tilde{\mathbf{T}})\right)^H \right\rangle, \tag{5.8}$$

where a typical element of the $M \times M$ matrix is given by

$$\left\langle \mathbf{D}\mathbf{D}^H \right\rangle_{\epsilon,\varphi} = \sum_{a=1}^{P} \bar{d}_{\epsilon a} \bar{d}^*_{\varphi a} \left\langle \exp\left(j\omega_k \left(\tilde{t}_{\epsilon a} - \tilde{t}_{\varphi a}\right)\right)\right\rangle, \tag{5.9}$$

in which $\bar{d}_{\epsilon a}$ and $\tilde{t}_{\epsilon a}$ are elements of $\overline{\mathbf{D}}$ and $\tilde{\mathbf{T}}$ respectively. The expectation of the exponential term can be formulated as a log-normal distribution, $X = \exp\left(\mu + \sigma\dot{Z}\right)$, where $\dot{Z}$ is a normal Gaussian distribution, the expectation of which is given as [75]:

$$\langle X \rangle = e^{\mu + \sigma^2/2}.$$

In which case $\left\langle \mathbf{D}\mathbf{D}^H \right\rangle_{\epsilon,\varphi}$ is evaluated as:

$$\begin{aligned}
\left\langle \mathbf{D}\mathbf{D}^H \right\rangle_{\epsilon,\varphi} &= \sum_{a=1}^{P} \bar{d}_{\epsilon a} \bar{d}^*_{\varphi a} \exp\left(-\frac{\omega_k^2}{2}\left\langle \left(\tilde{t}_{\epsilon a} - \tilde{t}_{\varphi a}\right)^2 \right\rangle\right) \\
&= \sum_{a=1}^{P} \bar{d}_{\epsilon a} \bar{d}^*_{\varphi a} \exp\left(-\frac{\omega_k^2}{2}\left(\left\langle \tilde{t}^2_{\epsilon a}\right\rangle + \left\langle \tilde{t}^2_{\varphi a}\right\rangle - 2\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle\right)\right).
\end{aligned} \tag{5.10}$$

From the above we see that in order to compute $\left\langle \mathbf{D}\mathbf{D}^H \right\rangle$, we need to determine the covariance between the elements of $\tilde{\mathbf{T}}$, $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle$, that represent the propagation time uncertainties from source a to microphones ϵ and φ respectively.

Each contribution to the time uncertainties model will be formulated into the form $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle$. We assume each uncertainty effect to be independent so that the covariance $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle$ may be expressed as a sum:

$$\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle = \left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle_S + \left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle_C, \tag{5.11}$$

where $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle_S$ and $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a}\right\rangle_C$ represent the uncertainty from position calibration errors and channel uncertainties respectively. Therefore each effect can considered separately.

The following sections detail the two contributions we have formulated, where array elements are not precisely located and where the channel speed is uncertain.

5.3.1 Position calibration errors

Each microphone or source in the array geometries may not be precisely located in the expected position. The change to the expected channel propagation path results in a change in propagation time and in turn causes a phase shift in the received signal. For example, we consider a single human talker and a microphone setup. In normal speech it is not unreasonable for the position of the human's mouth to oscillate around its mean by up to 10cm [76]. The resulting phase difference at 1 kHz is approximately 105°, which is enough to cause large performance degradations in array processing algorithms.

If we have a source at the position $\bar{\mathbf{p}}_a + \tilde{\mathbf{p}}_a$ and a microphone at the position $\bar{\mathbf{m}}_\epsilon + \tilde{\mathbf{m}}_\epsilon$ where $\tilde{\mathbf{p}}_a$ and $\tilde{\mathbf{m}}_\epsilon$ are zero-mean normally-distributed deviations from the nominal positions, then the change in path length due to $\tilde{\mathbf{p}}_a$ and $\tilde{\mathbf{m}}_\epsilon$ is the component of $(\tilde{\mathbf{m}}_\epsilon - \tilde{\mathbf{p}}_a)$ in the direction of $(\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a)$:

$$\delta(\epsilon, a) = \frac{(\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a)^T (\tilde{\mathbf{m}}_\epsilon - \tilde{\mathbf{p}}_a)}{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}. \tag{5.12}$$

The path length correlation along two different paths, from two different sources, is therefore:

$$\begin{aligned}\langle\delta(\epsilon,a)\delta(\varphi,b)\rangle &= \left\langle \frac{\left(\bar{\mathbf{m}}_\epsilon-\bar{\mathbf{p}}_a\right)^T\left(\tilde{\mathbf{m}}_\epsilon-\tilde{\mathbf{p}}_a\right)}{\|\bar{\mathbf{m}}_\epsilon-\bar{\mathbf{p}}_a\|}\frac{\left(\bar{\mathbf{m}}_\varphi-\bar{\mathbf{p}}_b\right)^T\left(\tilde{\mathbf{m}}_\varphi-\tilde{\mathbf{p}}_b\right)}{\|\bar{\mathbf{m}}_\varphi-\bar{\mathbf{p}}_b\|}\right\rangle \quad (5.13)\\ &= \frac{\left(\bar{\mathbf{m}}_\epsilon-\bar{\mathbf{p}}_a\right)^T\left\langle\left(\tilde{\mathbf{m}}_\epsilon-\tilde{\mathbf{p}}_a\right)\left(\tilde{\mathbf{m}}_\varphi-\tilde{\mathbf{p}}_b\right)^T\right\rangle\left(\bar{\mathbf{m}}_\varphi-\bar{\mathbf{p}}_b\right)}{\|\bar{\mathbf{m}}_\epsilon-\bar{\mathbf{p}}_a\|\,\|\bar{\mathbf{m}}_\varphi-\bar{\mathbf{p}}_b\|}.\end{aligned}$$

The numerator can be expanded using:

$$\left\langle\left(\tilde{\mathbf{m}}_\epsilon-\tilde{\mathbf{p}}_a\right)\left(\tilde{\mathbf{m}}_\varphi-\tilde{\mathbf{p}}_b\right)^T\right\rangle = \left\langle\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{m}}_\varphi^T-\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{p}}_b^T-\tilde{\mathbf{p}}_a\tilde{\mathbf{m}}_\varphi^T+\tilde{\mathbf{p}}_a\tilde{\mathbf{p}}_b^T\right\rangle. \quad (5.14)$$

The position variations between sources and microphones are assumed to be independent, $\left\langle\bar{\mathbf{m}}_\epsilon\bar{\mathbf{p}}_b^T\right\rangle=0$. Thus we can simplify the above:

$$\left\langle\left(\tilde{\mathbf{m}}_\epsilon-\tilde{\mathbf{p}}_a\right)\left(\tilde{\mathbf{m}}_\varphi-\tilde{\mathbf{p}}_b\right)^T\right\rangle = \left\langle\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{m}}_\varphi^T\right\rangle+\left\langle\tilde{\mathbf{p}}_a\tilde{\mathbf{p}}_b^T\right\rangle. \quad (5.15)$$

5.13 becomes:

$$\langle\delta(\epsilon,a)\delta(\varphi,b)\rangle = \frac{\left(\bar{\mathbf{m}}_\epsilon-\bar{\mathbf{p}}_a\right)^T\left(\left\langle\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{m}}_\varphi^T\right\rangle+\left\langle\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{m}}_\varphi^T\right\rangle\right)\left(\bar{\mathbf{m}}_\varphi-\bar{\mathbf{p}}_b\right)}{\|\bar{\mathbf{m}}_\epsilon-\bar{\mathbf{p}}_a\|\,\|\bar{\mathbf{m}}_\varphi-\bar{\mathbf{p}}_b\|}. \quad (5.16)$$

If the position deviations are isotropic, we can write $\left\langle\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{m}}_\varphi^T\right\rangle=\sigma_m^2\mathbf{I}$ for $\epsilon=\varphi$ and $\left\langle\tilde{\mathbf{m}}_\epsilon\tilde{\mathbf{m}}_\varphi^T\right\rangle=\sigma_p^2\mathbf{I}$ for $a=b$, which results in $\langle\delta(\epsilon,a)\delta(\epsilon,a)\rangle=\sigma_m^2+\sigma_p^2$.

We note that the change in path length causes a change in the propagation path time which is given by:

$$\left\langle\tilde{t}_{\epsilon,a}\tilde{t}_{\varphi,b}\right\rangle_S=\frac{1}{c^2}\langle\delta(\epsilon,a)\delta(\varphi,b)\rangle. \quad (5.17)$$

5.3.2 Channel speed uncertainty

When we consider the propagation path from a source to a microphone, in the traditional case, we assume that the propagation speed is constant, giving a predictable propagation delay. The path delay from a source at $\bar{\mathbf{p}}_a$ and a microphone at $\bar{\mathbf{m}}_\epsilon$

$$\bar{t}_{\epsilon a} = \frac{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}{c_0}, \tag{5.18}$$

where c_0 is the average propagation speed. For larger propagation distances the propagation speed can vary due to a number of variables such as temperature and air flows, [73, 74]. For example, a hot pocket of air will change the propagation speed in a localised space. Multiple acoustic channels which pass through this pocket will be similarly affected in terms of propagation time, whereas a channel not passing through will not have a similarly correlated time deviation. Generally, channels which are close together will be more correlated in terms of deviations from their expected propagation time, and widely separated channels will not be well correlated.

If there is a heat source present in a room, which varies the temperature by 10°, between 15° C and 25° C, the effective speed of propagation varies between 340 and 346 m/s. Over a 10 m channel the variance in speed of propagation corresponds to a 500 μs propagation time difference, corresponding to a 180° phase difference at 1 kHz, which could cause severe performance degradations.

In order to model this process, we can describe the propagation speed as a random quantity and as a function of position. The inverse propagation speed at any point can be modeled as follows:

$$\frac{1}{c(\mathbf{x})} = \frac{1}{c_0} + g(\mathbf{x}), \tag{5.19}$$

where the quantity $\frac{1}{c_0}$ is the mean inverse speed and the deviation from this value, $g(\mathbf{x})$, is zero mean.

The total path delay, $t_{\epsilon a}$, from source $\bar{\mathbf{p}}_a$ to microphone $\bar{\mathbf{m}}_\epsilon$ is given by the line integral along the propagation path:

$$\begin{aligned} t_{\epsilon a} &= \bar{t}_{\epsilon a} + \tilde{t}_{\epsilon a} \\ &= \frac{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}{c_0} + \|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\| \int_{s=0}^{1} g\left(\bar{\mathbf{p}}_a + (\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a)\, s\right) ds. \end{aligned} \tag{5.20}$$

We want to model the distribution of inverse speed in a simple way that captures its spatial correlation. We therefore assume that

$$g(\mathbf{x}) = \alpha u(\mathbf{x}) * \mathbf{F}(\mathbf{x}), \tag{5.21}$$

where α is an overall amplitude factor, u is uniform uncorrelated unit-variance Gaussian white noise, $*$ denotes 3-dimensional convolution and $\mathbf{F}$ is a 3-dimensional spatial low-pass filter. We now assume that $\mathbf{F}$ is an isotropic Gaussian distribution:

$$\mathbf{F}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x};\, \mathbf{0},\, \sigma^2\mathbf{I}\right) = \left(2\pi\sigma^2\right)^{-\frac{3}{2}} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right), \tag{5.22}$$

where σ^2 is the variance of the distribution. The covariance is given as:

$$\begin{aligned}
\langle g(\mathbf{x})g(\mathbf{y})\rangle &= \alpha^2 \left\langle \left(u(\mathbf{x}) * \mathbf{F}(\mathbf{x})\right)\left(u(\mathbf{y}) * \mathbf{F}(\mathbf{y})\right)\right\rangle \\
&= \left\langle \iiint \alpha u(\mathbf{x}')\mathcal{N}(\mathbf{x}-\mathbf{x}';\,\mathbf{0},\,\sigma^2\mathbf{I})d\mathbf{x}' \iiint \alpha u(\mathbf{y}')\mathcal{N}(\mathbf{y}-\mathbf{y}';\,\mathbf{0},\,\sigma^2\mathbf{I})d\mathbf{y}' \right\rangle \\
&= \alpha^2 \iiint \iiint \langle u(\mathbf{x}')u(\mathbf{y}')\rangle \,\mathcal{N}(\mathbf{x}-\mathbf{x}';\,\mathbf{0},\,\sigma^2\mathbf{I})\mathcal{N}(\mathbf{y}-\mathbf{y}';\,\mathbf{0},\,\sigma^2\mathbf{I})d\mathbf{x}'d\mathbf{y}' \\
&= \alpha^2 \iiint \mathcal{N}(\mathbf{x}-\mathbf{x}';\,\mathbf{0},\,\sigma^2\mathbf{I})\mathcal{N}(\mathbf{y}-\mathbf{x}';\,\mathbf{0},\,\sigma^2\mathbf{I})d\mathbf{x}' \\
&= \alpha^2\mathcal{N}(\mathbf{x}-\mathbf{y};\,\mathbf{0},\,2\sigma^2\mathbf{I}) \iiint \mathcal{N}(\mathbf{x}';\,\frac{1}{2}\left(\mathbf{y}+\mathbf{x}\right),\,\frac{1}{2}\sigma^2\mathbf{I})d\mathbf{x}' \\
&= \frac{\alpha^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\mathbf{x}-\mathbf{y}\right)^T\left(\mathbf{x}-\mathbf{y}\right)}{2\sigma^2}\right) \\
&= \kappa^2 \exp\left(-\frac{\left(\mathbf{x}-\mathbf{y}\right)^T\left(\mathbf{x}-\mathbf{y}\right)}{2\sigma^2}\right), \qquad (5.23)
\end{aligned}$$

which is characterised by the two parameters $\kappa^2 = \langle g(\mathbf{x})g(\mathbf{x})\rangle = \frac{\alpha^2}{\sqrt{2\pi\sigma^2}}$ and σ^2 which define its variance and spatial extent respectively.

The covariance between the two paths is given by:

$$\left\langle \tilde{t}_{\epsilon,a}\tilde{t}_{\varphi,b}\right\rangle_C = \left\langle \left(t_{\epsilon a} - \bar{t}_{\epsilon a}\right)\left(t_{\varphi b} - \bar{t}_{\varphi b}\right)\right\rangle . \qquad (5.24)$$

This can be simplified into the following form:

$$
\begin{aligned}
\left\langle \tilde{t}_{\epsilon,a}\tilde{t}_{\varphi,b}\right\rangle_C &= \left\langle \|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\| \int_{t=0}^{1} g\left(\bar{\mathbf{p}}_a + (\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a)\, t\right) dt \right. \\
&\qquad \left. \|\bar{\mathbf{m}}_\varphi - \bar{\mathbf{p}}_b\| \int_{s=0}^{1} g\left(\bar{\mathbf{p}}_b + (\bar{\mathbf{m}}_\varphi - \bar{\mathbf{p}}_b)\, s\right) ds \right\rangle \\
&= \|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|\, \|\bar{\mathbf{m}}_\varphi - \bar{\mathbf{p}}_b\| \\
&\qquad \iint_0^1 \left\langle g\left(\bar{\mathbf{p}}_a + (\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a)\, t\right) g\left(\bar{\mathbf{p}}_b + (\bar{\mathbf{m}}_\varphi - \bar{\mathbf{p}}_b)\, s\right)\right\rangle dtds \\
&= \|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|\, \|\bar{\mathbf{m}}_\varphi - \bar{\mathbf{p}}_b\|\, \kappa^2 \\
&\qquad \iint_0^1 \exp\left(-\frac{\left|\bar{\mathbf{p}}_a + (\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a)\, t - \bar{\mathbf{p}}_b + (\bar{\mathbf{m}}_\varphi - \bar{\mathbf{p}}_b)\, s\right|^2}{2\sigma^2} \right) dtds.
\end{aligned} \tag{5.25}
$$

5.3.2.1 Interpretation of $\left\langle \tilde{t}_{\epsilon,a}\tilde{t}_{\varphi,b}\right\rangle_C$

As we increase the channel distance between the source and the microphone, the channel uncertainty, $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\epsilon a}\right\rangle_C$, increases. This relationship is shown in Fig. 5.1. The initial

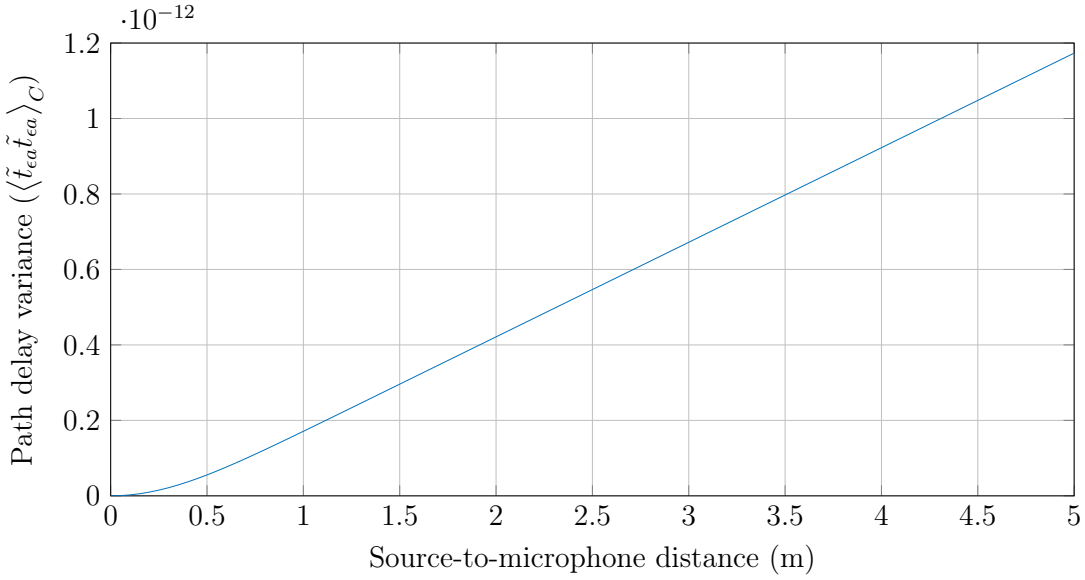

Figure 5.1: The expected channel uncertainty against the channel distance.

curvature is exponential based but the remainder tends towards a linear relationship.

The variance and spatial spread will vary depending on the medium under analysis, particularly in scenarios where the temperature can vary. For example, in sea based applications, such as SONAR, the speed of sound can vary significantly, between 1400-1600 m/s, and quickly between different spaces [73]. As such the values which govern (5.23) will need to be determined to suit the given environment.

5.4 Time uncertainty experiments

In order to obtain realistic values for the parameters used in the model in (5.23), a series of tests were conducted. The main focus of these tests was to estimate the parameters σ and κ from the inverse speed model for a typical room.

5.4.1 Test design

The parameters govern the covariance between the propagation time uncertainties across different channels. By measuring the propagation time across several different acoustic channels over an extend period we can compute the covariance in time uncertainty.

A single loudspeaker was used to produce a sound field across a room. A series of microphones recorded the sound field at various positions in the room. The propagation time across each acoustic channel was calculated over short time frames across a long time period.

5.4.1.1 Finding uncertainties matrix

To find the delay between two signals we can utilise their covariance:

$$\operatorname{cov}(x, y)_n = \sum_i x[i]\, y[i-n]\,. \tag{5.26}$$

The first large peak of the covariance function corresponds to the sample delay of the channel:

$$d_s = \arg\max_n \left[\mathrm{cov}\left(x, y\right)_n\right]. \tag{5.27}$$

To improve the resolution of the delay calculation, quadratic interpolation over three samples centred on the peak was used to find the peak position, with sub-sample resolution. The peak sample position was converted into a propagation time by dividing by the sampling frequency.

In order to improve the robustness of the time computation, the loudspeaker was driven with pseudo-random white noise, whose autocorrelation is an impulse.

Both the loudspeaker and microphone signals were divided into short frame times so that the propagation time can be computed as a function of time. The frame length is chosen so that the direct path peak is clearly distinguishable from the covariance. The short frame from the loudspeaker output was compared with a longer microphone frame which is also delayed to ensure that the entire loudspeaker frame can be found in the microphone frame.

Covariance matrix The resulting propagation times across all time frames were compared across all channels. As we are only interested in the variation from the expected propagation time, the mean delay along the channel was subtracted from all propagation times. The covariance in the time variations across two channels indicates how similarly coupled the two channels are.

5.4.1.2 Test setup

The tests were constructed using a single coil Fostex 6301B Personal Monitor and multiple DPA 4060 omni-directional microphones. All signals were routed through a RMS Fireface 800. All elements were secured on stands and set to 90cm from the ground. The microphones were orientated to point towards the loudspeaker. All elements were away from walls and other reflective surfaces, except the ceiling and

floor. The room that was used, measured 10.3x9.17x3.00m, with a T60 of 0.371s and the temperature was at 19°C. There was air conditioning present in the room, which was the only significant noise source, but was uncorrelated with the white noise of the loudspeaker, thus did not affect the results.

The tests were recorded at 44.1 kHz for approximately 10 minutes each. The frame length used to compute the propagation time in all tests was 0.5 s, giving approximately 1200 frames in total.

The propagation time measured in each frame includes contributions from all aspects of the acoustic channel, which includes positional uncertainties. In order to minimise the impact of position uncertainties the loudspeaker and microphones were attached as securely as possible. To ensure that all microphones were setup in a similar fashion, they were all placed in a similar location approximately 1m in front of the loudspeaker and the propagation time uncertainty covariance matrix was measured. In this case the channel uncertainty should be similar across all channels, so the observed differences between channels should be due to positional variations. The microphones were adjusted until a similar level of variance was seen across all channels. The microphones were then repositioned for the following tests.

5.4.2 Test 1: linear array

The first test considered four microphones positioned along a straight line directly in front of the loudspeaker. We expect as the channel length increases the resulting uncertainty increases. As the microphones are placed in a line, much of the propagation path to the furthest microphone was that of the other channels, thus there should be an increasing correlation between the uncertainties seen on co-located far away channels.

The microphone layout is shown in Fig. 5.2. A typical correlation frame is shown in Fig. 5.3. It is evident that the propagation time is well defined by the large peak. Over all frames the propagation time was found for source $P1$ to microphone $M1$, in Fig. 5.4 a histogram of the deviations from the mean propagation time are shown. The Shapiro-Wilk test for normality was run on these propagation time uncertainties

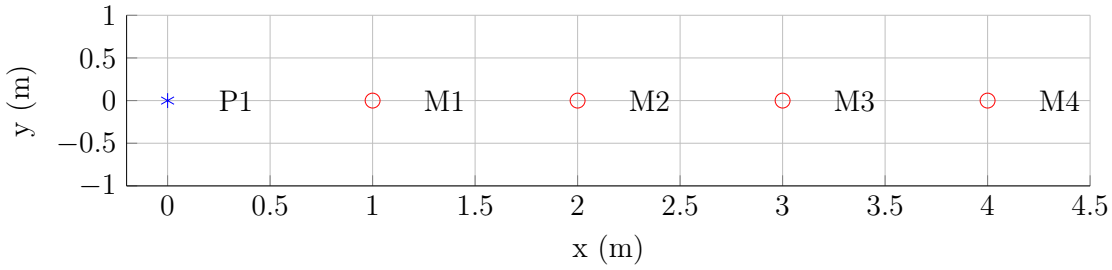

Figure 5.2: Layout of the linear array test, $P1$ indicates the loudspeaker location, M_ϵ indicates the location of microphone ϵ. Blue - sources, red - microphones.

and returned a p-value of 0.16; this justifies modeling this distribution as Gaussian. The resulting distribution shows that propagation time over a 1m channel in a typical room deviated by $\pm 1 \times 10^{-6}$ s.

The propagation time uncertainties for each frame were arranged into the vector $[M1\, M2\, M3\, M4]^T$. The resulting propagation time uncertainty covariance matrix over all frames was computed as:

$$\begin{bmatrix} 0.129 & 0.196 & 0.333 & 0.371 \\ 0.196 & 0.853 & 0.808 & 1.176 \\ 0.333 & 0.808 & 1.681 & 1.622 \\ 0.371 & 1.176 & 1.622 & 2.220 \end{bmatrix} \times 10^{-12}.$$

The deviations in propagation time across nearby channels had a higher correlation than those of further away channels. Thus there exists evidence to suggest the model fits real rooms. As expected, the uncertainty increased with distance. This is likely due to channel uncertainties rather than positional uncertainties. Position uncertainties would not increase with distance, as they contribute the same time uncertainty regardless of channel distance. Fig. 5.5 shows the variance in propagation time uncertainty against channel distance. The measurement show a near linear fit between channel distance and variance, with equation $\left\langle \tilde{t}_{\epsilon a} \tilde{t}_{\epsilon a} \right\rangle_C = (-0.6 + 0.71 \left\| \bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a \right\|) \times 10^{-12}$. As previously shown in Fig. 5.1, the

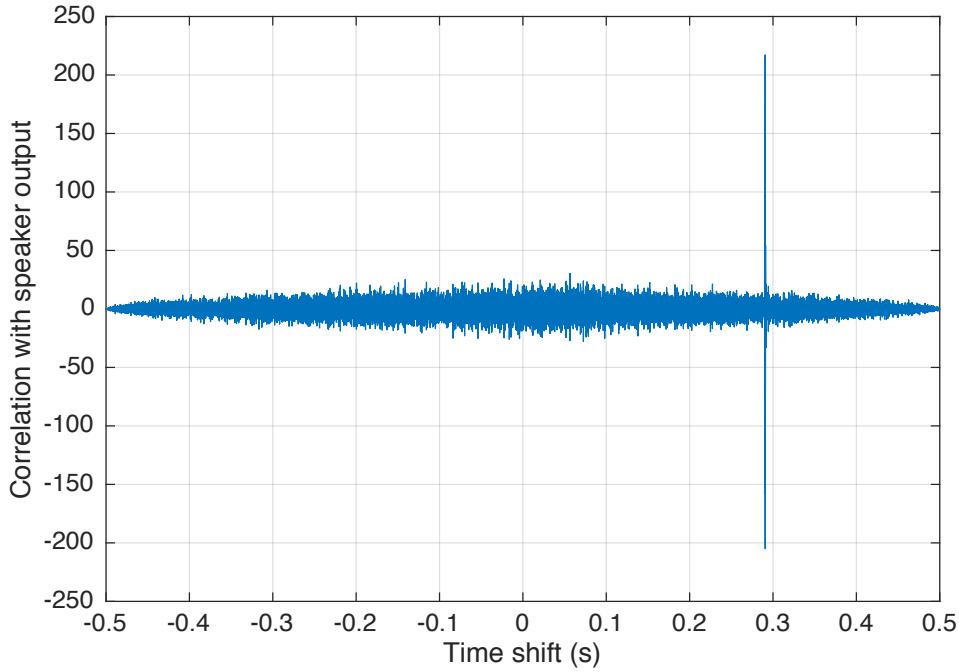

Figure 5.3: Typical correlation of a single microphone frame against the corresponding loudspeaker frame, the high time delay is due to latency.

channel uncertainties model tends towards a linear relationship when distances are large enough; from this graph we can assume that we are acting in the linear region.

5.4.3 Parameter estimation

With results of the above test we are able to compute reasonable values for both κ and σ from the channel uncertainties model 5.23. The value of κ changes the level of variance across all channels (the diagonal terms), whereas σ changes the correlation between neighbouring channels (the off-diagonal terms). Smaller σ values gives less correlation between channels.

As a closed form or least squares solution is not available and that derivatives are not easily computed, mapping the parameters to the results was achieved using an

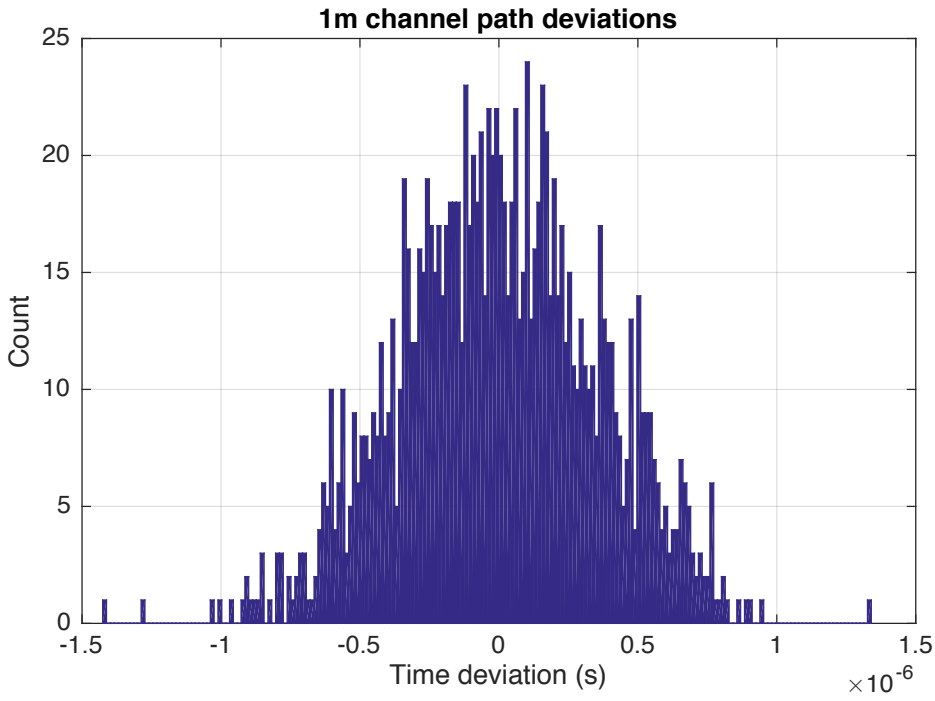

Figure 5.4: Propagation time deviations from the mean across all frames for a 1m source-microphone distance ($P1$ to $M1$).

optimiser from MATLAB [77]. Many random starting points were used and converged to the same point. The values were found to be $\kappa = 4.44 \times 10^{-7}$ and $\sigma = 1.61$. The resulting model prediction for the linear array in shown on Fig. 5.1. The resulting estimated covariance matrix is:

$$\begin{bmatrix} 0.1912 & 0.3507 & 0.4433 & 0.4807 \\ 0.3507 & 0.7013 & 0.9535 & 1.084 \\ 0.4433 & 0.9535 & 1.397 & 1.686 \\ 0.4807 & 1.084 & 1.686 & 2.167 \end{bmatrix} \times 10^{-12}.$$

This room represents small levels of uncertainty. In different room scenarios the distribution parameters will be different. For example, if there was a heat source,

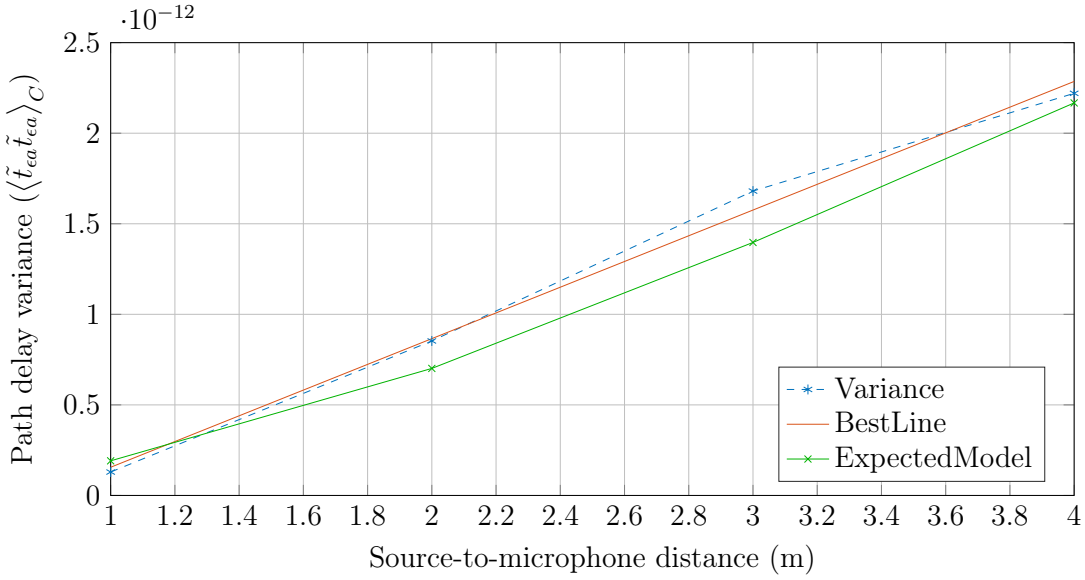

Figure 5.5: The variance in propagation delay uncertainty against channel distance

larger air currents or high levels of reverberation, the parameters would change to reflect the increase in variance.

5.4.4 Test 2: right angle setup

In order to assess the spatial diversity of the channel uncertainties, the microphones were placed along different channels. The microphone layout is shown in Fig. 5.6. The expectation is that the uncertainties along the channels to $M1$ and $M2$ are similar. Half of the channel to microphone $M2$ is the same as that to microphone $M1$, therefore the uncertainties of both channels should have some correlation; similarly to those of $M3$ and $M4$. However as the pairs of microphones are widely separated, we expect a lower correlation between the uncertainties seen on $M2$ and $M4$.

The propagation time uncertainties were arranged into the vector $[M1\ M2\ M3\ M4]^T$. The resulting propagation time uncertainty covariance matrix

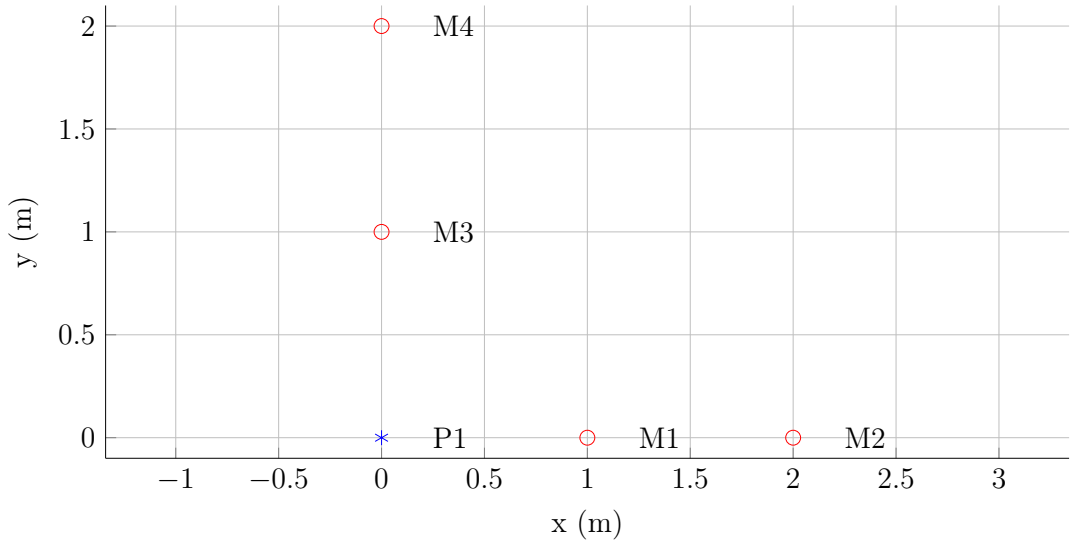

Figure 5.6: Layout of the right angle setup, $P1$ indicates the loudspeaker location, M_ϵ indicates the location of microphone ϵ.

was computed as:

$$\begin{bmatrix} 1.521 & 0.9384 & 1.644 & 1.361 \\ 0.9384 & 2.251 & 1.113 & 1.505 \\ 1.644 & 1.113 & 3.609 & 2.874 \\ 1.361 & 1.505 & 2.874 & 4.777 \end{bmatrix} \times 10^{-13}.$$

The above results do not exactly show the relationship we expect. In this case, it shows that the spatial relationship of the channel uncertainties is well correlated across the two far channels. The results are not symmetrical about the two channels. Over several trials of the same setup there was a large amount of variance in the estimated parameters between each set. The small order of magnitude of these measurements increases the significance of systematic error present in the results.

If we apply the previously found parameters, $\kappa = 4.44 \times 10^{-7}$ and $\sigma = 1.61$, the expected covariance matrix is given as:

$$\begin{bmatrix} 1.912 & 3.507 & 1.74 & 2.938 \\ 3.507 & 7.014 & 2.938 & 4.96 \\ 1.74 & 2.938 & 1.912 & 3.507 \\ 2.938 & 4.96 & 3.507 & 7.014 \end{bmatrix} \times 10^{-13}.$$

These results represent a lower limit for the uncertainties in the propagation channels. There were no moving objects, large temperature gradients or significant air flows. As the scale for the uncertainties are small, more accurate measurements may be required to estimate for realistic situation. It would be possible to measure the uncertainties across a series of rooms and scenarios in order to find the uncertainties relationship. These could then be used on unseen rooms.

5.5 Amplitude uncertainties

Similarly to propagation time uncertainties, the attenuation along the propagation channel can vary with time. As explained in Sec. 5.1, the channel attenuation in the conventional case is given as:

$$\bar{h}_{\epsilon a} \triangleq \frac{1}{\|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{p}}_a\|}. \tag{5.28}$$

There are several mechanisms that can cause a change in amplitude attenuation, many of which are frequency dependent. Many sources have different directivity patterns meaning that the amplitude observed at different directions from the source can vary significantly. Refraction of acoustic signals occurs primarily with temperature gradients. This occurs around air currents and underwater. Warmer mediums propagated sound faster, as such the wave front bends towards colder mediums. Wind speed gradients can act as refractors for sound waves over large distances [78]. The faster air currents at higher altitudes refract wave upwards or downwards depending on the direction of the sound wave, spreading or compacting the energy. Speaking into the wind causes a decrease in observed amplitude and vice versa. Occlusions can cause diffraction and spreading of the sound energy over a greater area. Whilst this does not affect the propagation time, the amplitude will decrease. This is most prominent when sound passes gaps which have comparable size to the sound wavelength. As such the effect is frequency dependent. Diffraction allows lower frequencies to travel further than higher frequencies through many environments.

5.5.1 Proposed model

We model the contribution of an amplitude uncertainty using a random variable, $\tilde{\mathbf{H}} \in \mathbb{R}^{M \times P}$:

$$\mathbf{D} \triangleq \overline{\mathbf{D}} \odot \tilde{\mathbf{H}} \odot \exp\left(j\omega_k \tilde{\mathbf{T}}\right), \tag{5.29}$$

where $\tilde{h}_{\epsilon a}$ represents the variation in the propagation path amplitude between a source at $\bar{\mathbf{p}}_a$ and a microphone at $\bar{\mathbf{m}}_\epsilon$. We do not impose a particular distribution on the amplitude uncertainties. In the case of no amplitude variations the uncertainties are unity, $\tilde{\mathbf{H}} = \mathbf{1}$.

A similar analysis to Sec. 5.3 of the propagation coefficients covariance is computed below.

The covariance of the propagation coefficients from (5.8) can be extended to include the new amplitude uncertainties term:

$$\left\langle \mathbf{D}\mathbf{D}^H \right\rangle = \left\langle \left(\overline{\mathbf{D}} \odot \tilde{\mathbf{H}} \odot \exp(j\omega_k \tilde{\mathbf{T}}) \right) \left(\overline{\mathbf{D}} \odot \tilde{\mathbf{H}} \odot \exp(j\omega_k \tilde{\mathbf{T}}) \right)^H \right\rangle. \quad (5.30)$$

In this case the elements of $\left\langle \mathbf{D}\mathbf{D}^H \right\rangle$ are given by

$$\left\langle \mathbf{D}\mathbf{D}^H \right\rangle_{\epsilon,\varphi} = \sum_{a=1}^{P} \bar{d}_{\epsilon a} \bar{d}^*_{\varphi a} \left\langle \tilde{h}_{\epsilon a} \tilde{h}_{\varphi a} \right\rangle \left\langle \exp\left(j\omega_k \left(\tilde{t}_{\epsilon a} - \tilde{t}_{\varphi a} \right) \right) \right\rangle. \quad (5.31)$$

Thus for each source, a, we require the covariance of the amplitude uncertainties: $\left\langle \tilde{h}_{\epsilon a} \tilde{h}_{\varphi a} \right\rangle$. In the following section we detail the amplitude uncertainties covariance matrix resulting from human head rotations.

5.5.2 Head rotations

In many applications for acoustic beamforming, the target source is a human talker. It is often assumed that acoustic sources radiate isotropically, but tests of the human head show that it is far from isotropic. The majority of the source energy propagates from the mouth in the direction the talker is facing, with the least energy propagating behind the head. This effect is frequency dependent, with higher frequencies being more directional than lower frequencies. The typical directivity pattern of a human talker is shown in Fig. 5.7. The directivity pattern is symmetrical around 0°, which

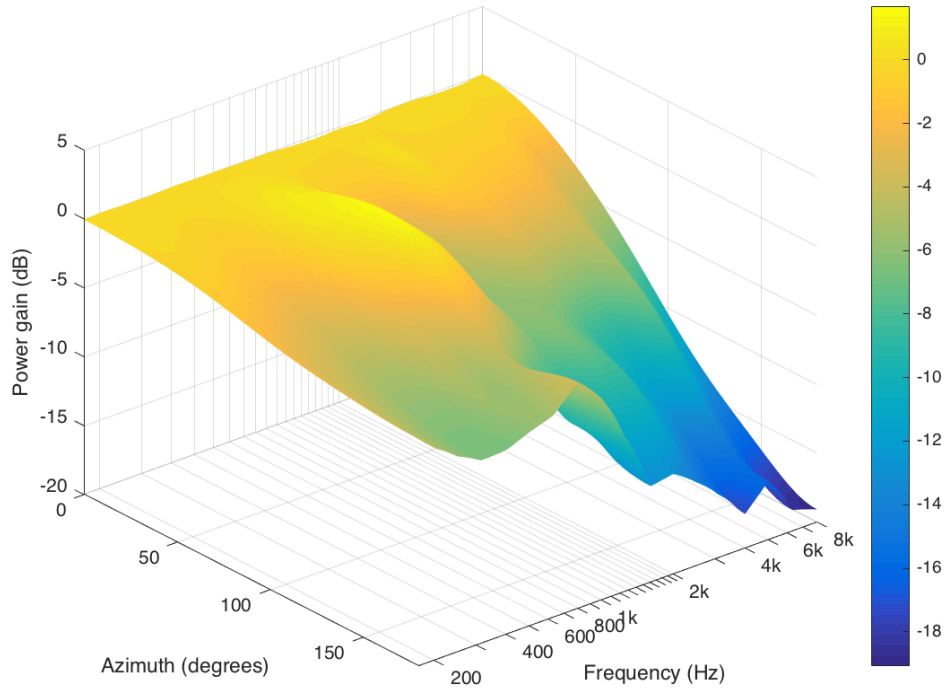

Figure 5.7: The typical directivity pattern from a human talker as a function of frequency.

corresponds to straight-ahead. The power have been normalised to give a gain of 0 dB at all frequencies in the straight-ahead direction, and do not, therefore, include the average power spectrum of speech. As frequency increases there is a larger difference between front and back amplitudes. At frequencies around 800 Hz, there is an increase in amplitude above 0 dB at azimuths around 75°.

If a human talker is facing in a random direction relative to a microphone, the effect on the amplitude of the propagation channel is substantial. When the talker is facing away from the microphone, there is an implicit low pass filter applied to the channel as well as a large reduction in observed amplitude. The directivity pattern

varies smoothly with azimuth, channels which are in a similar direction from the talker will have a similar amplitude attenuation. This is especially true at low frequencies.

Utilising signals in which the various sources have uncertain amplitudes can hinder the performance of multi-channel algorithms. By modeling the covariance in amplitude uncertainties as a result of a random head rotation, we can create algorithms that are robust to rotations.

5.5.2.1 Directivity pattern

We have utilised head directivity data from [79]. In the dataset the linear amplitude gain is defined with azimuths on a 15° grid and at discrete frequencies. We assume that the rate of change in response against azimuth and frequency is smooth. As the azimuth directivity pattern is periodic about the head we can express the gain against azimuth as a Fourier series. We compute the Fourier series coefficients, α_i, for the azimuth response at each frequency band given. The first 6 are used, energy in subsequent Fourier coefficients is small relative to the first 6. The energy of subsequent coefficients, $\tilde{h}_e$, is given as:

$$\tilde{h}_e(K) = 1 - \left(\frac{\sum_{i=0}^{K-1} |\alpha_i|^2}{\sum_{i=0}^{12} |\alpha_i|^2} \right),$$

where K is the number of coefficient used and 12 is the maximal number of coefficients. The error for all frequency bins is shown in Fig. 5.8. The coefficients at any arbitrary frequency band are estimated by using linear interpolation between neighbouring frequency coefficient sets. The use of the Fourier series allows us to compute the response at any azimuth by evaluating the Fourier series at the desired azimuth.

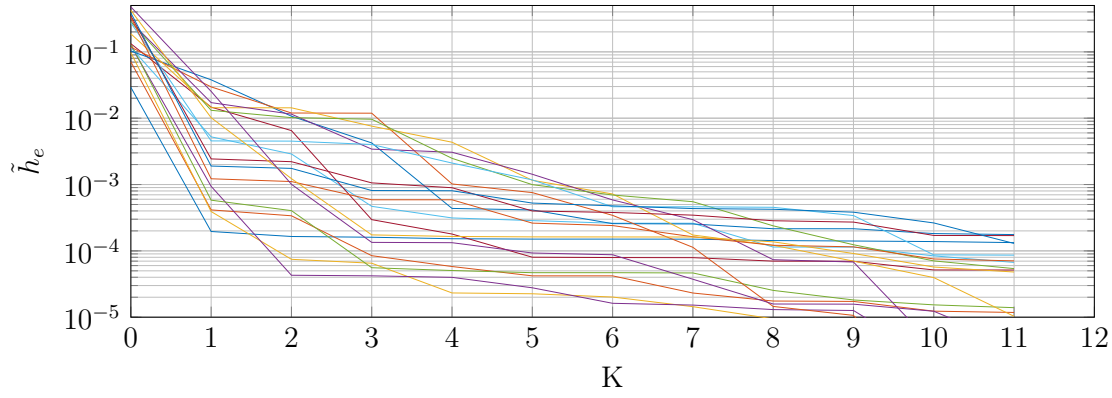

Figure 5.8: The error in energy of truncating the Fourier series coefficients of the head directivity pattern for each frequency bin specified.

5.5.2.2 Head rotation covariance

In the following sections we formulate the linear amplitude attenuation contribution to the propagation coefficients as follows:

$$\tilde{h}\left(\delta_{\phi_{\epsilon,a}}\right) = \sum_{i=0}^{5} \alpha_i \cos\left(k_i \delta_{\phi_{\epsilon,a}}\right), \tag{5.32}$$

where $\delta_{\phi_{\epsilon,a}}$ is the azimuth rotation angle of the channel from the source a to the microphone ϵ in radians, α_i are the Fourier coefficients of the directivity pattern and $k_i = i$ are the rotational frequencies.

We assume that the direction of the sources are independently randomly varying over time. The direction of each source is taken from a uniform distribution with mean μ_a and range $2\varrho_a$.

Linear expectation of head rotations To determine the linear expectation we compute the amplitude attenuation across all possible directions of the source, $\phi_a \in$

$[\mu_a - \varrho_a,\ \mu_a + \varrho_a]$:

$$\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right\rangle_{\phi_a} = \int_{\phi=\mu-\varrho}^{\mu+\varrho} \frac{1}{2\varrho}\tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) d\phi_a. \tag{5.33}$$

The expectation can be expanded as:

$$\begin{aligned}\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right\rangle_{\phi_a} &= \left\langle \sum_i \alpha_i \cos\left(k_i\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right)\right\rangle_{\phi_a} \\ &= \sum_i \alpha_i \left\langle \cos\left(k_i\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right)\right\rangle_{\phi_a}. \end{aligned} \tag{5.34}$$

The cosine expectation can be expressed using the following identity:

$$\begin{aligned}\left\langle \cos\left(k\phi_a + \delta\right)\right\rangle_{\phi_a} &= \frac{1}{2\varrho}\int_{\phi=\mu-\varrho}^{\mu+\varrho} \cos\left(k\phi_a + \delta\right) d\phi_a \\ &= \begin{cases} \frac{1}{2\varrho k}\left(\sin\left(k\left(\mu+\varrho\right)+\delta\right) - \sin\left(k\left(\mu-\varrho\right)+\delta\right)\right) & k \neq 0 \\ \cos\left(\delta\right) & k = 0 \end{cases} \\ &= \begin{cases} \frac{1}{\varrho k}\sin\left(k\varrho\right)\cos\left(k\mu+\delta\right) & k \neq 0 \\ \cos\left(\delta\right) & k = 0 \end{cases} \end{aligned} \tag{5.35}$$

The final result is expressed as:

$$\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right\rangle_{\phi_a} = \alpha_0 + \sum_{i=1} \alpha_i \frac{1}{k_i\varrho}\sin\left(k_i\varrho\right)\cos\left(k_i\mu + k_i\delta_{\phi_{\epsilon,a}}\right). \tag{5.36}$$

If the source rotation comes from uniformly distributions over a whole circle, $\phi_a \in [0,\ 2\pi]$ ($\varrho = \pi$), the expectation simplifies to:

$$\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right\rangle_{\phi_a} = \alpha_0. \tag{5.37}$$

Covariance of head rotations To compare the amplitude changes over two separate propagation paths we compute the double integral over the rotation of each source, $\phi_a \in [\mu_a - \varrho_a,\ \mu_a + \varrho_a]$ and $\phi_b \in [\mu_b - \varrho_b,\ \mu_b + \varrho_b]$:

$$\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_b + \delta_{\phi_{\varphi,b}}\right) \right\rangle_{\phi_{a,b}} = \iint_{\phi} \frac{1}{2\varrho_a} \frac{1}{2\varrho_b} \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_b + \delta_{\phi_{\varphi,b}}\right) d\phi_a d\phi_b. \tag{5.38}$$

We assume each source rotates independently of the others, $\langle \phi_a \phi_b \rangle = \langle \phi_a \rangle \langle \phi_b \rangle$ for $a \neq b$. In the case where the sources are different, $a \neq b$, the expectation can be separated into two parts and solved using (5.36):

$$\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_b + \delta_{\phi_{\varphi,b}}\right) \right\rangle_{\phi_{a,b}} = \left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \right\rangle_{\phi_a} \left\langle \tilde{h}\left(\phi_b + \delta_{\phi_{\varphi,b}}\right) \right\rangle_{\phi_b}. \tag{5.39}$$

In the case where the two paths start from the same source, $a = b$, the expectation is more complex:

$$\begin{aligned}
&\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_b + \delta_{\phi_{\varphi,b}}\right) \right\rangle_{\phi_{a,b}} \\
&\quad = \left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_a + \delta_{\phi_{\varphi,a}}\right) \right\rangle_{\phi_a} \\
&\quad = \left\langle \sum_i \alpha_i \cos\left(k_i\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right) \sum_j \alpha_j \cos\left(k_j\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\right) \right\rangle_{\phi_a} \\
&\quad = \left\langle \sum_i \sum_j \alpha_i \alpha_j \cos\left(k_i\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right) \cos\left(k_j\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\right) \right\rangle_{\phi_a} \\
&\quad = \sum_i \sum_j \alpha_i \alpha_j \left\langle \cos\left(k_i\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right) \cos\left(k_j\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\right) \right\rangle_{\phi_a}.
\end{aligned} \tag{5.40}$$

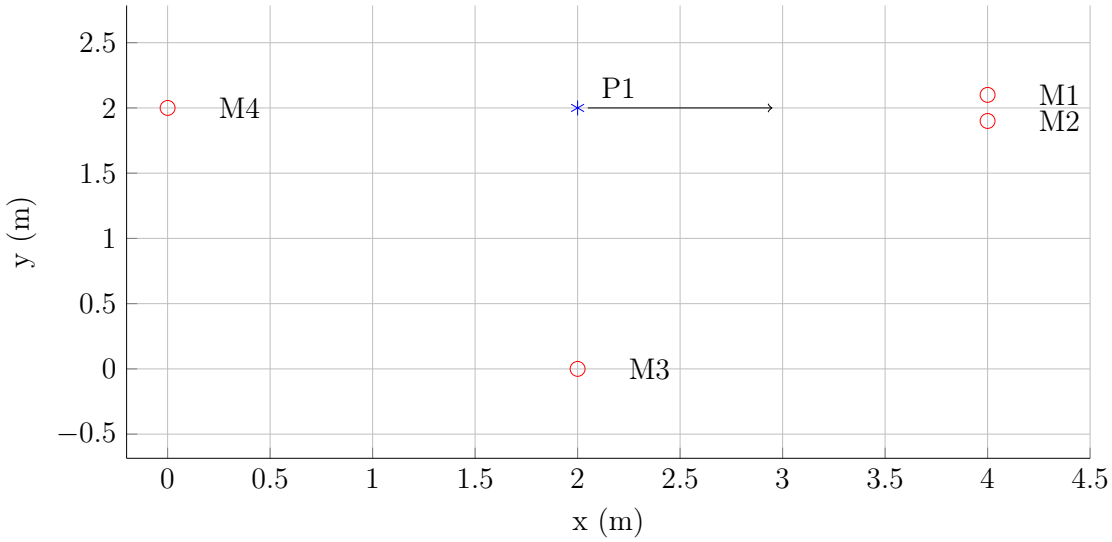

Figure 5.9: Layout of the amplitude uncertainties example, $P1$ indicates the talker location, M_ϵ indicates the location of microphone ϵ. $P1$ is pointing, on average, in the positive x-axis.

Use of the double cosine rule simplifies the above to:

$$\begin{aligned}&\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\right\rangle_{\phi_a} \\ &\quad = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j \left\langle \cos\left(\phi_a\left(k_i + k_j\right) + k_i\delta_{\phi_{\epsilon,a}} + k_j\delta_{\phi_{\varphi,a}}\right)\right. \\ &\qquad \left. + \cos\left(\phi_a\left(k_i - k_j\right) + k_i\delta_{\phi_{\epsilon,a}} - k_j\delta_{\phi_{\varphi,a}}\right)\right\rangle_{\phi_a}.\end{aligned} \tag{5.41}$$

Both of the cosines expectations can be expanded using the rule given in 5.35.

5.5.2.3 Illustrated example

Consider the array geometry shown in Fig. 5.9. The single source, $P1$, is assumed to rotate randomly, with an average direction of the positive x-axis. The resulting amplitude changes due to the random rotation will differ between each microphone. The amplitudes of the channels to the two microphones in front of the source, $M1$

and $M2$, will be well correlated as the channels are close in direction from the source. However, microphones that have large differences in their channel directions from the source, will be less correlated, such as $M3$ and $M4$. If we assume the source rotation follows a uniform distribution with a certain range, $\phi_1 \in [-\varrho_1,\ \varrho_1]$, we can compute the amplitude uncertainty covariances. The covariance in amplitude uncertainties along the paths from $P1$ to $M1$ and $P1$ to microphone φ is taken from 5.41 and notated as $\left\langle \tilde{h}_{11}\tilde{h}_{\varphi 1} \right\rangle = \left\langle \tilde{h}\left(\phi_1 + \delta_{\phi_{1,1}}\right) \tilde{h}\left(\phi_1 + \delta_{\phi_{\varphi,1}}\right) \right\rangle_{\phi_1}$, where $\delta_{\phi_{1,1}} = 0.05\,rad$, $\delta_{\phi_{2,1}} = -0.05\,rad$, $\delta_{\phi_{3,1}} = -\frac{\pi}{2}\,rad$ and $\delta_{\phi_{4,1}} = \pi\,rad$. The amplitude uncertainties covariance for three different distribution ranges, $2\varrho_1 = \left[\frac{\pi}{4},\ \frac{\pi}{2},\ \pi\right]$, are shown in Fig. 5.10. For all the source rotation ranges, the amplitude uncertainties to the microphones in front of the source, $M1$ and $M2$, are strongly correlated and indeed the plots of $\left\langle \tilde{h}_{11}\tilde{h}_{11} \right\rangle$ and $\left\langle \tilde{h}_{11}\tilde{h}_{21} \right\rangle$ are coincident on all three graphs. However the correlation between $M3$ and $M1$ is much lower especially at high frequencies. This effect is larger when the channels become more separated, and the correlation between $M1$ and $M4$ is even lower. As the source rotation range increases, the amplitude uncertainties start to follow a similar trend through the different channels.

If we were to build multi-channel algorithms that rely on the amplitudes through each channel, using $M1$ and $M2$ would be more reliable than those of $M1$ and $M4$.

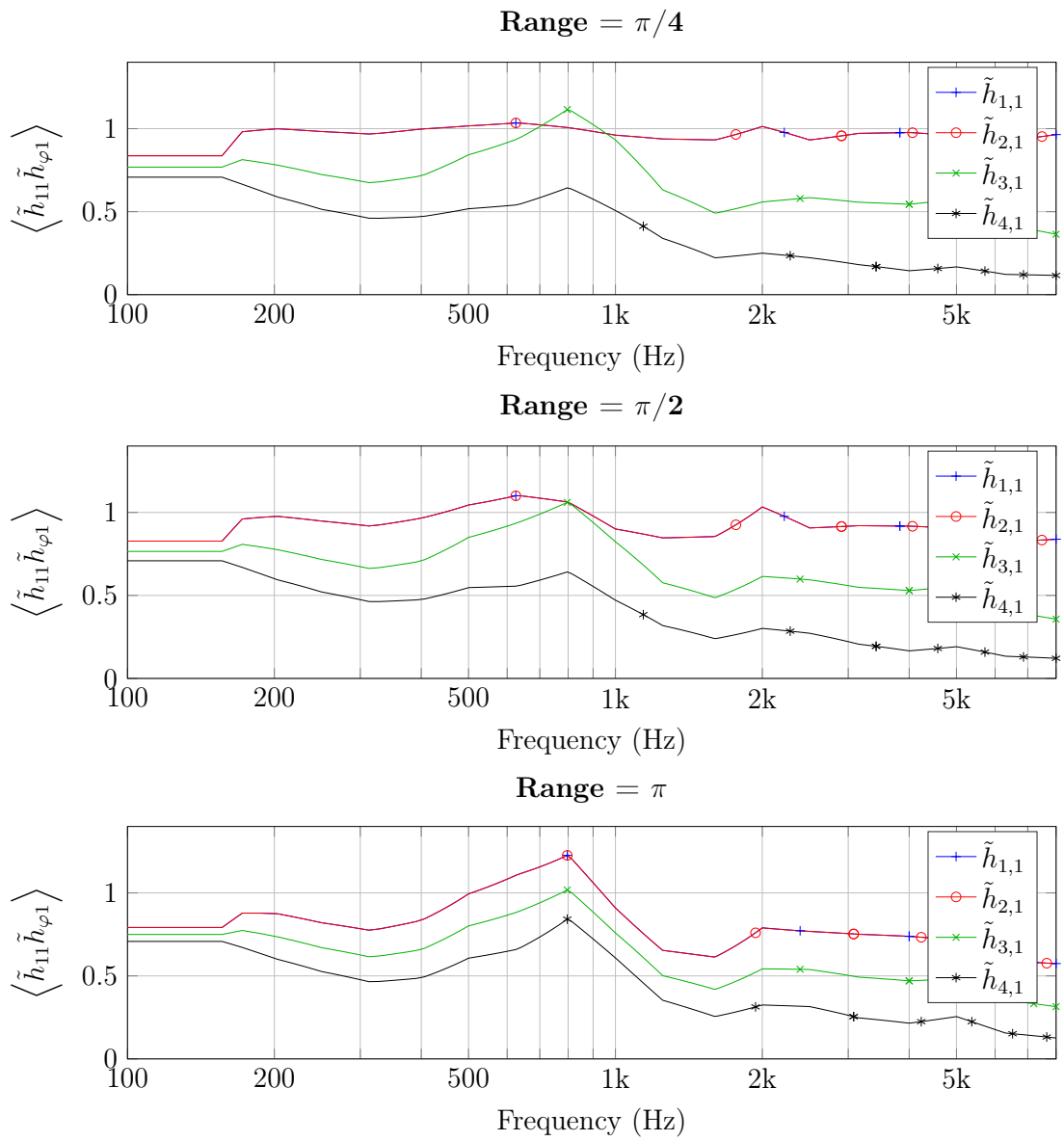

Figure 5.10: The covariance in amplitude uncertainties across different channels and source rotation ranges.

5.6 Quartic extension

In order to form the robust power domain beamformer that will be described in Chapter 8 we need to extend the uncertainties framework to include expectations of 6-dimensional quartic covariance terms of the form $\left\langle d_{\epsilon,a} d^*_{\varphi,a} d_{\rho,b} d^*_{\tau,b} \right\rangle$.

The uncertainties model is the same as that used above:

$$d_{\epsilon,a} \;=\; \overline{d}_{\epsilon,a} \tilde{h}_{\epsilon,a} \exp\left(j\omega \tilde{t}_{\epsilon,a}\right).$$

The quartic term can be expanded as:

$$\begin{aligned} \left\langle d_{\epsilon,a} d^*_{\varphi,a} d_{\rho,b} d^*_{\tau,b} \right\rangle \;=\; & \overline{d}_{\epsilon,a} \overline{d}^*_{\varphi,a} \overline{d}_{\rho,b} \overline{d}^*_{\tau,b} \left\langle \tilde{h}_{\epsilon,a} \tilde{h}_{\varphi,a} \tilde{h}_{\rho,b} \tilde{h}_{\tau,b} \right\rangle \\ & \left\langle \exp\left(j\omega_k \left(\tilde{t}_{\epsilon,a} - \tilde{t}_{\varphi,a} + \tilde{t}_{\rho,b} - \tilde{t}_{\tau,b}\right)\right)\right\rangle. \end{aligned} \tag{5.42}$$

We therefore need to derived expressions for the two expectation terms: $\left\langle \tilde{h}_{\epsilon,a} \tilde{h}_{\varphi,a} \tilde{h}_{\rho,b} \tilde{h}_{\tau,b} \right\rangle$ and $\left\langle \exp\left(j\omega_k \left(\tilde{t}_{\epsilon,a} - \tilde{t}_{\varphi,a} + \tilde{t}_{\rho,b} - \tilde{t}_{\tau,b}\right)\right)\right\rangle$.

5.6.1 Time uncertainties

The time uncertainties term can be expanding in a similar fashion to the quadratic version. The exponential term contains a sum of Gaussian random variables, which itself is a Gaussian. The expectation forms a log-normal distribution in which the variance is given as the square of the exponent. Therefore, we have:

$$\begin{aligned} & \left\langle \exp\left(j\omega_k \left(\tilde{t}_{\epsilon,a} - \tilde{t}_{\varphi,a} + \tilde{t}_{\rho,b} - \tilde{t}_{\tau,b}\right)\right)\right\rangle \\ & \quad = \exp\left(\frac{-\omega_k^2 \left\langle \left(\tilde{t}_{\epsilon,a} - \tilde{t}_{\varphi,a} + \tilde{t}_{\rho,b} - \tilde{t}_{\tau,b}\right)^2 \right\rangle}{2}\right) \\ & \quad = \exp\left(\frac{-\omega_k^2 \sigma_T^2}{2}\right), \end{aligned} \tag{5.43}$$

where the variance is expanded as:

$$\begin{aligned}\sigma_T^2 = \big\langle &\tilde{t}_{\epsilon,a}^2 + \tilde{t}_{\varphi,a}^2 + \tilde{t}_{\rho,b}^2 + \tilde{t}_{\tau,b}^2 - 2\tilde{t}_{\epsilon,a}\tilde{t}_{\varphi,a} + 2\tilde{t}_{\epsilon,a}\tilde{t}_{\rho,b} \\ &-2\tilde{t}_{\epsilon,a}\tilde{t}_{\tau,b} - 2\tilde{t}_{\varphi,a}\tilde{t}_{\rho,b} + 2\tilde{t}_{\varphi,a}\tilde{t}_{\tau,b} - 2\tilde{t}_{\rho,b}\tilde{t}_{\tau,b}\big\rangle .\end{aligned} \tag{5.44}$$

Each individual term can be taken from the time uncertainties covariance matrix, $\langle \tilde{t}_{\epsilon,a}\tilde{t}_{\varphi,b}\rangle$.

5.6.2 Amplitude uncertainties

The amplitude uncertainties term, $\left\langle \tilde{h}_{\epsilon,a}\tilde{h}_{\varphi,a}\tilde{h}_{\rho,b}\tilde{h}_{\tau,b}\right\rangle$, can be expressed using the head rotation model and the uniform distribution used previously:

$$\left\langle \tilde{h}_{\epsilon,a}\tilde{h}_{\varphi,a}\tilde{h}_{\rho,b}\tilde{h}_{\tau,b}\right\rangle = \left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\tilde{h}\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\tilde{h}\left(\phi_b + \delta_{\phi_{\rho,b}}\right)\tilde{h}\left(\phi_b + \delta_{\phi_{\tau,b}}\right)\right\rangle_{\phi_{a,b}} . \tag{5.45}$$

In the case where the sources are different, $a \neq b$, the expectation can be expressed as the product of two terms and solved using (5.41):

$$\begin{aligned}&\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\tilde{h}\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\tilde{h}\left(\phi_b + \delta_{\phi_{\rho,b}}\right)\tilde{h}\left(\phi_b + \delta_{\phi_{\tau,b}}\right)\right\rangle_{\phi_{a,b}} \\ &\quad = \left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\tilde{h}\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\right\rangle_{\phi_a}\left\langle \tilde{h}\left(\phi_b + \delta_{\phi_{\rho,b}}\right)\tilde{h}\left(\phi_b + \delta_{\phi_{\tau,b}}\right)\right\rangle_{\phi_b} .\end{aligned} \tag{5.46}$$

In the case where the two paths start from the same source, $a = b$, the expectation is more complex. Each term can be replaced by its cosine series, $\tilde{h}\left(\delta_{\phi_{\epsilon,a}}\right) =$

$\sum_i \alpha_i \cos\left(k_i \delta_{\phi_{\epsilon,a}}\right)$:

$$\begin{aligned}
&\left\langle \tilde{h}\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right) \tilde{h}\left(\phi_a + \delta_{\phi_{\varphi,a}}\right) \tilde{h}\left(\phi_b + \delta_{\phi_{\rho,b}}\right) \tilde{h}\left(\phi_b + \delta_{\phi_{\tau,b}}\right)\right\rangle_{\phi_a} \qquad (5.47)\\
&= \sum_{i,j,k,l} \alpha_i \alpha_j \alpha_k \alpha_l \left\langle \cos\left(k_i\left(\phi_a + \delta_{\phi_{\epsilon,a}}\right)\right) \cos\left(k_j\left(\phi_a + \delta_{\phi_{\varphi,a}}\right)\right)\right.\\
&\qquad \left. \cos\left(k_k\left(\phi_a + \delta_{\phi_{\rho,a}}\right)\right) \cos\left(k_l\left(\phi_a + \delta_{\phi_{\tau,a}}\right)\right)\right\rangle_{\phi_a}.
\end{aligned}$$

The cosine expectation terms can be solved using a quartic cosine identity:

$$\begin{aligned}
\cos(a)\cos(b)\cos(c)\cos(d) \;=\; &\frac{1}{8}\left(\cos(a+b+c+d) + \cos(a+b-c-d)\right.\\
&+ \cos(a+b+c-d) + \cos(a+b-c+d)\\
&+ \cos(a-b+c+d) + \cos(a-b-c-d)\\
&\left.+ \cos(a-b+c-d) + \cos(a-b-c+d)\right). \qquad (5.48)
\end{aligned}$$

Thus, the quartic head rotation term can be decomposed into a sum of linear cosine terms. Each of the cosines expectations can be calculated using the rule given in 5.35.

5.7 Conclusions

We have discussed conventional acoustic propagation models and described that in practical scenarios they may cause performance degradation in spatial filtering algorithms. We derived new models which include contributions for uncertainties in the phase and amplitude components of the propagation coefficients. The parameters of which were estimated from experimental data. The propagation coefficients model is extended to incorporate head rotations of human talkers. The directivity patterns are

used to find the expected covariances across different paths. In which case the propagation coefficients can be used in the design of beamformers to ensure robustness to both time and amplitude uncertainties.

Chapter 6

Beamforming

This chapter gives an outline of existing beamformer algorithms, both robust and non-robust. Beamforming is a popular multichannel signal processing technique used in conjunction with microphone arrays to spatially filter a sound field. It is commonly employed in applications such as hands-free telephony and teleconferencing for the acquisition of distant audio.

The first beamformers were implemented in the time domain. The delay-and-sum beamformer [80] aligns the desired source in each microphone signal and sums them together. This was extended to the Matched Filter [81, 82] which matches the amplitude as well as the time difference. In the STFT domain, the Minimum Variance Distortionless Beamformer (MVDR) [83, 84] minimises the output power of the beamformer whilst constraining a distortionless response to the steering vector for the desired source. The use of multiple constraints extended the MVDR beamformer to the linear constraint minimum variance (LCMV) beamformer [85]. The LCMV beamformer has been extended with different types of constraints for different applications [86, 87]. The Generalised Sidelobe Canceller [88] separates the constraints from the output variance minimisation.

In the presence of steering vector errors, the performance of conventional beamformers degrades. Robust beamformers are designed to limit the fall in performance

when errors are present in the steering vectors. Diagonally loaded beamformers use regularisation to smooth the beamformer response [89]. Derivative constraints extend the LCMV beamformer for less sensitivity to steering vector errors. Worst-case optimisation beamformers [90] assume the steering vector is contained within an uncertainty set and constrain the response of the whole set. Various other uncertainty sets have been considered [91, 92, 93] to better approximate the set of steering vector errors. Probabilistically constrained beamformers [94] extend the notion of the uncertainty set to a truncated distribution.

In the following chapter we first introduce the beamforming signal model used within the remainder of the thesis. Next we introduce the conventional non-robust beamformers. We then introduce the various methods of robust beamforming from the literature.

6.1 Signal model and problem formulation

Consider the case of P speech sources and M microphones in a noisy environment. The received signal at all M microphones in the short time Fourier transform (STFT) domain can be written as:

$$\mathbf{x}\left[l,k\right] \;=\; \bar{\mathbf{D}}\mathbf{s}\left[l,k\right] + \mathbf{v}\left[l,k\right], \tag{6.1}$$

where $\mathbf{x}, \mathbf{v} \in \mathbb{C}^{M\times 1}$, $\mathbf{s} \in \mathbb{C}^{P\times 1}$ and $\bar{\mathbf{D}} \in \mathbb{C}^{M\times P}$; l is a discrete-time frame index, k is a discrete-frequency index, s_p is the p-th source signal, as discussed at the start of Chapter 5 - Sec. 5.1. $\bar{d}_{\epsilon a}$ is an element of $\bar{\mathbf{D}}$ and represents the conventional deterministic complex channel propagation coefficient from source a to microphone ϵ, it is further described in (5.4). From this point on $\tilde{s} = s_1$ is the desired source. The source and noise STFT coefficients have independent, zero-mean real and imaginary parts. It is assumed that all the sources are independent, $\left\langle \mathbf{s}\mathbf{s}^H \right\rangle = \mathrm{diag}\left(\left\langle |s_1|^2 \right\rangle \ldots \left\langle |s_P|^2 \right\rangle\right)$. The additive noise component incident at microphone ϵ, v_ϵ, comprises sensor noise, v_η, and spatially diffuse acoustic noise, v_κ.

In each frequency band, k, we derive the beamformer output, y, as the weighted sum of the array data, $\mathbf{x}$, in order to optimally extract the wanted source:

$$y\left[l,k\right] = \mathbf{w}\left[k\right]^{H}\mathbf{x}\left[l,k\right], \tag{6.2}$$

where $\mathbf{w}(k)$ is a vector of complex-valued weights and $(\cdot)^{H}$ is the Hermitian transpose.

Since each frequency band is processed independently, the frequency index, k, will normally be omitted in the remainder of this thesis for clarity. Adaptive beamforming studies the case where the weights can vary with respect to time, $\mathbf{w}\left[l,k\right]$. In this thesis we concentrate on optimum beamforming with a time-invariant weight vector. For this reason, the time index, l, is also normally omitted.

6.1.1 SNR gain

The SNR gain of the beamformer is denoted as the gain in power from the desired steering vector compared to the power gain for noise and other interference sources:

$$G_a = \frac{\mathbf{w}^H \left\langle |\tilde{s}|^2 \right\rangle \left\langle \mathbf{d}_1 \mathbf{d}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}},$$

where $\mathbf{d}_1$ is the first column of $\mathbf{D}$ and is the steering vector for the desired source and $\mathbf{V} = \left\langle \mathbf{x}\mathbf{x}^H \right\rangle - \left\langle |\tilde{s}|^2 \right\rangle \left\langle \mathbf{d}_1 \mathbf{d}_1^H \right\rangle$ is the covariance for the array data without the desired source.

6.1.2 White noise gain

The white noise gain is the SNR gain of the desired source through the array in the presence of only spatially white noise. The higher the gain, the greater the desired source power and therefore the SNR. It is derived from the SNR gain, in the presence of spatially white noise, in which case $\mathbf{V} = \mathbf{I}$, and unity source power, $\left\langle |\tilde{s}|^2 \right\rangle = 1$. It

is defined as below [95]:

$$G_w = \frac{\mathbf{w}^H \left\langle \mathbf{d}_1 \mathbf{d}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{w}} \leq M. \tag{6.3}$$

Any errors that occur in the geometry and are uncorrelated will present in a similar fashion to white noise [95]. As such the white noise gain is a useful measure of robustness against uncorrelated errors.

6.1.3 Beampatterns

We illustrate many of the beamformers by showing their beampattern, such as in Fig. 6.2. This represents the gain of the array to each position in the room. The sources and microphones are considered on a flat plane. The steering vector for all points in the room, $\mathbf{d}$, are computed. The gain through the array is computed using the beamformer weights, $\mathbf{w}$, as $\mathbf{w}^H \left\langle \mathbf{d}\mathbf{d}^H \right\rangle \mathbf{w}$ over all frequencies. The A-weighted gain is calculated and converted into dB, which is represented as intensity.

6.2 Data-independent beamformers

This section describes beamformers that do not utilise the received sensor signals, just the location of the array elements and sources. Subsequent sections increase the amount of information we assume is known about the setup.

Early beamformers operated in the time domain. A prime example is that of the delay-and-sum beamformer [80]. It exploits the time differences of arrival of the desired source between the microphones in the array. By calculating the differences in time of arrival, using 5.1, we can add a compensating time delay to the microphone signals and then sum the results. We have high gain at our desired point as all the signals are in phase, whereas other angles are out of phase and are attenuated.

The performance was improved in the Matched Filter Beamformer [81, 82], in which the time domain signals are filtered with the impulse response for the desired

source in both the forwards and backwards directions. This removes any phase differences for the desired source, and is subsequently scaled for unity gain.

An equivalent delay-and-sum beamformer can be implemented in the STFT domain by using a phase shift in each frequency bin to align all the signals from the desired steering vector. We first calculate the propagation coefficients for the desired source to each microphone using the conventional model from 5.5 in Chapter 5. The weights are given as the phase shift of the propagation coefficients:

$$\mathbf{w} = \frac{1}{\left|\bar{\mathbf{d}}_1\right|^{\cdot 1}} \odot \bar{\mathbf{d}}_1 = \begin{bmatrix} \frac{\bar{d}_{11}}{|\bar{d}_{11}|} \\ \vdots \\ \frac{\bar{d}_{M1}}{|\bar{d}_{M1}|} \end{bmatrix}, \tag{6.4}$$

where $\left|\bar{\mathbf{d}}_1\right|^{\cdot 1} = \left[\bar{d}^1_{1,1} \dots \bar{d}^1_{M,1}\right]^T$.

The Matched Filter can be implemented in the STFT domain by constraining the response from our desired source to be undistorted [96]:

$$\mathbf{w}^H \bar{\mathbf{d}}_1 = 1. \tag{6.5}$$

The Matched Filter satisfies the above equation when the weights apply the inverse channel gain [82]:

$$\mathbf{w} = \frac{1}{\left|\bar{\mathbf{d}}_1\right|^{\cdot 2}} \odot \bar{\mathbf{d}}_1, \tag{6.6}$$

For example, using a two element linear microphone array. If we sum both microphone signals (sum based beamformer), we amplify all sound from straight in front and behind the array. If we have a source at 45° to the front of the array, we can use a delay-and-sum beamformer to steer the response to 45°. For each beamformer we can find the A-weighted [18] signal gain (dB), $\mathbf{w}^H \left\langle |\tilde{s}|^2 \right\rangle \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}$, from every position in a room. The two resulting beampatterns are shown in Fig. 6.1. The main lobe of the

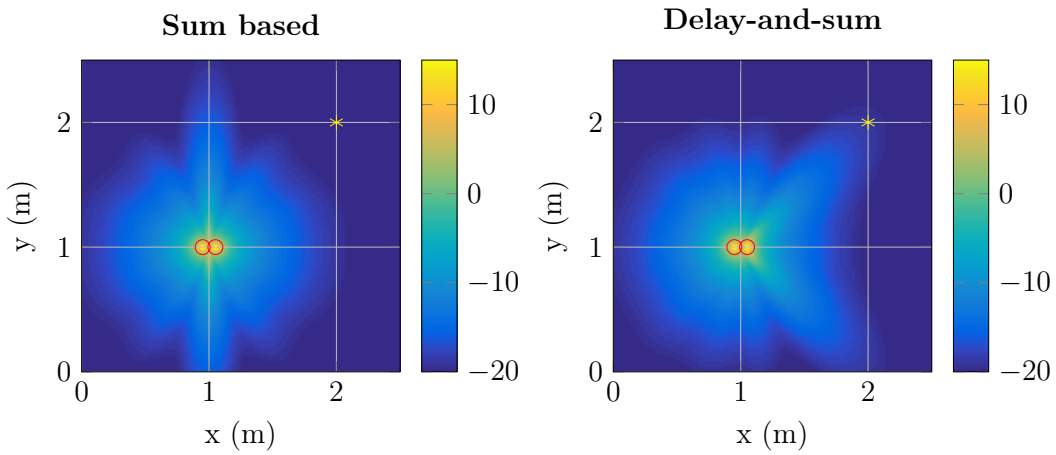

Figure 6.1: Spatial A-weighted gain of the array (dB) for the sum based beamformer (left) and the delay-and-sum beamformer (right), the desired source is located at $(2, 2)$.

delay-and-sum beamformer has rotated round relative to the sum based beamformer. It now matches the location of the source position. The gain of the desired source is greater without amplifying excess noise from other areas.

The Matched Filter beamformer ensures that there is no distortion to the direction of interest, however it cannot directly remove interference which originate from specific location. The performance is good for suppressing uncorrelated noise, such as sensor noise, but is not as successful for use with correlated noise.

The delay-and-sum beamformer has optimal white noise gain [95]. However it will not perform as well in the presence of correlated noise, such as diffuse noise or interference. In cases of correlated noise, alternative beamformers will offer suppression of noise whilst maintaining gain of the desired source.

6.3 Minimum variance distortionless response

A commonly used beamformer design which exploits knowledge of the speech covariance is the minimum variance distortionless response (MVDR) beamformer. It

maintains unity gain at the direction of the desired source whilst minimising the output variance [97, 83]:

$$\min_{\mathbf{w}} \mathbf{w}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle \mathbf{w} \qquad \text{subject to } \mathbf{w}^H \bar{\mathbf{d}}_1 = 1. \tag{6.7}$$

As with many other classical data-dependent beamformers, the weights are derived as a function of the array data covariance, $\left\langle \mathbf{x}\mathbf{x}^H \right\rangle$, where $\langle \cdots \rangle$ denotes the expected value. The resulting solution to (6.7) is [83]:

$$\mathbf{w}_{MVDR} = \left(\bar{\mathbf{d}}_1^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \bar{\mathbf{d}}_1 \right)^{-1} \left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \bar{\mathbf{d}}_1. \tag{6.8}$$

Similarly to the delay-and-sum beamformer, we constrain the direction of the desired source to be distortionless. However we also seek to minimise the output variance. The output variance of the beamformer is made of interference and noise in addition to the desired source. The number of microphones in the array determines the length of the weights vector, $\mathbf{w}$. Therefore the degrees of freedom to minimise the output variance is limited by the number of microphones in the array.

Whilst interference sources are not directly constrained in the optimum weights, the MVDR beamformer can successfully suppress interference sources. The interferers will be reflected in the array data covariance, $\left\langle \mathbf{x}\mathbf{x}^H \right\rangle$, which allows the MVDR beamformer to spurs them. As the output variance incorporates correlated noise terms, the performance will surpass that of the delay-and-sum beamformer in these cases.

It may be more intuitive to minimise the noise output power, which is equivalent to:

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{V} \mathbf{w} \qquad \text{subject to } \mathbf{w}^H \bar{\mathbf{d}}_1 = 1. \tag{6.9}$$

It can be shown that minimising the output variance achieves the same result as minimising the noise output power. First we expand the expected output power:

$$\begin{aligned}
\left\langle y^2 \right\rangle &= \left\langle \mathbf{w}^H \mathbf{x}\mathbf{x}^H \mathbf{w} \right\rangle \\
&= \mathbf{w}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle \mathbf{w} \\
&= \mathbf{w}^H \left\langle \left(\bar{\mathbf{D}}\mathbf{s} + \mathbf{v}\right)\left(\bar{\mathbf{D}}\mathbf{s} + \mathbf{v}\right)^H \right\rangle \mathbf{w} \\
&= \mathbf{w}^H \left\langle \left(\bar{\mathbf{D}}\mathbf{s}\mathbf{s}^H\bar{\mathbf{D}}^H + \bar{\mathbf{D}}\mathbf{s}\mathbf{v}^H + \mathbf{v}\mathbf{s}^H\bar{\mathbf{D}}^H + \mathbf{v}\mathbf{v}^H\right)\right\rangle \mathbf{w}
\end{aligned}$$

Using $\left\langle \mathbf{s}\mathbf{s}^H \right\rangle = \text{diag}\left(\left\langle |s_1|^2 \right\rangle \ldots \left\langle |s_P|^2 \right\rangle\right)$, we can simplify to:

$$\begin{aligned}
\left\langle y^2 \right\rangle &= \mathbf{w}^H \left\langle \left(\bar{\mathbf{D}}\mathbf{s}\mathbf{s}^H\bar{\mathbf{D}}^H + \mathbf{v}\mathbf{v}^H\right)\right\rangle \mathbf{w} \\
&= \mathbf{w}^H \left(\sum_p \bar{\mathbf{d}}_p \left\langle |s_p|^2 \right\rangle \bar{\mathbf{d}}_p^H + \left\langle \mathbf{v}\mathbf{v}^H \right\rangle \right) \mathbf{w} \\
&= \mathbf{w}^H \bar{\mathbf{d}}_1 \left\langle |\tilde{s}|^2 \right\rangle \bar{\mathbf{d}}_1^H \mathbf{w} + \mathbf{w}^H \left(\sum_{p=2} \bar{\mathbf{d}}_p \left\langle |s_p|^2 \right\rangle \bar{\mathbf{d}}_p^H + \left\langle \mathbf{v}\mathbf{v}^H \right\rangle \right) \mathbf{w} \\
&= \left\langle |\tilde{s}|^2 \right\rangle + \mathbf{w}^H \left(\sum_{p=2} \bar{\mathbf{d}}_p \left\langle |s_p|^2 \right\rangle \bar{\mathbf{d}}_p^H + \left\langle \mathbf{v}\mathbf{v}^H \right\rangle \right) \mathbf{w}.
\end{aligned}$$

As the constraint sets the gain for the desired source, $\mathbf{w}^H\mathbf{d}_1 = 1$, minimising the output variance only minimises the remaining noise and interference power.

6.4 SNR optimal beamformer

The SNR of a linear beamformer can be derived as the ratio of the output speech power to the output noise and interference power [98]:

$$\text{SNR} = \frac{\left\langle |\tilde{s}|^2 \right\rangle \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}}.$$

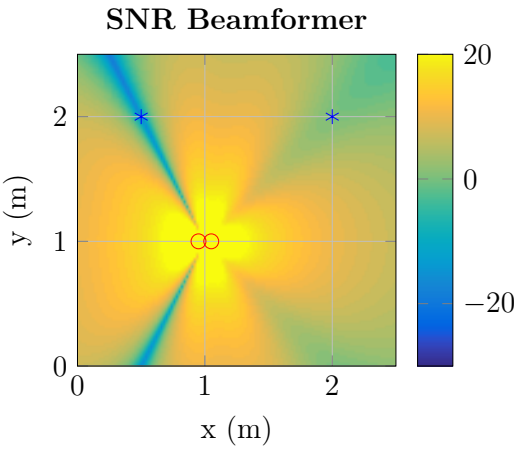

Figure 6.2: Spatial A-weighted signal gain (dB) for the SNR optimal beamformer, the desired source is located at $(2, 2)$ and an interferer is located at $(0.5, 2)$.

Optimising the weights to maximise the above SNR resulting the SNR optimal beamformer [95]. The weights are derived from the eigendecomposition of the inverse noise matrix multiplied by the source matrix:

$$\begin{aligned} \mathbf{B} &= \mathbf{V}^{-1} \left\langle |\tilde{s}|^2 \right\rangle \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \\ \mathbf{B}\mathbf{w} &= \lambda_{max} \mathbf{w}. \end{aligned} \tag{6.10}$$

A full derivation is shown in Appendix A. The optimal weights, $\mathbf{w}$, are the eigenvector that corresponds to the largest eigenvalue of $\mathbf{B}$, λ_{max}. The resulting beampattern is shown in Fig. 6.2. The gain is unconstrained, as such it is commonly constrained for unity gain for the desired source, $\mathbf{w}^H \bar{\mathbf{d}}_1 = 1$. The beamformer attenuates the interference source whilst maintaining unity gain for the desired source.

It can be shown that the SNR-optimal weights are equivalent to those of the MVDR beamformer [99], but they behave differently in the presence of steering vector errors [100]. As such the SNR in each case is the same. The difference being that the weights of the MVDR beamformer are scaled to force the distortionless constraint on the desired source location.

6.5 Linearly constrained minimum variance

Linearly constrained minimum variance (LCMV) beamformers are an extension of the MVDR beamformer. The formulation allows for multiple linear constraints, which allows for constraints on steering vectors from multiple different directions [85]. For example we can maintain distortionless response to the desired source whilst setting the gain to zero in the direction of an interference source. The problem is formulated as follows:

$$\min_{\mathbf{w}} \mathbf{w}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle \mathbf{w} \qquad \text{subject to } \mathbf{C}^H\mathbf{w} = \mathbf{f}, \tag{6.11}$$

where $\mathbf{C} = \left[\bar{\mathbf{d}}_1\ \bar{\mathbf{d}}_2 \dots \bar{\mathbf{d}}_L\right]$ are the propagation coefficients for each of L constrained positions and $\mathbf{f} = [1\ 0 \dots 0]^T$ and the constrained response gains. The resulting weights that solve the above are [85]:

$$\mathbf{w}_{LCMV} = \left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \left(\mathbf{C}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \mathbf{C}\right)^{-1} \mathbf{C}\mathbf{f}. \tag{6.12}$$

Given L constraints and M microphones, there are $M - L$ degrees of freedom in order to reduce the output variance. The more constraints present, the fewer degrees of freedom that can be used to reduce the output variance. This causes a tradeoff between reducing interference and reducing noise. As such the overall performance of the MVDR beamformer with a single constraint can, in some cases, surpass that of the LCMV beamformer.

For example, we consider the case of two sources of equal power, one of which in an interferer, located at [0.5, 2] m. We apply both the MVDR and LCMV to the situation. Both are constrained for distortionless response to the desired source, and the LCMV beamformer constrains no gain at the location of the interferer. The two resulting beampatterns are shown in Fig. 6.3. The two beampatterns are very similar. The MVDR beamformer accurately suppressed the interference source. However the suppression on the LCMV beamformer is deeper and more narrow. This will be more

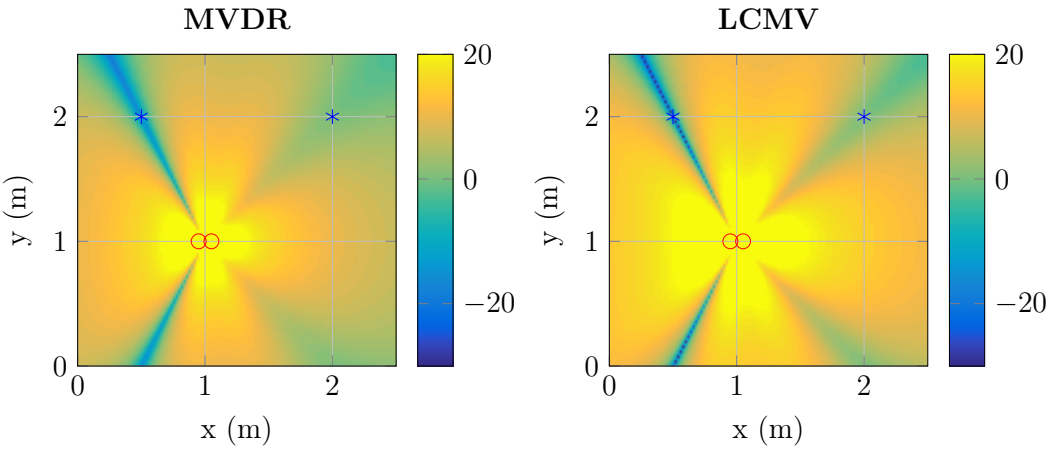

Figure 6.3: Spatial A-weighted signal gain (dB) for the MVDR beamformer (left) and the LCMV beamformer (right), the desired source is located at $(2, 2)$ and an interferer is located at $(0.5, 2)$.

sensitive to propagation coefficient errors. There can be a tradeoff between interference suppression and noise suppression; in this example, the LCMV beamformer has slightly lower white noise gain (A-weighted: 1.84) than the MVDR case (A-weighted: 1.85), which will increase noise that is uncorrelated across the microphones. Having two constraints, which are close together in terms of location, will increase the weights magnitude and in turn reduce the white noise gain.

The LCMV beamformer weights can be decomposed into two components, the constrained subspace and an orthogonal subspace:

$$\mathbf{w}_{LCMV} \quad = \quad \left(\mathbf{P}_{\mathbf{C}} + \mathbf{P}_{\mathbf{C}}^{\perp}\right) \mathbf{w}_{LCMV},$$

where $\mathbf{P}_{\mathbf{C}} \triangleq \mathbf{C}\left(\mathbf{C}^H\mathbf{C}\right)^{-1}\mathbf{C}^H$ and $\mathbf{P}_{\mathbf{C}}^{\perp} \triangleq \mathbf{I} - \mathbf{C}\left(\mathbf{C}^H\mathbf{C}\right)^{-1}\mathbf{C}^H$ are projection matrices for the two subspaces. The rank of $\mathbf{P}_{\mathbf{C}}$ is the number of constraints used. This derivation leads to the generalized sidelobe canceller (GSC) beamformer [101].

6.6 Robust beamforming

In practical situations, the performance of the MVDR beamformer is hard to achieve. As outlined in Section 5.2, there are several mechanisms that prevent the expectation of the array data matching the statistics of the actual array data. Many of these effects introduce an error into the steering vector of each source. Much of the robust beamforming literature considers an additive disturbance to the steering vector. Whilst we consider the model introduced in Chapter 5 in the subsequent technical chapters, for the following discussion of the literature we denote the actual steering vector, $\mathbf{d}$, as the expected steering vector, $\bar{\mathbf{d}}$, perturbed by an error vector, $\tilde{\mathbf{d}}$:

$$\mathbf{d} \triangleq \bar{\mathbf{d}} + \tilde{\mathbf{d}}. \tag{6.13}$$

Robust beamformers are designed to limit the fall in performance when errors are present in the steering vectors. The ideal MVDR beamformer, where all steering vectors are as expected, sets an upper limit on the performance that robust beamformers can achieve. There are various existing methods in the literature to create robust beamformers.

In the case of the LCMV beamformer 6.12, linear constraints can be added to the neighbourhood of the expected steering vectors to increase the width of the main beam, which reduces the chances that the desired source will be suppressed, [85]. Derivative constraints are used on the beamformer gain with respect to the expected steering vector angle in order to force the gain to be flat with angle, this prevents the main beam width from collapsing sharply, [87].

Eigenspace beamformers, such as the SNR optimal beamformer (6.10), can be powerful in many scenarios when the SNR is positive, [90, 102]. However when the Signal-to-Noise ratio (SNR) is poor the performance suffers dramatically due to subspace swaps, [103, 104], in which the largest eigenvalue may not correspond to the desired source. In order to design eigenspace beamformers, the dimension of the signal-plus-interference subspace needs to be low and known. This can cause

performance degradations in situations with source scattering and reverberant environments, [62].

Adaptive beamforming Conventional beamforming weight design can be divided into either optimum (data-independent) or adaptive designs. Adaptive beamformers are deployed when the acoustic environment is changing with time. They are based on optimum beamformer designs [95] but are more computationally expensive and can suffer from performance losses when the environment does not match the expectation, or the desired signal is indistinguishable from the interference or noise. Many robust beamformers are either designed as adaptive or can be reformulated in an adaptive manner. Indeed, similarly to robust beamformers, many adaptive beamformers will aim to match the performance of the ideal MVDR beamformer when no steering vector errors are present.

Insufficient training data [59] and low SNR [90] amongst other issues can cause the adaptive beamformers to confuse interferences with the desired source; and cancel the desired source, which is known as self-nulling, [105]. Optimum beamformers are preferred in scenarios where the position of speech sources are known or unlikely to move, such as hearing aids, teleconferencing or in-car communication, [106]. The beamformers in this thesis concentrate on the area of optimum beamforming.

6.6.1 Diagonal loading

Diagonal loading is a form of regularisation which helps to form robust beamformers [89]. It follows the same method as the MVDR beamformer, but the objective function includes a loading parameter, ξ, which scales Euclidean norm of the weight vector and add its to the cost function:

$$\min_{\mathbf{w}} \mathbf{w}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle \mathbf{w} + \xi \left\| \mathbf{w} \right\|^2 \qquad \text{subject to } \mathbf{w}^H \bar{\mathbf{d}}_1 = 1, \tag{6.14}$$

where $\|\cdot\|$ is the Euclidian norm. All the eigenvalues of the inverse covariance matrix are raised by the loading parameter, which in turn reduces its condition number. The solution of the above is given as [95]:

$$\mathbf{w}_{LMVDR} = \left(\bar{\mathbf{d}}_1^H \left(\left\langle \mathbf{x}\mathbf{x}^H \right\rangle + \xi \mathbf{I}\right)^{-1} \bar{\mathbf{d}}_1\right)^{-1} \left(\left\langle \mathbf{x}\mathbf{x}^H \right\rangle + \xi \mathbf{I}\right)^{-1} \bar{\mathbf{d}}_1. \tag{6.15}$$

The loading parameter is added to the diagonal of the covariance matrix. The diagonal loading reduces the singularity of the problem and limits narrow peaks and troughs around each source. Hence the response against changes in the steering vector is smoother, and more robust to errors in the steering vector. The same regularisation can be applied to other beamformers such as the LCMV and SNR optimal.

The loading parameter is not generally specified in a deterministic way and is found empirically [91]. Choosing the loading parameter to be too large reduces the suppression of noise and interference, whilst choosing it to be too small leads to sensitivity to steering vector errors. Many robust beamforming approaches can be reformulated as a diagonally loaded beamformer, in which the loading parameter is better defined.

A commonly used value for the loading parameter is ten times the total noise power: $\xi = 10\mathrm{tr}\left(\mathbf{V}\right)$ [107]. A diagonally loaded MVDR was created for the same geometry shown in Fig. 6.3. The weights, shown in Fig. 6.4, are smoother around the interference source than either MVDR or LCMV. The response is zoomed at the location of the interference source in Fig. 6.5. The location of the interference source is not constrained to such a small region, therefore it will be more robust to changes in the steering vector. The robustness is shown in an increase in A-weighted white noise gain of 2.06 versus 1.84 for the LCMV beamformer. However, the larger and more shallow lobes mean more interference and noise power is observed in the case when no errors are present, in which case the performance is worse than that of the MVDR beamformer.

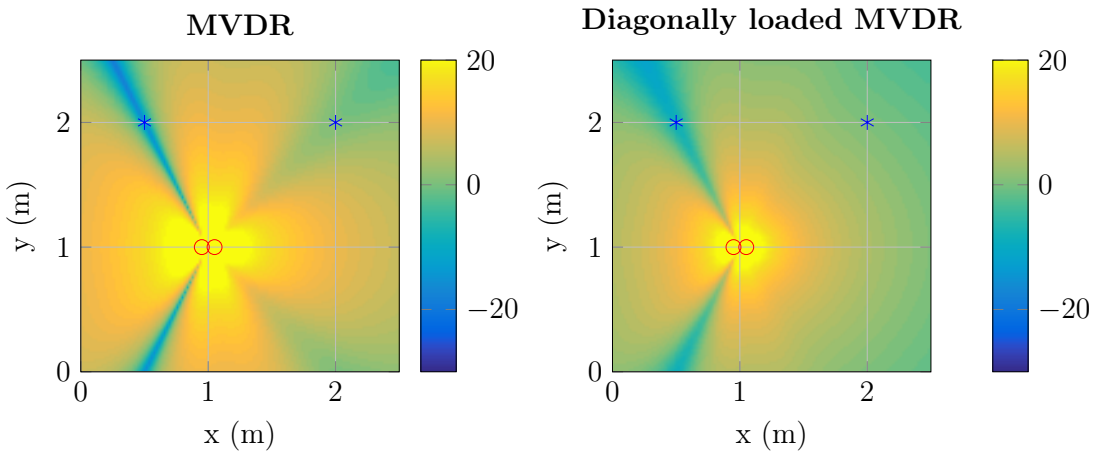

Figure 6.4: Spatial A-weighted gain of the array (dB) for the MVDR beamformer (left) compared to the diagonally loaded MVDR beamformer (right).

6.6.2 Worst-case optimisation

Beamformers which optimise the worst-case performance are designed in [108, 90]. The steering vector is assumed to be perturbed from its expected position by an vector which is bounded by its Euclidean norm:

$$\left\|\tilde{\mathbf{d}}\right\| \leq \epsilon.$$

The actual steering vector belongs to the set

$$\mathcal{A}(\epsilon) \triangleq \left\{\mathbf{d} \mid \mathbf{d} = \bar{\mathbf{d}} + \tilde{\mathbf{d}},\ \left\|\tilde{\mathbf{d}}\right\| \leq \epsilon\right\}.$$

The response of the beamformer should be greater than unity for all steering vectors from the above set. The problem is formulated as:

$$\min_{\mathbf{w}} \mathbf{w}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle \mathbf{w} \qquad \text{subject to } \left|\mathbf{w}^H \bar{\mathbf{d}}\right| \geq 1 \quad \forall\, \mathbf{d} \in \mathcal{A}(\epsilon). \tag{6.16}$$

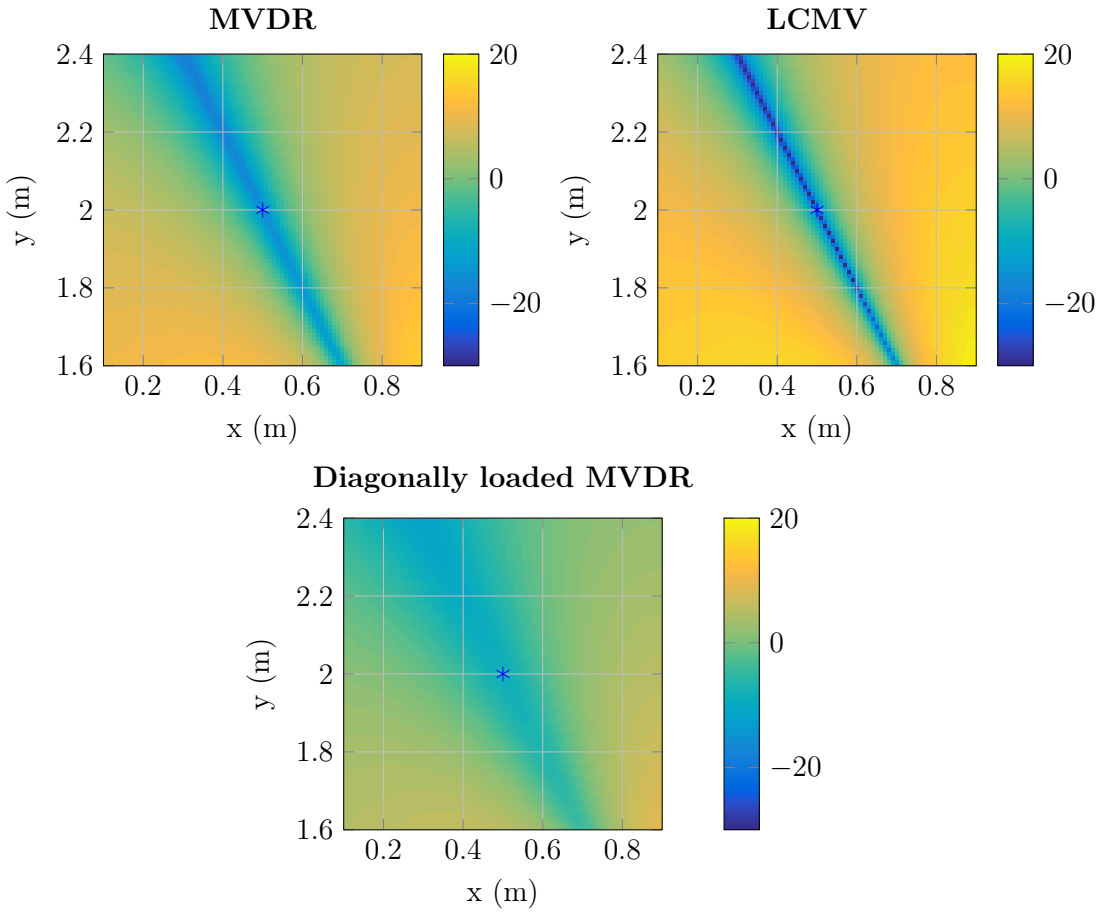

Figure 6.5: Spatial A-weighted signal gain (dB) for the MVDR (left), LCMV (right) and the diagonally loaded MVDR (bottom) beamformers, zoomed at the interference location, $(0.5, 2)$.

The resulting beamformer is conservative over the constrained region and ensures the desired source always has at least unity gain. In the worst-case scenario, when the error is at the boundary, $\left\|\tilde{\mathbf{d}}\right\| = \epsilon$, the source has distortionless response. The high gain over the set $\mathcal{A}(\epsilon)$ comes at the expense of noise and interference suppression. The error term maps to a spherical uncertainty set. The steering vectors of interference sources can also be modeled under a spherical uncertainty set. If the set of an interference source overlaps with that of the desired source, we may maintain high gain for the interference source, as thus degrade performance. The process of finding the boundary value, ϵ, needs to be defined in each application.

There are various extensions to the above model. The spherical uncertainty set is extended to a polyhedron set in [93]. It is also extended to a nondegenerate ellipsoidal uncertainty set in [91, 92]:

$$\mathcal{A}(\epsilon) \triangleq \left\{ \mathbf{d} \mid \tilde{\mathbf{d}}^H \mathbf{C}^{-1} \tilde{\mathbf{d}} \leq 1 \right\},$$

where $\mathbf{C}$ is a positive definite matrix that represents the ellipsoid. In [91] the uncertainty set is used to find an equivalent diagonal loading parameter.

In all the approaches, the steering vector is assumed from an uncertainty set. All steering vectors in the set are constrained, regardless of the distribution of steering vectors within the set. This suits well the case when the steering vectors are uniformly distributed across the set as shown in Fig. 6.6. However, if the distribution has small tails near the boundary and a large peak, such as a triangular distribution, we will constrain the tails equally to the peak. The steering vectors are unlikely at the tails and thus, by constraining high gain at the tails, we will increase the chance of amplifying noise and interference.

The uncertainty set is also truncated to a particular boundary, ϵ. If the actual distribution of steering vector errors is not truncated then we cannot ensure at least unity gain at all times for the desired source, therefore the SNR may abruptly drop when the steering vector error exceeds the boundary.

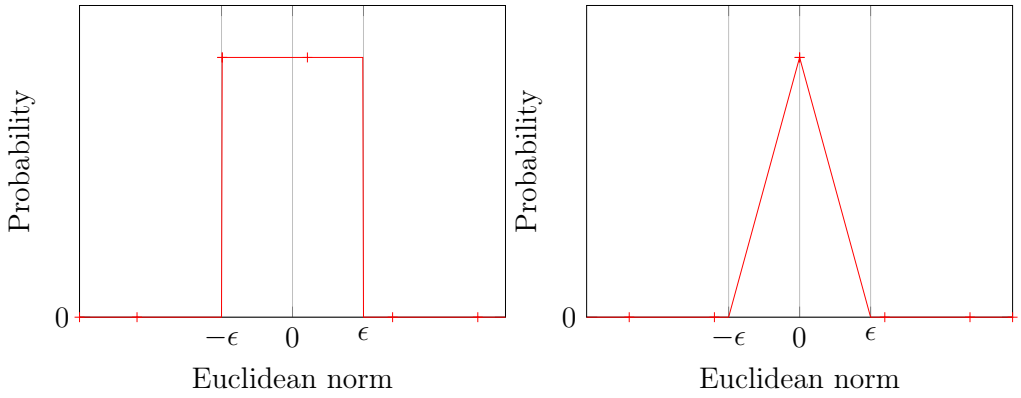

Figure 6.6: Uniform and triangular distributions for the Euclidean norm of the error vector, $\mathbf{d}_\delta$.

Doubly constrained The approach outlined by Li et al in [109] uses a doubly constrained MVDR beamformer to ensure robustness to steering vector errors. The construction is similar to that of worst-case optimisation, but an additional constraint is applied to the norm of the estimated steering vector. When applied to plane wave sources, the steering vector norm is constrained to the number of microphones: $\|\mathbf{d}\|^2 = M$. The method attempts to find $\mathbf{d}$ that maximises the power from the desired signal covariance, $\left\langle |\tilde{s}|^2 \right\rangle \mathbf{d}\mathbf{d}^H$, whilst the remaining covariance is positive semidefinite. The weights are found by using the resulting $\mathbf{d}$ in an MVDR beamformer:

$$\mathbf{w} = \frac{\left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \mathbf{d}}{\mathbf{d}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \mathbf{d}}.$$

As with worst-case optimisation, the error is bounded in a spherical set and we still have a fixed parameter, ϵ, to estimate. The work of [110] extends the model to an ellipsoidal uncertainty set.

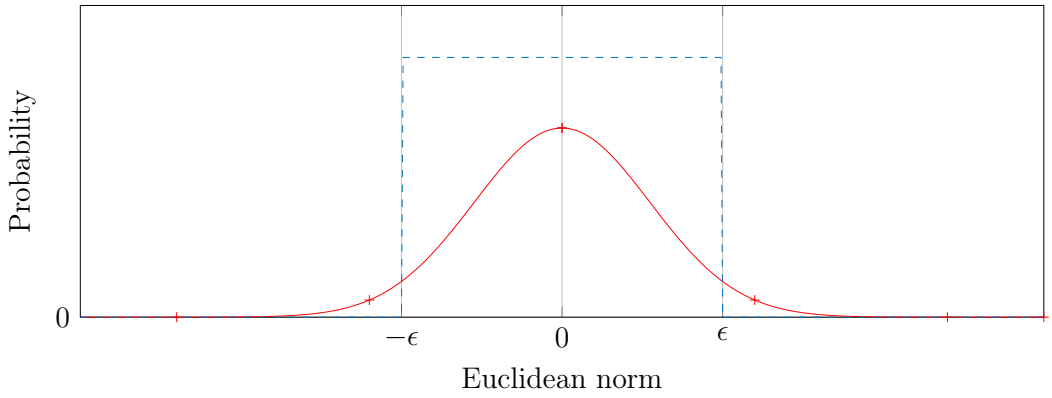

Figure 6.7: Gaussian distributions for the Euclidean norm of the error vector, $\mathbf{d}_\delta$. Only the area within the dashed line is constrained.

6.6.3 Probabilistically constrained

For the case where the distribution of steering vector errors is not bounded, the probabilistically constrained robust beamformer was formulated [94]. The authors constrain a set of steering vectors which lie within a subset of all the possible steering vectors. The array gain in the distribution tails is not constrained. The probability of the steering vector being present in the tails of the distribution is denoted the non-outage probability, p_0.

For example, $\tilde{\mathbf{d}}$ are formed from the Gaussian distribution shown in Fig. 6.7. The distribution between the bounds, $\left\|\tilde{\mathbf{d}}\right\| \leq \epsilon$, is constrained to at least unity gain, whereas the distribution outside the bounds is the non-outage probability and is not constrained.

The problem is formulated as:

$$\min_{\mathbf{w}} \mathbf{w}^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle \mathbf{w} \qquad \text{subject to } \mathrm{P}\left\{\left|\mathbf{w}^H\mathbf{d} \geq 1\right|\right\} \geq p_0. \tag{6.17}$$

The beamformer is not as conservative as the previous worst-case beamformers. The distribution which lies inside the constrained set is treated as uniform, where the area

outside is not constrained. In order to select the probability value, p_0, the distribution of $\tilde{\mathbf{d}}$ may need to be known.

6.7 Parameter estimation

In order to design optimal beamformers we require values for the steering vector and the various covariance matrices. We can either determine these from modeling, as in Chapter 5, or else adaptively.

6.7.1 Sample matrix inversion

The MVDR beamformer relies on the inversion of the array data covariance matrix, $\left\langle \mathbf{x}\mathbf{x}^H \right\rangle$. In some cases the array data covariance is not well known and is estimated from the observed array data, generally when the desired source is not active:

$$\left\langle \mathbf{x}\mathbf{x}^H \right\rangle \approx \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^H,$$

where $\mathbf{x}_n$ is the array data observed at time instance n. If the number of observed samples is small the inversion of the array data covariance matrix may not accurately reflect that of the expected array data covariance.

In the work of [59], errors in the sample matrix inversion process are modeled. It poses that the errors, which occur when the sample covariance matrix is estimated in the presence of the desired source, are similar to those of steering vector perturbations. The authors design a projection beamformer which is robust to both steering vector errors and sample matrix inversion errors. To prevent performance degradation, the expected steering vector is projected onto the signal-plus-interference subspace.

6.7.2 Steering vector estimation

Steering vector estimation methods have been proposed to avoid having to make the assumptions of the worst-case optimisation and probabilistically constrained beamformers. In the approaches of [111, 112], the authors constrain the steering vector of the desired source to an angular sector and we assume that this sector does not overlap with any interference sources.

The aim is to maximise the SNR with respect to the steering error, $\mathbf{d}_\delta$. The problem is equivalent to minimising the following:

$$\min_{\tilde{\mathbf{d}}} \left(\bar{\mathbf{d}} + \tilde{\mathbf{d}}\right)^H \left\langle \mathbf{x}\mathbf{x}^H \right\rangle^{-1} \left(\bar{\mathbf{d}} + \tilde{\mathbf{d}}\right) \quad \text{subject to} \quad \begin{aligned} \left\|\bar{\mathbf{d}} + \tilde{\mathbf{d}}\right\| &= \sqrt{M} \\ \& \quad \mathbf{P}^\perp \left(\bar{\mathbf{d}} + \tilde{\mathbf{d}}\right) &= \mathbf{0}, \end{aligned} \tag{6.18}$$

where $\mathbf{P}^\perp$ is a projection matrix to force the steering vector to stay within the angular sector subspace. The approach to find the steering vector is iterative. Once converged, the updated steering vector is used within an MVDR beamformer.

The methods do not require the norm of the error, $\tilde{\mathbf{d}}$, or its distribution, but require an angular range for the desired source.

6.8 Conclusions

We have described conventional beamformers, both robust and non-robust, in increasing requirements of information about the array. The non-robust MVDR beamformer establishes the upper performance limit of an ideal beamformer, but its performance fails in the presence of errors in the steering vectors. Robust beamformers have been designed to offer performance that does not degrade abruptly in the presence of errors. This is at the cost of performance when no errors at present. All the robust beamformers require some form of parameter that needs to be determined. The choice of the parameters creates a tradeoff between optimal performance and robustness. In speech based applications many of the parameters are not easily defined.

The beamformers designed use increasing amounts of information. The data-independent beamformers only require the location of the array elements and the sources. The common ideal beamformers use the second order statistics of the sources and the noise in order to improve performance. There are no ideal beamformers that utilise the fourth order statistics of the sources or the complex kurtosis of speech in order to further improve performance. The robust beamformers assume different amount of knowledge to determine how the steering vectors vary in practice, but not based on models of the acoustic propagation channels.

Chapter 7

Robust Beamforming in the STFT Domain

In this chapter, we design a beamformer that is robust to position uncertainties in the sources and microphones and also to the channel variations discussed in Chapter 5. It differs from the robust beamformers described in Sec. 6.6 in having design parameters that are directly related to the causes of performance variation. We derive the optimal beamformer weights that maximise the expected SNR. Initially, we consider only the effects of time delay uncertainties when deriving the optimal weights. The resulting robust beamformer is compared over random simulated geometries to other beamformers in the presence of phase uncertainties. The beamformer design method is then extended to utilise the directivity patterns of a human talker and we show that this results in further performance gains over traditional beamformers.

7.1 Beamformer weights design

In this chapter, we utilise the extended array data model, (5.7), from Chapter 5 to introduce an unknown random phase contribution to each of the propagation coefficients in order to create robust beamformers. This is given by

$$\mathbf{D} \;=\; \overline{\mathbf{D}} \odot \exp(j\omega_k \tilde{\mathbf{T}}), \tag{7.1}$$

where $\odot$ denotes element-by-element multiplication, ω_k represents the angular frequency relating to frequency index k, $\tilde{\mathbf{T}}$ represents the zero-mean Gaussian variations in the propagation path time delays and $\overline{\mathbf{D}}$ represents the propagation coefficients from the conventional deterministic model.

In the following we utilise the vector form of the array data:

$$\mathbf{x}\left(l,k\right) \;=\; \mathbf{D}\mathbf{s}\left(l,k\right) + \mathbf{v}\left(l,k\right), \tag{7.2}$$

where $\mathbf{x}, \mathbf{v} \in \mathbb{C}^{M\times 1}$, $\mathbf{s} \in \mathbb{C}^{P\times 1}$ and $\mathbf{D} \in \mathbb{C}^{M\times P}$.

Beamformer designs usually rely on the expected value of the array data covariance, $\left\langle \mathbf{x}\mathbf{x}^H \right\rangle$. Using (7.2) and (7.1) we can expand $\left\langle \mathbf{x}\mathbf{x}^H \right\rangle$ as

$$\left\langle \mathbf{x}\mathbf{x}^H \right\rangle \;=\; \left\langle \mathbf{D}\mathbf{s}\mathbf{s}^H\mathbf{D}^H \right\rangle + \left\langle \mathbf{v}\mathbf{v}^H \right\rangle \tag{7.3}$$

Assuming that all sources are independent we have:

$$\left\langle \mathbf{s}\mathbf{s}^H \right\rangle \;=\; \mathrm{diag}\left(\left\langle |s_1|^2 \right\rangle \ldots \left\langle |s_P|^2 \right\rangle\right). \tag{7.4}$$

As the source power covariance matrix is diagonal and the propagation coefficients are independent of the source power we can expand the first term in (7.3) as

$$\left\langle \mathbf{D}\mathbf{s}\mathbf{s}^H\mathbf{D}^H \right\rangle = \sum_a \left\langle |s_a|^2 \right\rangle \left\langle \mathbf{d}_a\mathbf{d}_a^H \right\rangle . \tag{7.5}$$

The propagation coefficient covariance matrix, $\left\langle \mathbf{d}_a\mathbf{d}_a^H \right\rangle$, is expanded in (5.10). Thus the first term is simplified as

$$\left\langle \mathbf{D}\mathbf{s}\mathbf{s}^H\mathbf{D}^H \right\rangle_{\epsilon,\varphi} = \sum_{a=1}^{P} \left\langle |s_a|^2 \right\rangle \bar{d}_{\epsilon a}\bar{d}^*_{\varphi a} \exp\left(-\frac{\omega_k^2}{2}\left(\left\langle \tilde{t}^2_{\epsilon a} \right\rangle + \left\langle \tilde{t}^2_{\varphi a} \right\rangle - 2\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a} \right\rangle \right) \right) . \tag{7.6}$$

Each term in the above expression is known or can be approximated. The propagation uncertainty covariances, $\left\langle \tilde{t}_{\epsilon a}\tilde{t}_{\varphi a} \right\rangle$, depend on the paths from source a to microphones ϵ and φ and are derived in Chapter 5.

7.2 Simulations

Applying (7.3) to a distributed beamformer shows how the weights are changed when we consider the different types of propagation uncertainties. To validate the performance of the proposed robust beamformer we first use an illustrative example consisting of one source and four microphones considering only propagation time uncertainties. We then consider the average results over many different source and microphone geometries with up to 50 microphones. Finally, we consider the performance over the same geometries but in the presence of position uncertainties in addition to the path delay uncertainties.

The following results were generated using the parameters $\sigma = 0.2$ and $\kappa^2 = 1.7 \times 10^{-8}$. We consider two types of noise, $\mathbf{v} = \mathbf{v}_\eta + \mathbf{v}_\kappa$, where $\mathbf{v}_\eta$ is sensor noise and $\mathbf{v}_\kappa$ is spatially diffuse noise. White Gaussian sensor noise, $\mathbf{v}_\eta$, is added to each microphone in the time domain at -90 dB, relative to the desired source power,

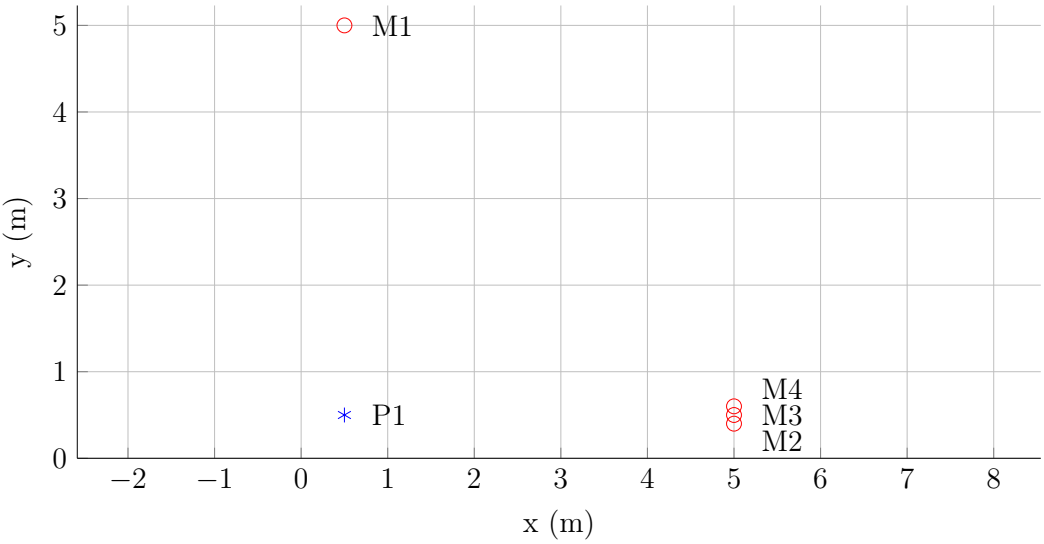

Figure 7.1: Microphone and source locations within a room

before being transformed into the STFT domain. Diffuse noise is generated at each microphone in the STFT domain using the following covariance matrix, [113, 114]:

$$\langle \mathbf{v}_\kappa \mathbf{v}_\kappa^* \rangle_{(\epsilon,\varphi)} = \phi_\kappa \frac{\sin\left(\frac{\omega_k}{c} \|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{m}}_\varphi\|\right)}{\frac{\omega_k}{c} \|\bar{\mathbf{m}}_\epsilon - \bar{\mathbf{m}}_\varphi\|}, \tag{7.7}$$

where c is the expected propagation speed and $\phi_\kappa = -80\,\text{dB}$ is the expected power of the diffuse field.

7.2.1 Four element array

We consider the distributed beamformer consisting of four microphones, as illustrated in Fig. 7.1. A single source is at position (0.5 m 0.5 m). Microphone $M1$ is placed at (0.5 m 5 m). Microphones $M2$, $M3$ and $M4$, are placed, perpendicular to $M1$, approximately along the same path, at distances of 0.4 m, 0.5 m and 0.6 m from the y-axis respectively. We initially consider uncertainties due to channel speed uncertainties. The corresponding covariances from the channel uncertainty model, (5.25),

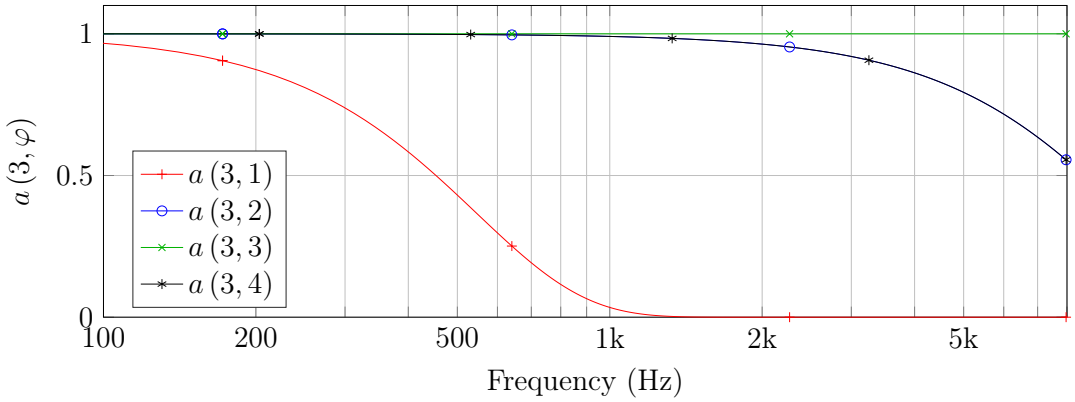

Figure 7.2: Covariance between microphone signals x_3 and x_φ in the absence of noise, $a\,(3,\varphi)$.

are given by

$$\left\langle \tilde{t}_{\epsilon 1}\tilde{t}_{\varphi 1}\right\rangle_C = 5.15\times 10^{-8}\begin{bmatrix} 1 & 0.041 & 0.041 & 0.043 \\ 0.041 & 1 & 0.981 & 0.924 \\ 0.041 & 0.981 & 1 & 0.981 \\ 0.043 & 0.924 & 0.981 & 1 \end{bmatrix}. \tag{7.8}$$

The paths from the source to microphones $M2$, $M3$ and $M4$ approximately coincide and we see from the above covariance matrix that their delays are strongly correlated. In contrast, the path from the source to $M1$ does not overlap the paths to the other microphones and we see from the low off-diagonal values in the first row and column that the delays are largely uncorrelated. Fig. 7.2 shows the correlation of microphones M_3 and M_φ in the absence of noise, as a function of frequency:

$$a\,(\epsilon,\varphi) \triangleq \exp\left(-\frac{\omega_k^2}{2}\left(\left\langle \tilde{t}_{\epsilon 1}^2\right\rangle + \left\langle \tilde{t}_{\varphi 1}^2\right\rangle - 2\left\langle \tilde{t}_{\epsilon 1}\tilde{t}_{\varphi 1}\right\rangle\right)\right).$$

The bottom line represents microphone $M1$, its covariance decreases with frequency faster than that of $M2$ and $M4$, which are represented by the middle lines.

This shows that the path delay deviations from $P1$ are similar across $M2$, $M3$ and $M4$. Thus a beamformer that uses the phase difference between microphones is still able to rely on $M2$, $M3$ and $M4$, but not on $M1$. As the phase errors are proportional to frequency this distinction becomes more important at higher frequencies. At the highest frequencies, the paths to $M2$ and $M4$ are not well correlated with that of $M3$, in which case we should no longer combine any microphones.

7.2.1.1 Beamformer weights

Utilising the change in covariance we can design beamformer weights that are robust to these channel deviations in a given array geometry.

To illustrate these effects, we compare two alternative designs of the SNR optimal beamformer from Sec. 6.4. The optimal weights are the entries of the eigenvector corresponding to the maximal eigenvalue, λ_{max}, of the matrix $\mathbf{B}$ [95]:

$$\begin{aligned} \mathbf{B} &= \mathbf{V}^{-1} \left\langle |\tilde{s}|^2 \right\rangle \left\langle \mathbf{d}_1 \mathbf{d}_1^H \right\rangle \\ \mathbf{B}\mathbf{w} &= \lambda_{max} \mathbf{w}. \end{aligned} \tag{7.9}$$

The weights are constrained so that there is unity gain for the desired source, $\mathbf{w}^H \bar{\mathbf{d}}_1 = 1$.

First, the original weights, when the propagation coefficients are taken from the conventional deterministic model,

$$\mathbf{D} \triangleq \overline{\mathbf{D}}, \tag{7.10}$$

and second, the proposed weights, when we include both channel speeds and positional uncertainties from (7.1):

$$\mathbf{D} \triangleq \overline{\mathbf{D}} \odot \exp(j\omega_k \tilde{\mathbf{T}}). \tag{7.11}$$

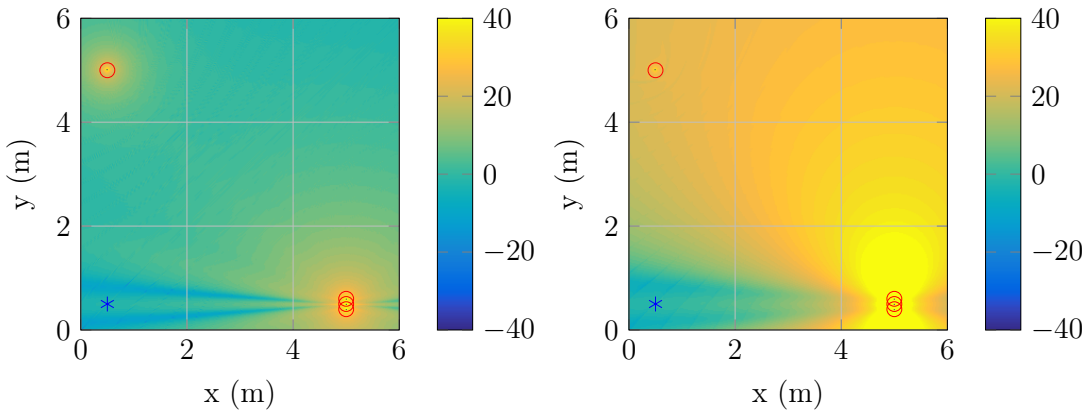

Figure 7.3: Spatial A-weighted gain of the array (dB) for the traditional SNR beamformer, (7.10), (left) and the robust beamformer (right), (7.11).

To compare the two approaches we compute the SNR of each beamformer:

$$\text{SNR} = \frac{\left\langle |\tilde{s}|^2 \right\rangle \mathbf{w}^H \left\langle \mathbf{d}_1 \mathbf{d}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}},$$

where $\mathbf{D} \triangleq \overline{\mathbf{D}} \odot \exp(j\omega_k \tilde{\mathbf{T}})$. The expected speech power, $\left\langle |\tilde{s}|^2 \right\rangle$, is taken from the long-term average speech spectra (LTASS) [20] as described in Chapter (3).

Applying a traditional, (7.10), and robust beamformer, (7.11), to the example shown in Sec. 7.2.1 demonstrates the robustness of the proposed beamformer. The spatial response of each beamformer was computed. Figure 7.3 shows the resulting A-weighted gain of the array, $\mathbf{w}^H \left\langle \mathbf{d}\mathbf{d}^H \right\rangle \mathbf{w}$, for each position, $\mathbf{d}$, in the plane of the source. In the left plot, the traditional beamformer, (7.10), shows an increase in gain in the region surrounding microphone $M1$, which indicates that there is significant gain in $M1$. The proposed beamformer, (7.11), shown in the right plot, does not have a large gain in the region around microphone $M1$. The signal from this microphone has been given a low weight because the propagation path uncertainties mean that its phase relative to the other microphone signals is very variable.

In the zoomed response of the traditional beamformer, shown in the left plot of Fig. 7.4, we can see that there is large variation in the gain with respect to position

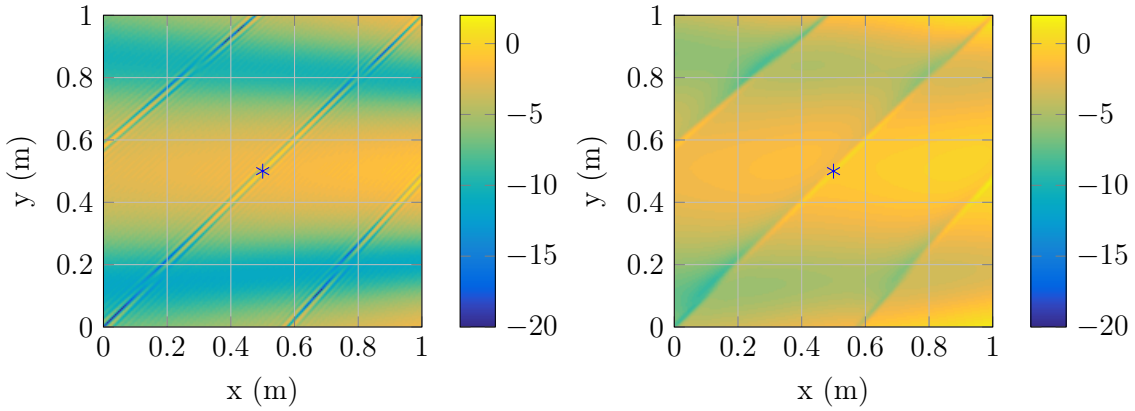

Figure 7.4: Spatial A-weighted gain of the array (dB) for the traditional SNR beamformer, (7.10), (left) and the robust beamformer (right), (7.11), zoomed in at the source position.

around the source location. If the source moves by 1 cm, the response changes by up to 15 dB. The robust beamformer, (7.11), however, has a much smoother response, as such a similar change in position gives rise to a change of around 5 dB. Therefore the proposed beamformer is more robust to time based uncertainties.

7.2.2 Performance over random geometries

Fig. 7.5 shows the improvement in SNR comparing beamformers designed with (7.10) and (7.11) across 300 random array geometries, with a single source and up to 50 microphones. The positions of the source and microphones were chosen independently from a uniform distribution. The mean gain in SNR of the robust beamformer weights against the conventional weights increases with frequency. The more microphones present, the more beneficial the robust weights. As frequency increases, the effect of time delay variations on phase become larger. Thus the performance of the conventional method, (7.10), starts to degrade. However the robust weights design, (7.11), still performs well, thus the SNR improvement increases with frequency. Also the robust beamformer always provides a positive average SNR improvement, even when uncertainties are not an issue, such as at low frequencies. The performance

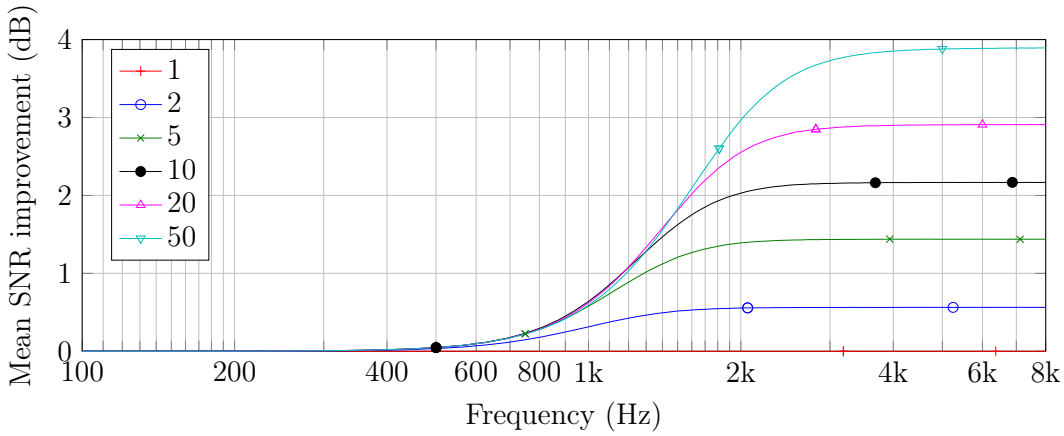

Figure 7.5: The mean expected SNR gain of the robust beamformer, (7.11), including channel deviations, compared to the conventional beamformer, (7.10), against frequency, for different numbers of microphones.

benefits taper off at high frequency. When the uncertainties are large enough the beamformer only uses the microphone with the best SNR, and additional path delay uncertainties do not cause further degradation. The benefit of the robust beamformer design increases with the number of microphones in the array; with a single microphone there is no benefit at all whereas with 50 microphones the approach gives an SNR improvement of up to 4 dB.

Position uncertainties In addition to the effects of path delay uncertainty, we now consider uncertainties in the the positions of the sources and microphones. The position errors are independently chosen from Gaussian distributions using parameters $\sigma_m = 0.05\,\text{m}$ and $\sigma_p = 0.05\,\text{m}$, and the contribution to $\tilde{\mathbf{T}}$ is defined in (5.17). Fig. 7.6 shows the results from including source and microphone position errors. We see further gains in the performance over the conventional beamformer, particularly at frequencies below 2 kHz. The position errors cause an additional phase difference that increases with frequency. The performance of the proposed robust weights, (7.11), are not affected as much. Thus the SNR gains are further improved with frequency.

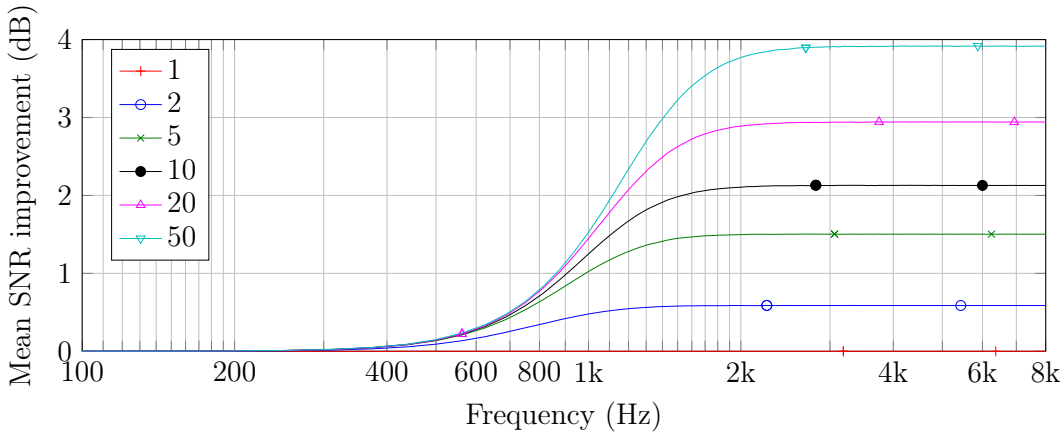

Figure 7.6: The mean expected SNR gain of the robust beamformer, (7.11), including channel deviations and position errors, compared to the conventional beamformer, (7.10), against frequency, for different numbers of microphones.

7.3 Amplitude uncertainties

We can extend the robust time domain beamformer to the amplitude uncertainties arising from the rotation of the talker's head that were discussed in Sec. 5.5. This will improve robustness to a random head rotation for each source. In order to utilise the head rotations model we incorporate the amplitude uncertainties matrix into the array data covariance:

$$\mathbf{D} \triangleq \overline{\mathbf{D}} \odot \tilde{\mathbf{H}} \odot \exp\left(j\omega_k \tilde{\mathbf{T}}\right). \tag{7.12}$$

The contribution of the sources to the array data covariance matrix, $\left\langle \mathbf{x}\mathbf{x}^H \right\rangle$, can be expanded using (5.31) as:

$$\left\langle \mathbf{D}\mathbf{s}\mathbf{s}^H\mathbf{D}^H \right\rangle_{\epsilon,\varphi} = \sum_{a=1}^{P} \left\langle |s_a|^2 \right\rangle \bar{d}_{\epsilon a} \bar{d}^*_{\varphi a} \left\langle \tilde{h}_{\epsilon a} \tilde{h}_{\varphi a} \right\rangle \exp\left(-\frac{\omega_k^2}{2} \left(\left\langle \tilde{t}^2_{\epsilon a} \right\rangle + \left\langle \tilde{t}^2_{\varphi a} \right\rangle - 2 \left\langle \tilde{t}_{\epsilon a} \tilde{t}_{\varphi a} \right\rangle \right) \right).$$

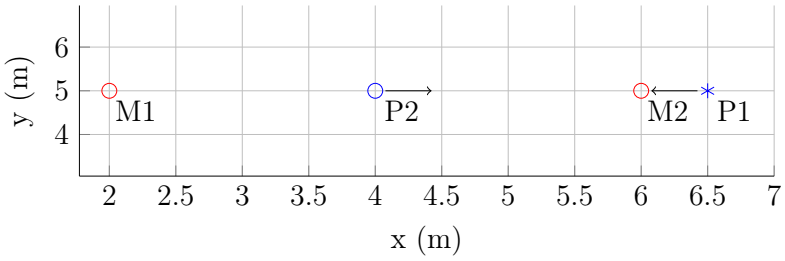

Figure 7.7: The sources are directed on average towards each other.

The head rotations which make up the amplitude uncertainties, $\left\langle \tilde{h}_{\epsilon a}\tilde{h}_{\varphi a}\right\rangle$, are formed from a known uniform distribution and are computed following the derivation in Sec. 5.5.2.2.

7.3.1 Illustrative example

Consider the array geometry consisting of two sources ($P2$ is an interferer) and two microphones, shown in Fig. 7.7. The sources independently rotate with an orientation uniformly distributed in the range $\pm 45°$and on average are both facing towards $M2$. The amplitude uncertainty covariances, $\left\langle \tilde{h}_{\epsilon a}\tilde{h}_{\varphi a}\right\rangle$, for each source are shown in Fig. 7.8. Since both microphones are in the same direction relative to source P1, they will experience the same amplitude variations as the source rotates, as illustrated in the upper plot. However, since the microphones are on opposite sides of the interferer, $P2$, the channel directions are different and the amplitude uncertainties will not be well correlated, as shown in the lower plot. The microphone, $M1$, behind $P2$ has a larger attenuation compared to $M2$.

The magnitudes of the weights resulting from a traditional SNR beamformer, (7.10), and an amplitude uncertainties robust beamformer, (7.12), are shown in Fig. 7.9. The traditional SNR beamformer, shown in the left plot, attempts to remove the contribution of the interferer, $P2$, by subtracting $M1$ from $M2$ and so that the weights of the two microphone signals have opposite signs. As the distance from source $P2$ to each microphone is the same, the magnitude of each weight is the same. However when the effect of the head rotation model is considered the interferer is no

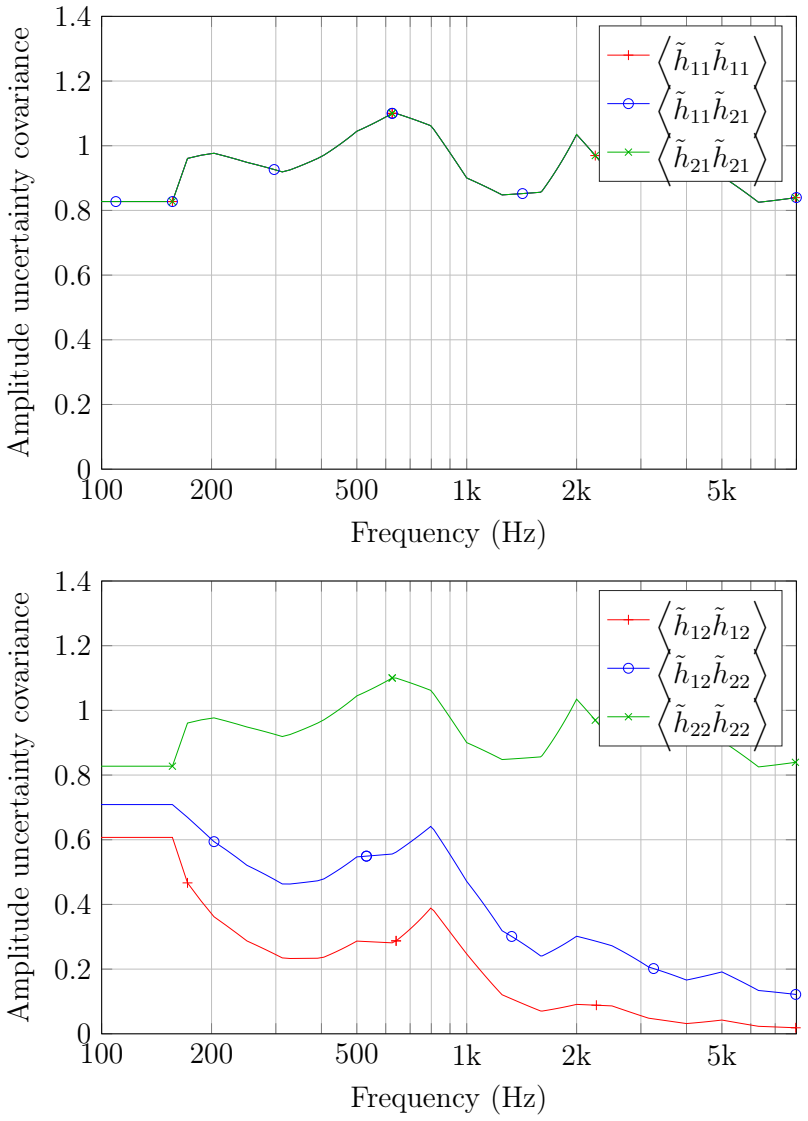

Figure 7.8: The covariance in amplitude uncertainties across different channels for each source, $P1$ - top, $P2$ - bottom.

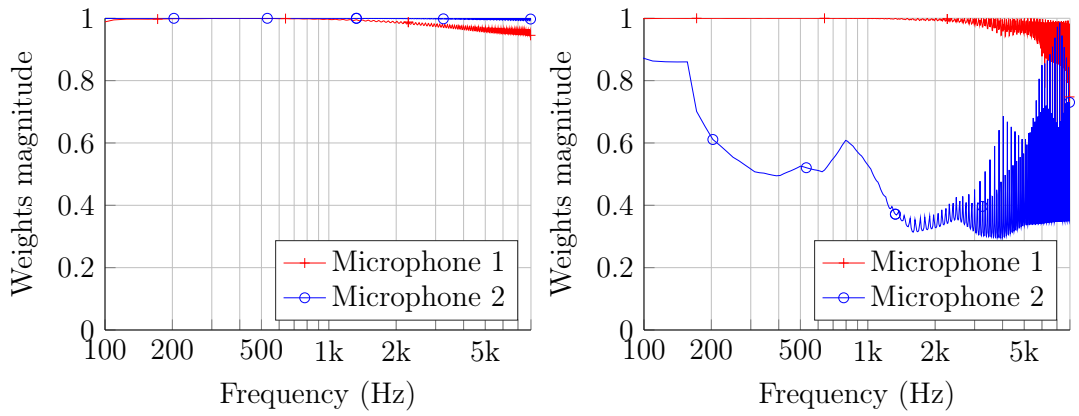

Figure 7.9: The weights magnitude for a traditional, (7.10), and robust beamformer, (7.12).

longer matched across the two microphones. The amplitude at $M1$ is much less than that of $M2$, therefore it is not fully cancelled. The resulting expected A-weighted SNR for the traditional beamformer is 14.7 dB. Generally, beamformers which attempt to create spatial zeros for interference will suffer larger performance degradations.

The robust beamformer, (7.12), attenuates the weight for $M2$ in order to rescale the expected amplitude of the interferer to match $M1$. As such the cancellation of the interferer is more successful and gives an expected A-weighted SNR of 22.9 dB. At higher frequencies the contribution of the interference power from $P2$ that is present at $M1$ is small enough to be similar in power to the noise contributions, from diffuse and sensor noise. Therefore we no longer wish to scale $M2$ to match the power contributions of the interference as the noise will reduce the SNR. The weights magnitude from 2 kHz to 8 kHz varies rapidly. This is due to the diffuse noise, which is a function of the position of the two microphones, as shown in (7.7). The correlation in diffuse noise power varies based on the relationship between the wavelength and the microphone separation, which causes the periodic oscillation seen at high frequencies.

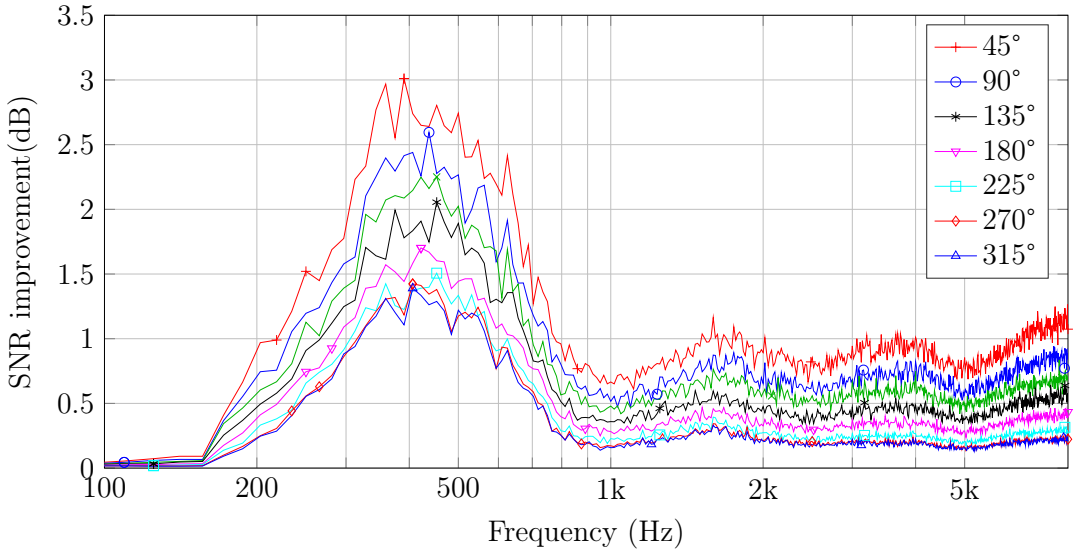

Figure 7.10: The median expected SNR gain of the amplitude uncertainties robust beamformer, (7.12), compared to the conventional beamformer, (7.10), against frequency, for different head rotation ranges.

7.3.2 Performance over random geometries

The rotation range of each source will change the amplitude uncertainties across each channel, as seen in Fig. 5.10. Thus the resulting beamformer weights and SNR performance will vary based on the rotation range. The gain in SNR of the amplitude uncertainties robust beamformer, (7.12), over the traditional beamformer, (7.10), was calculated over 100 random array geometries. The 100 geometries were chosen to cover a range of SNRs between -15 and 15 dB for the best case microphone. In each scenario there were 10 microphones and 4 sources. Each source rotates randomly and uniformly within a specified range, and is directed on average to a random direction. The median SNR gain was taken over all geometries and the results are shown for each rotation range in Fig. 7.10. As the rotation range increases, the performance gains decrease. A smaller rotation range creates a larger difference in expected amplitude between different channel directions, which increases the performance degradation of the traditional beamformer. As previously noted, the performance gains are limited

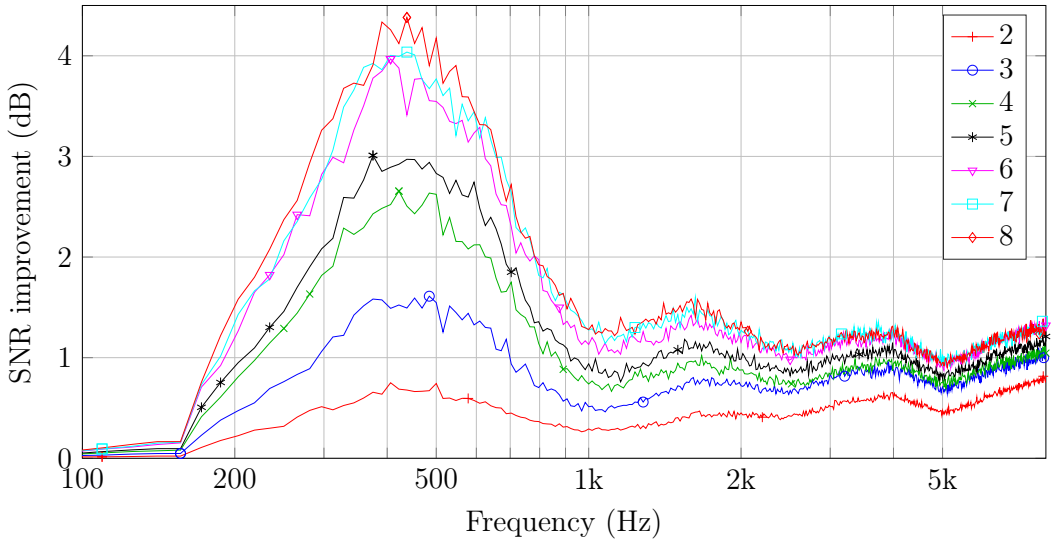

Figure 7.11: The median expected SNR gain of the amplitude uncertainties robust beamformer, (7.12), compared to the conventional beamformer, (7.10), against frequency, for different numbers of sources.

at higher frequencies where the directivity means the low received source power is comparable to that of the noise.

To assess the effect of varying the number of interfering sources, the average performance of the amplitude uncertainties robust beamformer, (7.12), was considered for different number of sources over 300 random array geometry simulations. The setup consisted of 10 microphones and different number of sources. Each source has a random expected direction and a rotation range of $\pm 22.5°$. The 300 geometries were chosen to cover a range of SNRs between -15 and 15 dB for the best case microphone. The performance improvement in dB of the robust beamformer over the traditional beamformer was taken and averaged over all rooms. The results are shown in Fig. 7.11. The more sources present, the greater the SNR improvement of the robust beamformer over the traditional beamformer.

The median A-weighted results over all frequencies and all 300 geometries relative to the traditional SNR beamformer are shown in Fig. 7.12. As we introduce more sources, the performance gains over the traditional beamformer increase. The more

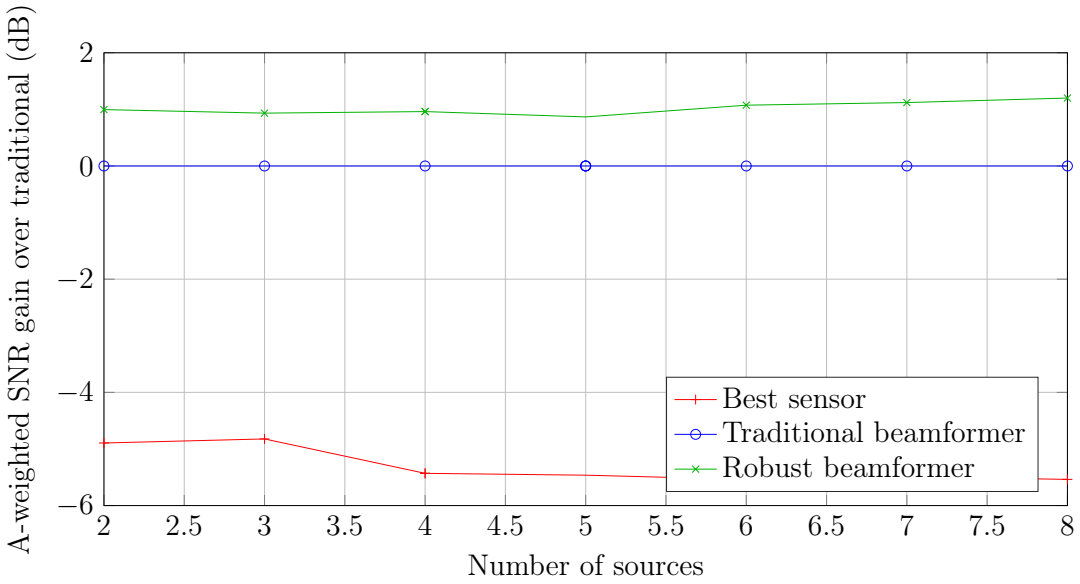

Figure 7.12: The median A-weighted expected SNR gain of the amplitude uncertainties robust beamformer, (7.12), compared to the conventional beamformer, (7.10), against frequency and the best microphone, for different numbers of sources.

sources present, the more errors the traditional beamformers makes when attempting to cancel the signal from interfering source.

7.4 Conclusions

In this chapter we have incorporated the propagation phase uncertainties model from Chapter 5 into the design of a SNR optimal beamformer. We have demonstrated that it is robust to the phase uncertainties and always results in performance improvements over the traditional beamformer, especially at high frequencies where the resultant phase uncertainties are greatest. We have extended the robust beamformer to incorporate human head directivity patterns to further improve its robustness and performance gains over the traditional beamformer.

Chapter 8

Robust Beamforming in the Power Domain

We proposed a novel new beamformer approach which is based in both the linear and power domains. The non-linear power domain processing of the microphone signals gives a benefit to beamformer performance that cannot be exploited through linear processing. Through simulation and experimental data the power beamformer is compared with optimal amplitude domain beamformers. A statistical model is used for the channel propagation and array geometry in order to create a robust beamformer.

8.1 Introduction

In the last chapter, we described an SNR-optimal beamformer in the presence of uncertainties in the propagation coefficients. The propagation channel and the element positions are modeled as probabilistic functions by realistic distributions. A well-defined closed-form beamformer was created which is robust to both.

Many speech enhancement algorithms exist that manipulate speech signals in a non-linear fashion, such as spectral subtraction [25, 26, 23], neural networks [115] and particle filtering [116]. Beamforming, however, has been primarily limited to linear processing. In this paper, we propose a novel two stage beamformer which separates the processing of the phase and magnitude of the output signal. As we will show, this non-linear approach allows for greater performance gains. The intelligibility of human hearing is more sensitive to amplitude variations than phase changes [117], therefore the weight design will concentrate on the derivation of optimal amplitude weights.

The chapter will introduce the new framework of our power domain beamformer, following this the optimal weights will be derived. Then, we will then present comprehensive simulation results and draw conclusions.

8.2 Two stage beamformer

In this section we proposed a novel two stage beamformer. The beamformer manipulates the magnitude and phase of the separately as shown in the block diagram of Figure 8.1. The M microphone signals are first transformed into the time-frequency domain using the STFT (see Sec. 2.1.1), each frequency bin is processed independently in the remainder of the block diagram. In each frequency bin, $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is a vector of complex STFT coefficients.

Linear beamforming stage The first stage is a set of N linear beamformers in the STFT domain. The output of this first stage is $\mathbf{y} \in \mathbb{C}^{N\times 1}$, where each element represents the output of a different beamformer. The weights used to obtain the first stage beamformers will be denoted $\mathbf{W} \in \mathbb{C}^{M\times N}$. The output from the first stage is formulated as:

$$\mathbf{y} \;=\; \mathbf{W}^H\mathbf{x}. \tag{8.1}$$

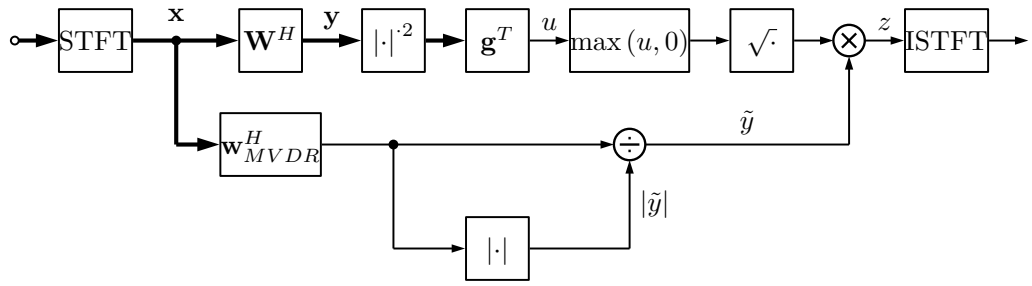

Figure 8.1: Block diagram of two stage beamformer, a thick line indicates a vector, a thin line indicates a scalar.

Power beamformer stage The second stage beamformer acts in the power domain to combine the outputs of the first stage beamformers. The first stage beamformer outputs are converted into the power domain, before being combined with another set of weights, $\mathbf{g}$, to give the power domain beamformer output, u:

$$u = \mathbf{g}^T |\mathbf{y}|^{\cdot 2}, \tag{8.2}$$

where $|\cdot|^{\cdot 2}$ represents the element by element magnitude squared. Without loss of generality we can take the elements of $\mathbf{g}$ to have absolute value of 1 by scaling the corresponding columns of W appropriately.

Substituting (8.1) into (8.2) gives

$$\begin{aligned} u &= \mathbf{g}^T \mathrm{diag}\left(\mathbf{W}^H \mathbf{x}\mathbf{x}^H \mathbf{W}\right) \\ &= \mathrm{tr}\left(\mathbf{G}\mathbf{W}^H \mathbf{x}\mathbf{x}^H \mathbf{W}\right), \text{ where } \mathbf{G} = \begin{bmatrix} g_1 & & 0 \\ & \ddots & \\ 0 & & g_N \end{bmatrix} \\ &= \mathrm{tr}\left(\mathbf{W}\mathbf{G}\mathbf{W}^H \mathbf{x}\mathbf{x}^H\right), \end{aligned} \tag{8.3}$$

where $\mathrm{diag}\left(\mathbf{A}\right) = \begin{bmatrix} a_{11} & \dots & a_{NN} \end{bmatrix}^T$ and $\mathrm{tr}\left(\cdot\right)$ is the trace operator.

We introduce the Hermitian matrix

$$\mathbf{F} = \mathbf{W}\mathbf{G}\mathbf{W}^H \in \mathbb{C}^{M\times M} \tag{8.4}$$

which represents the combined weights of the first and second stage beams:

$$u \;=\; \mathrm{tr}\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right). \tag{8.5}$$

Output signal construction As u is in the power domain, we need to construct the time domain output which requires a choice of amplitude and phase. The power domain weights, $\mathbf{g}$, are not constrained to force the power, u, to be non-negative; accordingly, as is done in spectral subtraction [23], we half-wave rectify u using the max operator.

To determine the phase of the output signal we use another linear beamformer, where the weights, $\mathbf{w}_{MVDR}$ are chosen from a MVDR beamformer, with the distortionless response for the desired source. The weights, $\mathbf{w}_{MVDR}$, are applied to the microphone signals as a normal beamformer to form a single output: $\tilde{y} = \mathbf{w}_{MVDR}^H\mathbf{x}$, of which only the phase is used.

The output of the power beamformer, z, in the amplitude domain is constructed as:

$$z = \sqrt{\max\left(u, 0\right)}\frac{\mathbf{w}_{MVDR}^H\mathbf{x}}{|\mathbf{w}_{MVDR}^H\mathbf{x}|}. \tag{8.6}$$

8.3 Optimal weights formulation

The combined weights, $\mathbf{F}$, are chosen to minimise the minimum squared error (MSE) between the power beamformer output and the power of the desired speech source:

$$\min\left[\left\langle\left(u - |\tilde{s}|^2\right)^2\right\rangle\right] = \min_{\mathbf{F}}\left[\left\langle u^2\right\rangle - 2\left\langle u\tilde{s}\tilde{s}^H\right\rangle + \left\langle|\tilde{s}|^4\right\rangle\right], \tag{8.7}$$

$$\underbrace{\begin{bmatrix} \langle x_1x_1^*x_1x_1^*\rangle & \dots & \langle x_Mx_1^*x_1x_1^*\rangle & \langle x_1x_2^*x_1x_1^*\rangle & \dots & \langle x_Mx_M^*x_1x_1^*\rangle \\ \vdots & & \vdots & \vdots & & \vdots \\ \langle x_1x_1^*x_Mx_1^*\rangle & \dots & \langle x_Mx_1^*x_Mx_1^*\rangle & \langle x_1x_2^*x_Mx_1^*\rangle & \dots & \langle x_Mx_M^*x_Mx_1^*\rangle \\ \langle x_1x_1^*x_1x_2^*\rangle & \dots & \langle x_Mx_1^*x_1x_2^*\rangle & \langle x_1x_2^*x_1x_2^*\rangle & \dots & \langle x_Mx_M^*x_1x_2^*\rangle \\ \vdots & & \vdots & \vdots & & \vdots \\ \langle x_1x_1^*x_Mx_M^*\rangle & \dots & \langle x_Mx_1^*x_Mx_M^*\rangle & \langle x_1x_2^*x_Mx_M^*\rangle & \dots & \langle x_Mx_M^*x_Mx_M^*\rangle \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} f_{1,1} \\ \vdots \\ f_{1,M} \\ f_{2,1} \\ \vdots \\ f_{M,M} \end{bmatrix}}_{\mathbf{F}} = \underbrace{\begin{bmatrix} \langle x_1x_1^*\tilde{s}\tilde{s}^H\rangle \\ \vdots \\ \langle x_Mx_1^*\tilde{s}\tilde{s}^H\rangle \\ \langle x_1x_2^*\tilde{s}\tilde{s}^H\rangle \\ \vdots \\ \langle x_Mx_M^*\tilde{s}\tilde{s}^H\rangle \end{bmatrix}}_{\mathbf{r}} \tag{8.11}$$

where $|\cdot|$ is the magnitude operator, $\tilde{s}$ is the desired source taken from $\mathbf{s}$ and $\langle\cdot\rangle$ is the expectation operator. Substituting $u = \mathrm{tr}\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right)$ from (8.5) gives

$$\begin{aligned} &\min\left[\left\langle\left(u-|\tilde{s}|^2\right)^2\right\rangle\right] = \\ &\min_{\mathbf{F}}\left[\left\langle\mathrm{tr}^2\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right)\right\rangle - 2\left\langle\mathrm{tr}\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right)\tilde{s}\tilde{s}^H\right\rangle + \left\langle|\tilde{s}|^4\right\rangle\right]. \end{aligned} \tag{8.8}$$

The optimisation is solved by setting the derivative of the MSE in (8.7) to zero:

$$\frac{d}{d\mathbf{F}}\left\langle\left(u-|\tilde{s}|^2\right)^2\right\rangle = 2\left\langle\mathrm{tr}\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right)\mathbf{x}^*\mathbf{x}^T\right\rangle - 2\left\langle\mathbf{x}^*\mathbf{x}^T\tilde{s}\tilde{s}^H\right\rangle = \mathbf{0}, \tag{8.9}$$

where $\mathbf{x}^*$ is the element by element complex conjugate of $\mathbf{x}$. Indexing the elements of the matrix equation (8.9) with $\varepsilon, \varphi \in [1, M]$ gives a set of M^2 linear equations:

$$\left\langle\mathrm{tr}\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right)x_\varepsilon x_\varphi^*\right\rangle = \sum_{\tau,\rho}\left\langle x_\varepsilon x_\varphi^* x_\rho x_\tau^*\right\rangle = f_{\tau,\rho}\left\langle x_\varepsilon^* x_\varphi \tilde{s}\tilde{s}^H\right\rangle, \tag{8.10}$$

where $f_{\tau,\rho}$ is the (τ,ρ) element of $\mathbf{F}$. This set of M^2 linear equations in the M^2 unknown elements of $\mathbf{F}$ can be written in the form $\mathbf{Q}\mathbf{F} := \mathbf{r}$, where $\mathbf{Q} \in \mathbb{C}^{M^2\times M^2}$ and $\mathbf{F}$: is the vectorization of $\mathbf{F}$, obtained by concatenating its columns $\mathbf{F} := [f_{1,1}\dots f_{M,1}\dots f_{M,M}]^T$. This is expanded in (8.11) at the top of the page. Since $\mathbf{F}$ is Hermitian, by decomposing the off-diagonal $f_{\tau,\rho}$ into real and imaginary parts, (8.10) can be converted into a set of M^2 real-valued equations. There are several

symmetries present in $\mathbf{Q}$ that can be exploited to reduce the computation needed to solve the equations.

The solution $\mathbf{F}$: $= \mathbf{Q}^{-1}\mathbf{r}$, can then be converted back into the unique matrix, $\mathbf{F}$, that solves (8.7).

8.3.1 Decomposing F

In order to implement the beamformer we need to decompose $\mathbf{F}$ into its constituent parts: $\mathbf{F} = \mathbf{WGW}^H$, where $\mathbf{G}$ must be a diagonal matrix consisting of values from the set $\{-1, 0, +1\}$.

$\mathbf{F}$ can be decomposed with eigendecomposition, where $\mathbf{B}$ are eigenvectors and $\mathbf{C}$ is a diagonal matrix of eigenvalues:

$$\mathbf{F} = \mathbf{BCB}^H. \tag{8.12}$$

As $\mathbf{F}$ is Hermitian the eigenvalues are real. In order to satisfy the constraints on $\mathbf{G}$ we rescale the eigenvalues, $\mathbf{C}$, to form our desired $\mathbf{G}$:

$$\mathbf{C} = \hat{\mathbf{C}}\mathbf{G}\hat{\mathbf{C}}^H, \tag{8.13}$$

where the component parts can be computed as:

$$\begin{aligned} \hat{\mathbf{C}} &= \sqrt{|\mathbf{C}|} \\ \mathbf{G} &= \operatorname{sign}(\mathbf{C}), \end{aligned} \tag{8.14}$$

where $\sqrt{|\mathbf{C}|}$ represents the element by element square root of $\mathbf{C}$. The magnitude of $\mathbf{C}$ is taken because $\mathbf{F}$ is not necessarily positive definite and may have some negative

eigenvalues. Substituting back into $\mathbf{F}$ gives:

$$\begin{aligned} \mathbf{F} &= \mathbf{B}\hat{\mathbf{C}}\mathbf{G}\hat{\mathbf{C}}^H\mathbf{B}^H \\ &= \left(\mathbf{B}\hat{\mathbf{C}}\right)\mathbf{G}\left(\mathbf{B}\hat{\mathbf{C}}\right)^H, \end{aligned} \tag{8.15}$$

which now matches the desired form of the beamformer weights:

$$\mathbf{F} = \mathbf{W}\mathbf{G}\mathbf{W}^H. \tag{8.16}$$

This allows us to set the first stage beamformer weights to the scaled eigenvectors:

$$\mathbf{W} = \mathbf{B}\hat{\mathbf{C}}. \tag{8.17}$$

$\mathbf{G}$ will have maximum rank M, from the number of columns of $\mathbf{W}$, therefore the number of beamformers, N, required to generate this optimum $\mathbf{F}$ is bounded by $N \leq M$.

8.4 Finding component expectations

In order to form (8.11) we needs to compute the expectations which form $\mathbf{Q}$ and $\mathbf{r}$, specifically $\left\langle x_\varepsilon x_\varphi^* x_\rho x_\tau^* \right\rangle$ and $\left\langle x_\varepsilon x_\varphi^* \tilde{s}\tilde{s}^H \right\rangle$. This section details the various expansions needed to compute these terms.

8.4.1 Quartic in x

The first term, needed to form $\mathbf{Q}$, $\left\langle x_\varepsilon x_\varphi^* x_\rho x_\tau^* \right\rangle$, can be expanded by substituting the expression for $x_\epsilon = \mathbf{d}_\epsilon^T\mathbf{s} + v_\epsilon$, from (6.1), where the row vector $\mathbf{d}_\epsilon^T = \begin{bmatrix} d_{\epsilon,1} & \dots & d_{\epsilon,P} \end{bmatrix}$. The noise and speech terms are assumed to be independent and zero-mean, $\langle s_\epsilon v \rangle = 0$,

and also have independent real and imaginary parts with equal variances; it follows that, $\langle s_\epsilon s_\epsilon \rangle = 0$, $\langle v_\epsilon v_\epsilon \rangle = 0$. If $s = u + jv$, $\langle s^2 \rangle = \langle u^2 \rangle - \langle v^2 \rangle + 2j \langle uv \rangle$ which is zero provided $\langle u^2 \rangle = \langle v^2 \rangle$. The sources are assumed to be independent, so that $\langle s_a s_b^* \rangle = 0$, $a \neq b$. With the above assumptions we can simplify the quartic expression to the following:

$$\begin{aligned}\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle = & \sum_{a,b,c,d} \left\langle d_{\epsilon,a} d_{\varphi,b}^* d_{\rho,c} d_{\tau,d}^* \right\rangle \langle s_a s_b^* s_c s_d^* \rangle \\ & + \left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle \\ & + \sum_a \langle s_a s_a^* \rangle \left(\left\langle d_{\epsilon,a} d_{\varphi,a}^* \right\rangle \langle v_\rho v_\tau^* \rangle + \left\langle d_{\epsilon,a} d_{\tau,a}^* \right\rangle \left\langle v_\varphi^* v_\rho \right\rangle \right. \\ & \qquad \left. + \left\langle d_{\varphi,a}^* d_{\rho,a} \right\rangle \langle v_\epsilon v_\tau^* \rangle + \left\langle d_{\rho,a} d_{\tau,a}^* \right\rangle \left\langle v_\epsilon v_\varphi^* \right\rangle \right).\end{aligned} \tag{8.18}$$

A full derivation is given in Appendix B.1.

Quartic speech term A similar analysis can be applied to the quartic speech term, which appears in the above so that it simplifies to:

$$\begin{aligned}\sum_{a,b,c,d} \left\langle d_{\epsilon,a} d_{\varphi,b}^* d_{\rho,c} d_{\tau,d}^* \right\rangle \langle s_a s_b^* s_c s_d^* \rangle = & \sum_{a=1}^{P} \left\langle d_{\epsilon,a} d_{\varphi,a}^* d_{\rho,a} d_{\tau,a}^* \right\rangle \left\langle |s_a|^4 \right\rangle \\ & + \sum_{a=1}^{P-1} \sum_{b=a+1}^{P} \left\langle |s_a|^2 \right\rangle \left\langle |s_b|^2 \right\rangle \\ & \left\langle d_{\epsilon,a} d_{\varphi,a}^* d_{\rho,b} d_{\tau,b}^* + d_{\epsilon,a} d_{\varphi,b}^* d_{\rho,b} d_{\tau,a}^* \right. \\ & \quad \left. + d_{\epsilon,b} d_{\varphi,a}^* d_{\rho,a} d_{\tau,b}^* + d_{\epsilon,b} d_{\varphi,b}^* d_{\rho,a} d_{\tau,a}^* \right\rangle.\end{aligned} \tag{8.19}$$

A full derivation is shown in Appendix B.1.1.

8.4.2 Quadratic in x

The second term, needed to form $\mathbf{r}$, the right side of (8.11):

$$\begin{aligned}\mathbf{r} \triangleq \left\langle x_\varepsilon x_\varphi^* \tilde{s}\tilde{s}^H \right\rangle &= \sum_p \left\langle d_{\varepsilon,p} d_{\varphi,p}^* \right\rangle \left\langle s_p s_p^* \tilde{s}\tilde{s}^* \right\rangle + \left\langle v_\varepsilon v_\varphi^* \right\rangle \left\langle \tilde{s}\tilde{s}^* \right\rangle . \\ &= \left\langle d_{\varepsilon,1} d_{\varphi,1}^* \right\rangle \left\langle |\tilde{s}|^4 \right\rangle + \left\langle v_\varepsilon v_\varphi^* \right\rangle \left\langle |\tilde{s}|^2 \right\rangle \\ &\quad + \left\langle |\tilde{s}|^2 \right\rangle \sum_{p=2}^{P} \left\langle d_{\varepsilon,p} d_{\varphi,p}^* \right\rangle \left\langle |s_p|^2 \right\rangle .\end{aligned} \tag{8.20}$$

8.4.3 Noise expectations

The noise expectations required in (8.18) and (8.20) are the quartic term, $\left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle$, and the quadratic term $\left\langle v_\epsilon v_\varphi^* \right\rangle$. For the remainder of the chapter and the simulations we model the additive noise as a combination of independent Gaussian acoustic sensor noise, $\mathbf{v}_\eta$, and diffuse noise, $\mathbf{v}_\kappa$:

$$\mathbf{v} = \mathbf{v}_\eta + \mathbf{v}_\kappa. \tag{8.21}$$

The quadratic noise term, $\left\langle v_\epsilon v_\varphi^* \right\rangle$, can be split into its component parts, as each noise contribution is independent:

$$\left\langle v_\epsilon v_\varphi^* \right\rangle = \left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* \right\rangle + \left\langle v_{\kappa,\epsilon} v_{\kappa,\varphi}^* \right\rangle , \tag{8.22}$$

where each term is taken directly from the noise covariance matrices, in which $\left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* \right\rangle = 0$, for $\epsilon \neq \varphi$.

The quartic noise covariance, $\left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle$, can also be expanded with (8.21) to produce the following result:

$$\begin{aligned}\left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle &= \left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* v_{\eta,\rho} v_{\eta,\tau}^* \right\rangle + \left\langle v_{\kappa,\epsilon} v_{\kappa,\varphi}^* v_{\kappa,\rho} v_{\kappa,\tau}^* \right\rangle \\ &\quad + \left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* \right\rangle \left\langle v_{\kappa,\rho} v_{\kappa,\tau}^* \right\rangle + \left\langle v_{\eta,\epsilon} v_{\eta,\tau}^* \right\rangle \left\langle v_{\kappa,\rho} v_{\kappa,\varphi}^* \right\rangle \\ &\quad + \left\langle v_{\kappa,\epsilon} v_{\kappa,\tau}^* \right\rangle \left\langle v_{\eta,\rho} v_{\eta,\varphi}^* \right\rangle + \left\langle v_{\kappa,\epsilon} v_{\kappa,\varphi}^* \right\rangle \left\langle v_{\eta,\rho} v_{\eta,\tau}^* \right\rangle .\end{aligned} \tag{8.23}$$

As the noise terms are multivariate Gaussians, the remaining quartic terms $\left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* v_{\eta,\rho} v_{\eta,\tau}^* \right\rangle$ and $\left\langle v_{\kappa,\epsilon} v_{\kappa,\varphi}^* v_{\kappa,\rho} v_{\kappa,\tau}^* \right\rangle$ can be expanded using a complex form of Isserlis' theorem, [118, 119]:

$$\begin{aligned}\left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* v_{\eta,\rho} v_{\eta,\tau}^* \right\rangle &= \left\langle v_{\eta,\epsilon} v_{\eta,\varphi}^* \right\rangle \left\langle v_{\eta,\rho} v_{\eta,\tau}^* \right\rangle \\ &\quad + \left\langle v_{\eta,\epsilon} v_{\eta,\tau}^* \right\rangle \left\langle v_{\eta,\varphi}^* v_{\eta,\rho} \right\rangle ,\end{aligned} \tag{8.24}$$

$$\begin{aligned}\left\langle v_{\kappa,\epsilon} v_{\kappa,\varphi}^* v_{\kappa,\rho} v_{\kappa,\tau}^* \right\rangle &= \left\langle v_{\kappa,\epsilon} v_{\kappa,\varphi}^* \right\rangle \left\langle v_{\kappa,\rho} v_{\kappa,\tau}^* \right\rangle \\ &\quad + \left\langle v_{\kappa,\epsilon} v_{\kappa,\tau}^* \right\rangle \left\langle v_{\kappa,\varphi}^* v_{\kappa,\rho} \right\rangle .\end{aligned} \tag{8.25}$$

8.4.4 Uncertainties model

In order to ensure that the beamformer is robust to the uncertainties in the propagation channels, we use the uncertainties model from Chapter 5, in which the propagation coefficients, $\mathbf{D}$, are modeled each with a probabilistic distribution:

$$\mathbf{D} \triangleq \overline{\mathbf{D}} \odot \tilde{\mathbf{H}} \odot \exp\left(j\omega_k \tilde{\mathbf{T}} \right) . \tag{8.26}$$

The uncertainty covariance matrices, $\left\langle d_{\epsilon,a} d_{\varphi,a}^* d_{\rho,b} d_{\tau,b}^* \right\rangle$ and $\left\langle d_{\varepsilon,a} d_{\varphi,a}^* \right\rangle$, that are required in (8.18) to (8.20) have been derived in Chapter 5, (5.31) and (5.42).

8.5 Simulations

The following sections describe the setup for the simulations.

LTASS assumption For the tests we assume the sources in the room are speech. In practice the expected values for the speech power, $\left\langle |s|^2 \right\rangle$, and the speech squared-power, $\left\langle |s|^4 \right\rangle$, may not be known. We assume the talkers are represented by the long-term average speech spectra (LTASS) [20]. This gives an estimate for the speech power, $\left\langle |s|^2 \right\rangle$. The speech squared-power, $\left\langle |s|^4 \right\rangle$, was found using the complex kurtosis, α_ω, described in Chapter 3, (3.2), based on the TIMIT dataset [56]:

$$\alpha_\omega = \frac{\left\langle |s_\omega|^4 \right\rangle}{\left\langle |s_\omega|^2 \right\rangle^2}. \tag{8.27}$$

STFT frame length Considering the case of a fixed weighted beamformer. The weights are not changing with time. For time domain beamformers we do not require that the sources are stationary with time. Therefore we have no constraints on the frame length of the STFT analysis window. We could apply the weights along the FFT of the whole time domain signal.

In the power domain case, we utilise the complex kurtosis from Chapter 3. The complex kurtosis relies on the statistics of speech sources and therefore assumes that the speech source power is stationary with time. The analysis window length must ensure that the speech signal can be assumed stationary within the frame. The window length will direct affect the complex kurtosis.

If the channel propagation delays are long enough, energy from the source signal can be shifted to a later analysis frame of the beamformer output. Longer frames increase the chances that the analysis window will contain frames of the source and output of the same point. The frame length chosen for the analysis of the power domain beamformer is 64 ms.

Expected MSE After computing the optimal beamformers weights, the expected MSE in each frequency band can be computed as a function of $\mathbf{F}$ as follows from (8.8):

$$\text{MSE} = \varsigma^4 \left\langle \text{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle - 2\varsigma^2 \left\langle \text{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \tilde{s}\tilde{s}^H \right\rangle + \left\langle |\tilde{s}|^4 \right\rangle, \tag{8.28}$$

where ς is a constant to scale the output signal. One can expand $\left\langle \text{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle$ to the following:

$$\left\langle \text{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle = \sum_{\varphi,\epsilon,\tau,\rho} f_{\varphi,\epsilon} f_{\tau,\rho} \left\langle x_\epsilon x_\varphi^H x_\rho x_\tau^H \right\rangle, \tag{8.29}$$

where the quartic, $\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle$, can be simplified as in (8.18).

The second term, $\left\langle \text{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \tilde{s}\tilde{s}^H \right\rangle$, is expanded as follows:

$$\left\langle \text{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \tilde{s}\tilde{s}^H \right\rangle = \text{tr} \left(\mathbf{F} \left\langle \mathbf{x}\mathbf{x}^H \tilde{s}\tilde{s}^H \right\rangle \right), \tag{8.30}$$

which is a function of $\left\langle x_\varepsilon x_\varphi^* \tilde{s}\tilde{s}^H \right\rangle$, which was expanded above in (8.20).

The MSE from (8.28) is computed for each frequency band independently. The MSE is normalised with respect to the expected signal power and then A-Weighted [18] to emphasis the importance of the ear as a receiver:

$$\text{E}_{\text{STFT}} = \sum_{\omega=1}^{N_\omega} A_\omega \left(\frac{\text{MSE}_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \right). \tag{8.31}$$

The results are presented in dB, as the MSE is in the power-squared domain, a factor of 5 is used:

$$\text{E}_{\text{STFT}}\,(\text{dB}) = 5 \log_{10} \text{E}_{\text{STFT}}, \tag{8.32}$$

where the expression for dB includes a factor of 5 because the MSE is in the power-squared domain.

The $\mathrm{E}_{\mathrm{STFT}}$ and other metrics are affected by scaling the amplitude of the beamformer output. The metric reduces if there is mismatch between the desired signal in the reference and the beamformer output. The optimal scaling factor, ς, to minimise the $\mathrm{E}_{\mathrm{STFT}}$ can be found from:

$$\varsigma^2 = \frac{\sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left\langle \mathrm{tr}\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right)_\omega \tilde{s}_\omega \tilde{s}_\omega^H \right\rangle}{\sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left\langle \mathrm{tr}^2\left(\mathbf{F}\mathbf{x}\mathbf{x}^H\right) \right\rangle_\omega}, \tag{8.33}$$

where a full derivation is shown in B.2.

Expected SNR The expected SNR from the beamformer weights can be computed as the weighted ratio of speech power to noise and interference power:

$$\mathrm{SNR}\,(\mathrm{dB}) = 10 \log_{10} \frac{\sum_{\omega=1}^{N_\omega} A_\omega \mathrm{tr}\left(\mathbf{F}\left\langle \mathbf{d}_1 \mathbf{d}_1^H \right\rangle \left\langle |\tilde{s}|^2 \right\rangle\right)_\omega}{\sum_{\omega=1}^{N_\omega} A_\omega \mathrm{tr}\left(\mathbf{F}\mathbf{V}\right)_\omega}. \tag{8.34}$$

We assume the half-wave rectification of u has no effect on the SNR. The power domain beamformer implicitly performs spectral subtraction which in some cases results in a very high expected SNR since the expected interference power is exactly cancelled.

Generating input signals In the tests which required actual speech signals, the following procedure was used. Each source, both desired and interferences, was formed by concatenating TIMIT sentences [56] from a single talker comprising approximately 9s of speech each. The STFT of each source is taken in order to form $\mathbf{s}$. The source and microphone positions were used to generate the expected propagation coefficients, $\overline{\mathbf{D}}$, per frequency per channel and also the uncertainties matrices from Sec. 8.4.4. The variance in source position is set to 10 cm and the variance in microphone position is set to 0.5 cm.

The time uncertainties covariance matrices, $\langle \tilde{t}_{\epsilon,a} \tilde{t}_{\varphi,b} \rangle$, were used to generate time uncertainties, $\tilde{\mathbf{T}} \in \mathbb{C}^{M \times P}$, for each time frame. Gaussian noise was low-pass filtered across time frames using a second order Butterworth filter whose cut-off frequency was chosen to limit the rate of change in the channel coefficients, $\tilde{t}_{\epsilon,a}$, to 0.2 ms/s, thus the channels are randomly evolving with time in a controlled manner.

We consider two types of noise, $\mathbf{v} = \mathbf{v}_\eta + \mathbf{v}_\kappa$, where $\mathbf{v}_\eta$ is sensor noise and $\mathbf{v}_\kappa$ is spatially diffuse noise. White Gaussian sensor noise, $\mathbf{v}_\eta$, is added to each microphone in the time domain at -90 dB relative to the desired source power, before being transformed into the STFT domain. Diffuse noise is generated at each microphone in the STFT domain using the following relationship, [113, 114]:

$$\langle \mathbf{v}_\kappa \mathbf{v}_\kappa^* \rangle_{(\epsilon,\varphi)} \quad = \quad \phi_\kappa \frac{\sin\left(\frac{\omega_k}{c} \left\| \bar{\mathbf{m}}_\epsilon - \bar{\mathbf{m}}_\varphi \right\|\right)}{\frac{\omega_k}{c} \left\| \bar{\mathbf{m}}_\epsilon - \bar{\mathbf{m}}_\varphi \right\|}, \tag{8.35}$$

where c is the expected propagation speed and $\phi_\kappa = -80\,\text{dB}$ is the expected power of the diffuse field relative to the desired source power.

Comparison beamformers The performance of the proposed **R**obust **P**ower **D**omain **B**eamformer (**RPDB**) is compared to five competing methods as detailed below.

BestM (Best microphone): the expected SNR at each microphone is computed and the microphone with the highest SNR is used, all other microphones are ignored.

MVDR: the MVDR beamformer weights from (6.8), [97].

SNRrob: the robust SNR beamformer from Chapter 7, [66].

Oracle: an adaptive tracking beamformer was designed which is given oracle knowledge of the true source positions and propagation delays.. It follows the low-passed true time uncertainties applied to each propagation channel. The weights are computed using an MVDR beamformer. It always tracks the desired source and interferers successfully and therefore represents an upper limit on the performance of

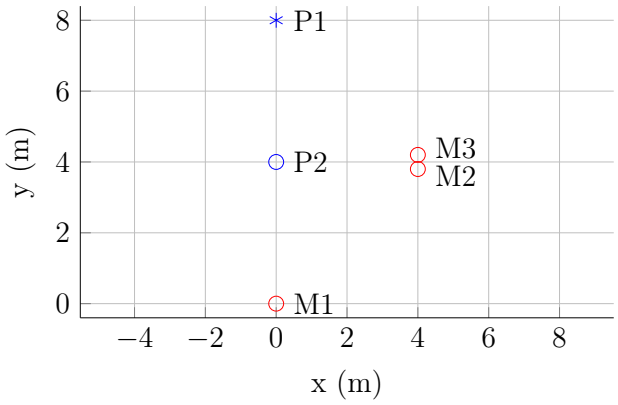

Figure 8.2: Geometry for the explained example

an adaptive tracking beamformer. In practice, low SNR situations cause the performance of tracking algorithms to fail and so a practical beamformer would perform less well that this.

8.6 Simulation results

We first consider the case of time-based uncertainties only, leading to

$$\mathbf{D} \triangleq \overline{\mathbf{D}} \odot \exp\left(j\omega_k \tilde{\mathbf{T}}\right).$$

In the subsequent section we will expand the model to include the head rotations model on each source.

8.6.1 Explained example

Consider the geometry shown in Fig. 8.2. This is an example of a distributed array containing three microphones, one desired source ($P1$) and an interference source ($P2$). Microphones $M2$ and $M3$ are separated by 40 cm.

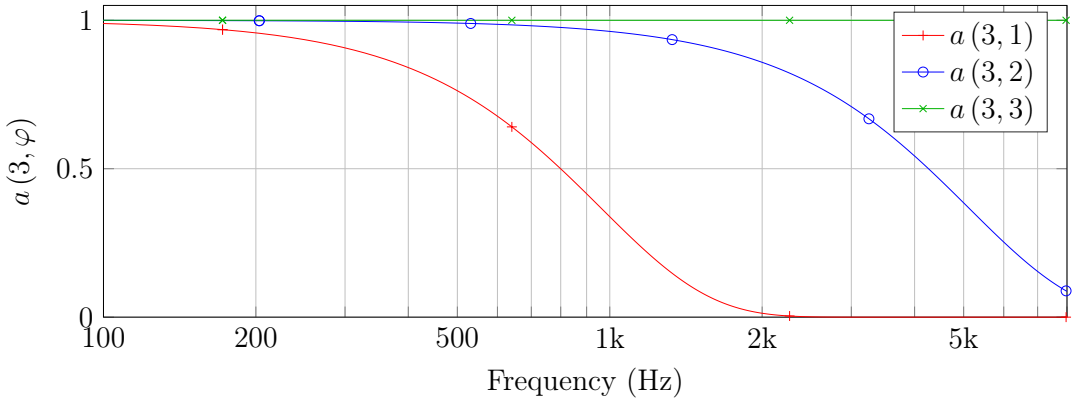

Figure 8.3: Covariance between microphone signals x_3 and x_φ in the absence of noise, $a\,(3,\varphi)$.

The correlation between channel uncertainties increases with frequency. As microphones $M2$ and $M3$ are located along a similar channel from each source the channel uncertainties are well correlated across both. However as $M1$ is located along a different channel, the uncertainties along this channel are largely uncorrelated with those along the channel to the other two microphones. Therefore the resulting phase difference along the channel to $M1$ is random relative to the phase difference of $M2$ or $M3$. The uncertainties covariance:

$$a\,(\epsilon,\varphi) \triangleq \exp\left(-\frac{\omega_k^2}{2}\left(\left\langle \tilde{t}_{\epsilon 1}^2\right\rangle + \left\langle \tilde{t}_{\varphi 1}^2\right\rangle - 2\left\langle \tilde{t}_{\epsilon 1}\tilde{t}_{\varphi 1}\right\rangle\right)\right),$$

between $M3$ and the other microphones is shown in Fig. 8.3, where it can be seen that the covariance reduces at high frequencies more rapidly for the widely spaced microphones $M3$ and $M1$. A beamformer that combines all microphones will result in performance degradation.

A traditional SNR-optimal beamformer designs weights that utilise all three microphones as shown in Fig. 8.4. The weights oscillate in accordance to the phase difference between the two close microphones; as the frequency axis is logarithmic this spacing decreases in the positive x-axis. The robust SNR beamformer from

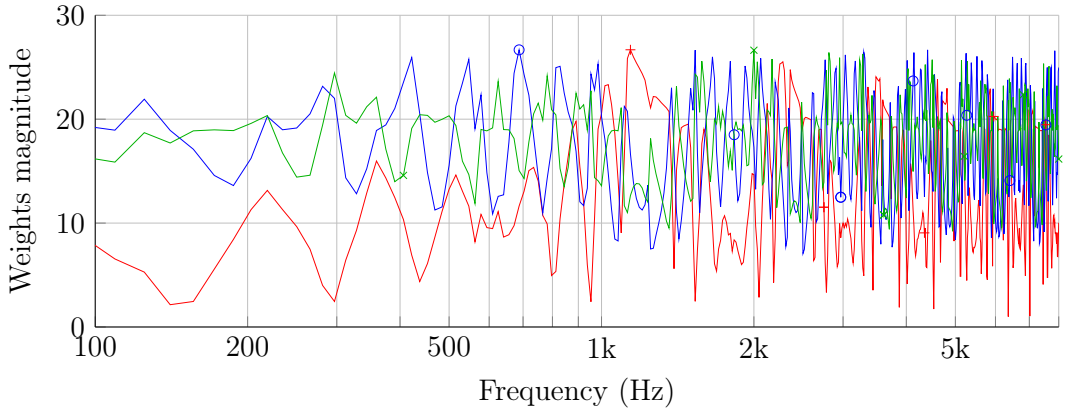

Figure 8.4: Traditional SNR-optimal beamformer

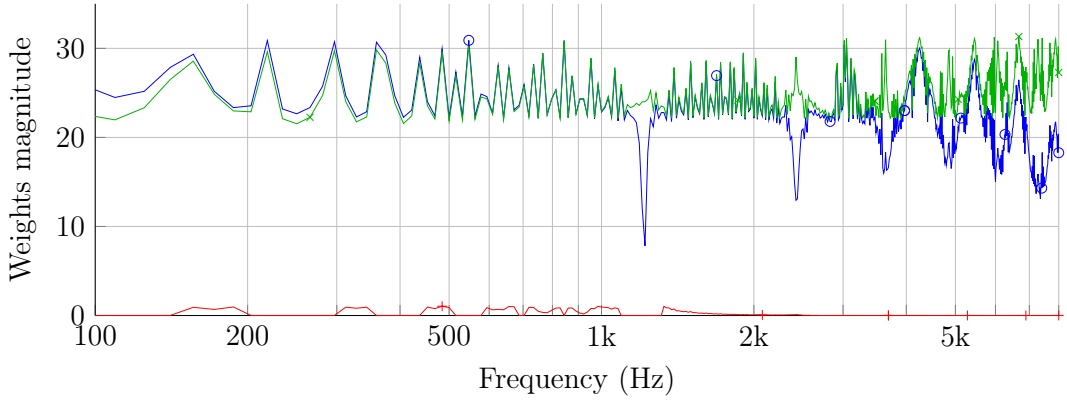

Figure 8.5: Robust SNR-optimal beamformer

Chapter 7 uses the time uncertainties correlation to reduces the emphasis on combining microphones as frequency increases. As such we combine only $M2$ and $M3$ as shown in Fig. 8.5, in which it can be seen that the weight applied to the $M1$ signal is near zero at all frequencies.

The robust power domain beamformer weights utilise spectral subtraction to remove the interfering source. It creates three separate beamformers, two that combine to select the desired source and one that focuses on the interference source. In the power domain the interference beamformer is subtracted from the others resulting in an improvement in SNR. The weights are shown in Fig. 8.6. The title reflects the

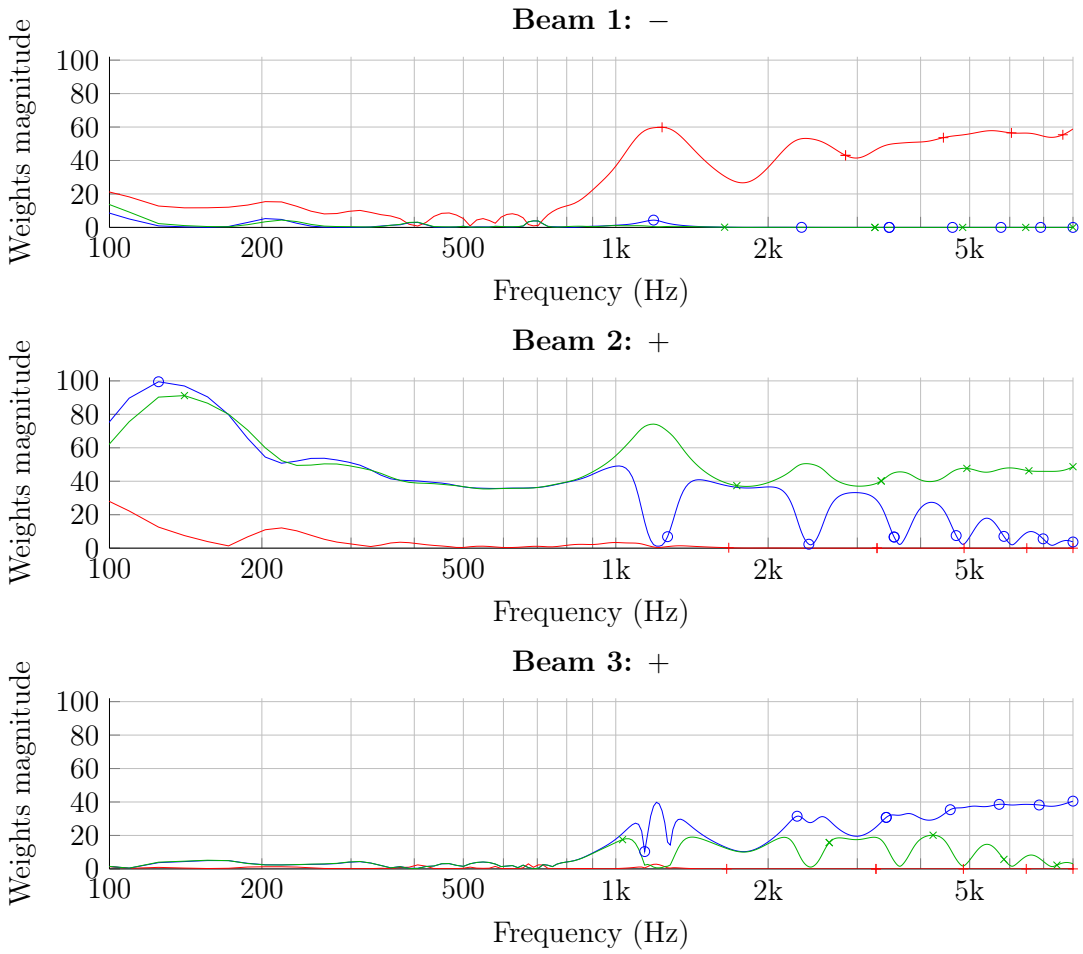

Figure 8.6: Power domain beamformer weights, which create three separate beamformers. $M1$ is red, $M2$ is blue and $M3$ is green, the plot title represents the median elements of g.

beams median weight in the power domain, where a $+$ refers to addition and a $-$ refers to subtraction.

Each beam acts in a similar nature to the robust SNR-optimal beamformer, in that they utilise combining fewer microphones at higher frequencies. However as the phase uncertainties have a much smaller impact on the power of each frame, we can still combine the powers of the beamformer outputs at high frequencies. As such the

power domain beamformer subtracts the interference beam, beam 1, whilst adding the other two beams. This causes spectral subtraction and helps to further remove the interference source.

The beamformers were applied to simulated speech signals. The resulting signals were evaluated with six metrics. These include the previously mentioned STFT domain MSE, (8.32), the expected SNR, (8.34), Perceptual Evaluation of Speech Quality (PESQ) [12] and the Short-Time Objective Intelligibility [16]. To ensure the signals were correctly time aligned, the weights for each beamformer were normalised so that the desired source is processed with the same phase as a reference signal consisting of only the desired source. The results are shown in the Table 8.1. The RPDB

Weights	E_{STFT}	SNR	PESQ	STOI
BestM	-0.47	-3.04	1.35	0.51
MVDR	-0.13	0.15	1.52	0.63
SNRrob	-0.69	4.95	1.78	0.69
RPDB	**-2.85**	**12.36**	1.99	0.77
Oracle	-0.27	5.35	**2.00**	**0.79**

Table 8.1: Explained Example results

out-performs the other robust beamformers in all categories and offers similar performance to the Oracle beamformer. This is further evident from the spectrograms, shown in Fig. 8.7. The power domain beamformer is closer to the original reference signal than the MVDR beamformer output.

8.6.2 Performance over random geometries

In order to access the performance of the Robust Power Domain Beamformer a series of tests were conducted over different array geometries. The number of sources was either one, two or four, and there were two, four or eight microphones. For each scenario, the sources and microphones were placed in 50 different random positions. The rooms were varied in order to get a large range of SNRs for the best microphone case.

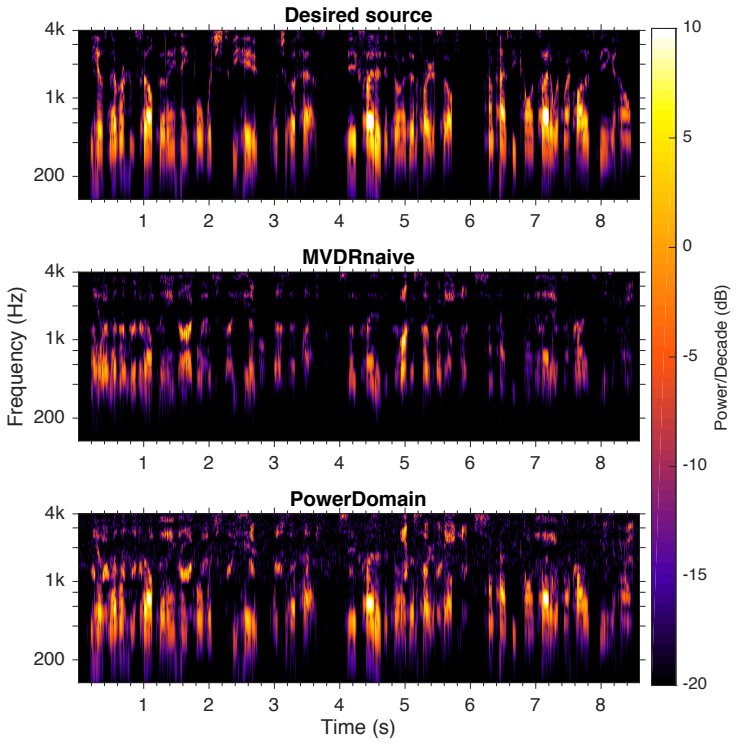

Figure 8.7: Spectrograms of output signals

For the case of eight microphones and four sources, the expected SNR of the output signal is taken for each test and compared with the SNR of the best case microphone in each test. The difference in performance is shown in the box plots in Figs. 8.8, and the difference in PESQ is shown in Fig. 8.9. When there is no uncertainty present in the propagation coefficients, all beamformer designs produce a gain in expected SNR against the best microphone. The RPDB produces an expected gain in SNR greater than all other beamformers, including the oracle beamformer (Welch T-test p-value 3×10^{-3}). As the beamformer is implemented in the power domain it can use spectral subtraction to reduce interference and noise further than any linear beamformer.

When uncertainty is introduced into the element positions and the channel propagation velocity, the performance of many of the beamformers degrades. Combining

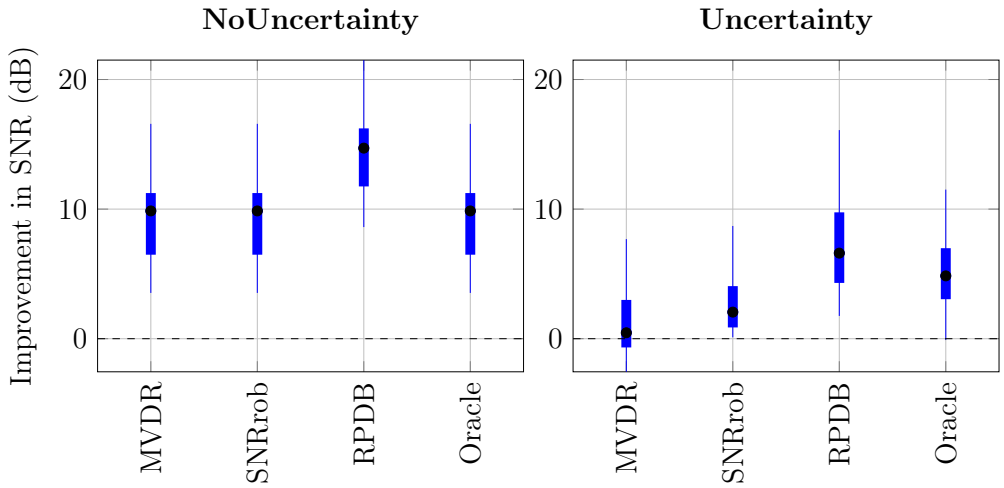

Figure 8.8: Improvement in expected SNR relative to the best microphone over 50 simulations, excluding and including uncertainties.

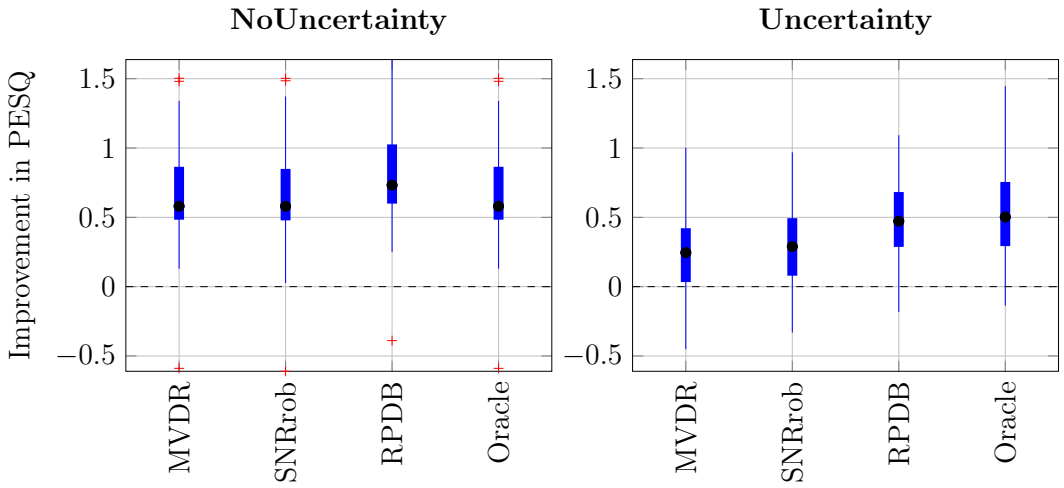

Figure 8.9: Improvement in PESQ relative to the best microphone over 50 simulations, excluding and including uncertainties.

multiple microphones, when the phase of each is not well known, causes poor results. At high SNRs the best microphone outperforms the MVDR beamformer. The high levels of uncertainty adversely affects the performance of the MVDR beamformer

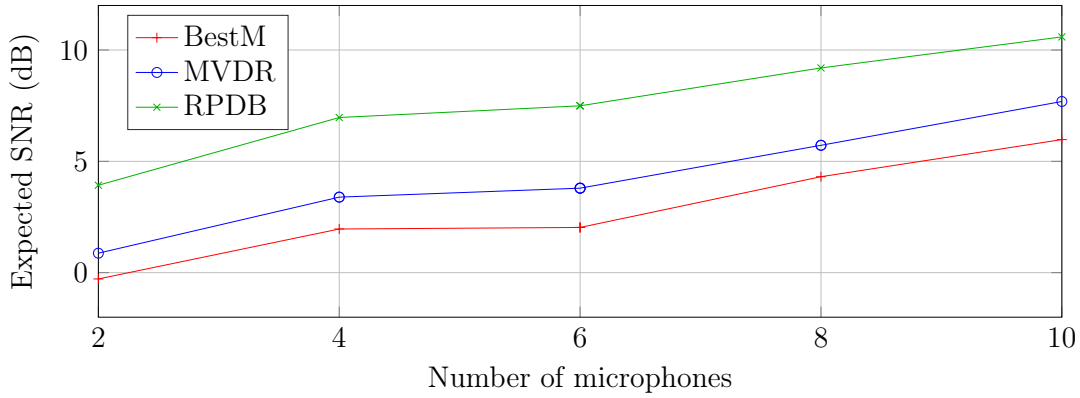

Figure 8.10: The median Expected SNR against the number of microphones in the array.

when combining microphones together, whereas this does not effect the single microphone case.

However the beamformers that are robust to the uncertainties do not suffer in the same way. The RPDB beamformer is the best out of all the conventional beamformers (RPDB>SNRrob Welch T-test p-value 0.015). The oracle beamformer represents the best possible linear beamformer performance with a near perfecting tracking beamformer, in reality this performance would not be achieved. However, the RPDB beamformer offers comparable performance to the oracle beamformer in SNR performance and is close in PESQ.

Number of microphones The number of microphones in the array changes the performance benefits of the RPDB over the competing beamformers. The results in Fig. 8.10 shows the median A-weighted expected SNR for 50 random geometries of different number of microphones, in the presence of two sources, one of which is an interferer. As the number of microphones increases the SNR of all beamformers increase, as there are more channels available to extract the desired source. The RPDB has similar performance gains over competing beamformers regardless of the number of microphones.

General results Across all 450 geometries the performance difference between each beamformer and the best microphone case was found. The resulting performance deviations were averaged over all geometries and presented below, where $\mathrm{E_{STFT}}$ is lower-the-better and the remaining columns are higher-the-better.

The results shown in Table 8.2 include contributions for position and channel uncertainties. The proposed RPDB demonstrates performance close to the oracle

Weights	$\triangle \mathrm{E_{STFT}}$ (dB)	$\triangle$SNR (dB)	$\triangle$PESQ	$\triangle$STOI
MVDR	1.07	1.28	0.16	0.17
SNRrob	1.09	1.75	0.18	0.17
RPDB	**-2.50**	**3.46**	0.32	0.20
Oracle	0.64	3.20	**0.34**	**0.22**

Table 8.2: Average results over 450 geometries comparing the performance of each beamformer.

beamformer and improves on average SNR over the oracle beamformer.

Robustness The RPDB requires knowledge of the ratio of fourth order moments to second moments squared of the speech, $\left\langle |\tilde{s}|^4 \right\rangle / \left\langle |\tilde{s}|^2 \right\rangle^2$, as estimated using LTASS and the complex kurtosis, (3.2). The performance when under and over scaling the above ratio is investigated. The ratio is scaled by a scaling factor before the weights are computed and the performance is measured. In this test four microphones and two sources, one of which is an interferer, are used. Ten different geometries are used and results are averaged over all 10 for each scaling value, the results are shown in Fig. 8.11. The results show a small change in PESQ with respect to large changes in $\left\langle |\tilde{s}|^4 \right\rangle$, thus the RPDB is not particularly sensitive to errors in $\left\langle |\tilde{s}|^4 \right\rangle$.

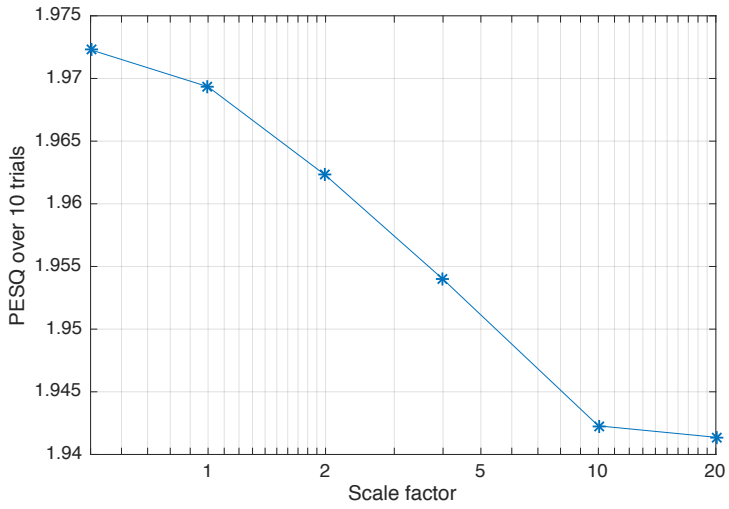

Figure 8.11: PESQ results for different scaling of $\langle |\tilde{s}|^4 \rangle$. Four microphones, two sources, 10 geometries.

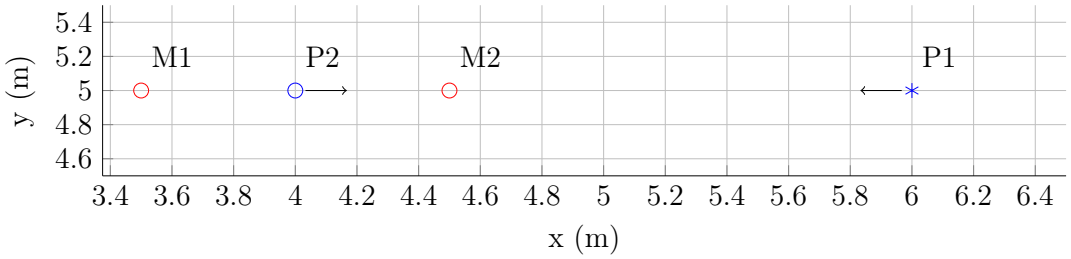

Figure 8.12: The two sources, $P1$ and $P2$, are on average facing each other.

8.7 Amplitude uncertainties

In the follow section we introduce amplitude uncertainties as well as time uncertainties into the propagation coefficients:

$$\mathbf{D} \triangleq \bar{\mathbf{D}} \odot \tilde{\mathbf{H}} \odot \exp\left(j\omega_k \tilde{\mathbf{T}}\right). \tag{8.36}$$

8.7.1 Explained example

Consider the geometry consisting of two microphones and two sources shown in Fig. 8.12. The two sources are on average facing towards each other, each has a range of rotation equal to $\pm 22.5°$. The positions and channels are assumed time varying with known statistics.

The weights of a power domain beamformer designed without the knowledge of the amplitude uncertainties (RPDBnon) are shown in Fig. 8.13. Assuming no amplitude uncertainties, the SNR at each microphone is -14 dB and -9.5 dB respectively. Therefore, the beamformer attempts to subtract the signal $M1$ from $M2$ in the power domain. The interference power is assumed equal in each microphone, thus subtracting one from the other should remove the interference. However, when we consider the amplitude uncertainties the interference will not cancel correctly.

The amplitude uncertainties for the geometry are shown in Fig. 8.14. Similarly to the example shown in Sec. 7.3.1, the power due to amplitude uncertainties from the

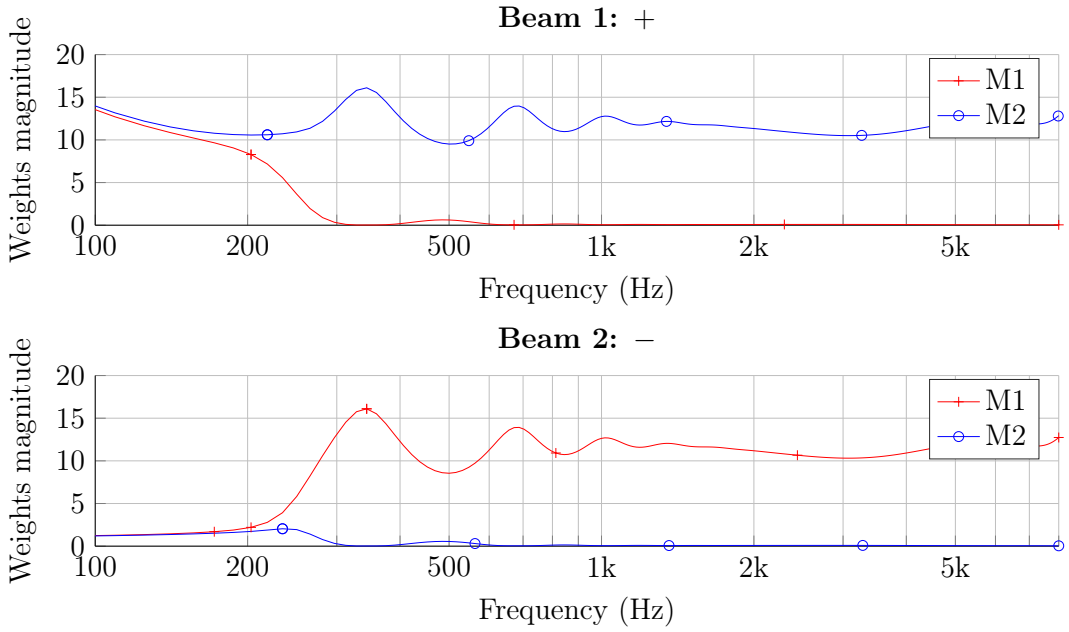

Figure 8.13: The weights of each beamformer that comprise the RPDBnon. The top beamformer is on average added in the power domain, the bottom beamformer is on average subtracted in the power domain.

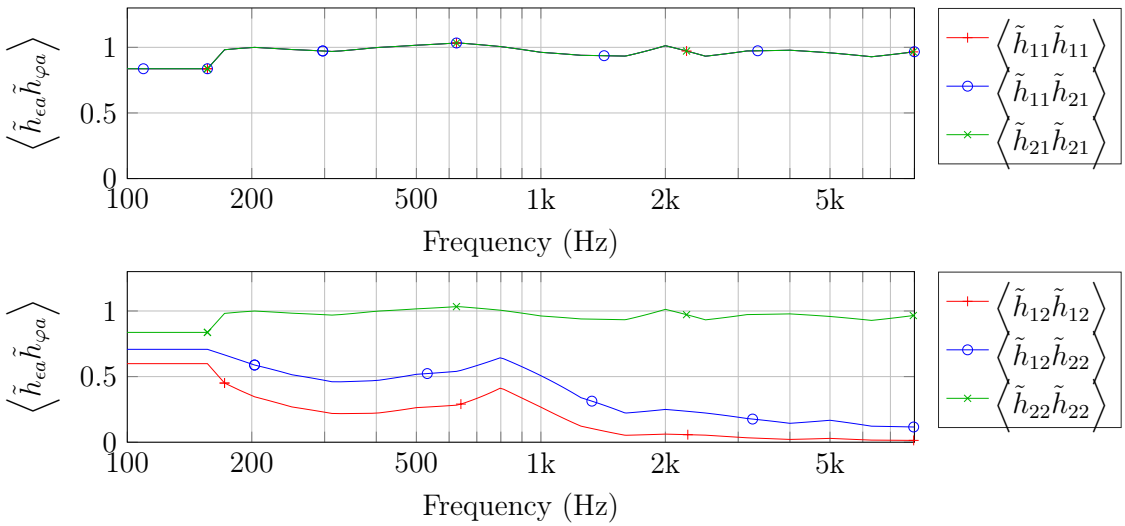

Figure 8.14: The covariance in amplitude uncertainties across different channels for $P1$ (top) and $P2$ (bottom).

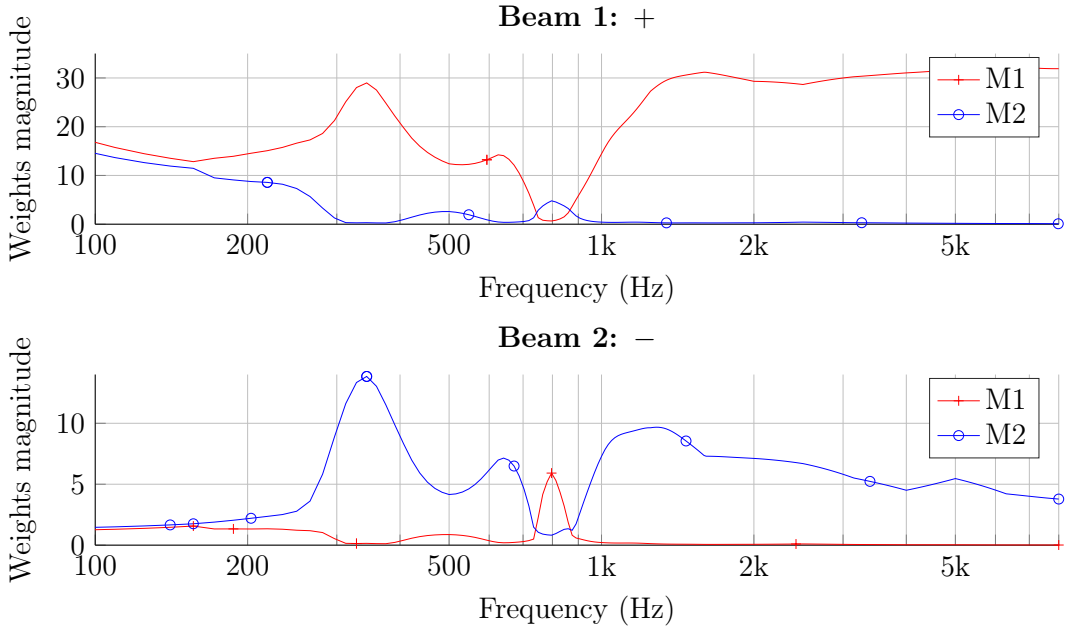

Figure 8.15: The weights of each beamformer that comprise the RPDB. The top beamformer is on average added in the power domain, the bottom beamformer is on average subtracted in the power domain.

interference, $P2$, in the forwards direction, $\left\langle \tilde{h}_{22}\tilde{h}_{22} \right\rangle$, is on average 15 times greater than the power due to amplitude uncertainties observed behind the source, $\left\langle \tilde{h}_{12}\tilde{h}_{12} \right\rangle$. The power due to amplitude uncertainties observed from the desired source, $P1$, is the same in each microphone, $\left\langle \tilde{h}_{11}\tilde{h}_{11} \right\rangle = \left\langle \tilde{h}_{12}\tilde{h}_{12} \right\rangle$. As much less power is seen behind the interference source than in-front, the SNR for $M1$ has improved and the SNR at each microphone is now -6.4 dB and -9.5 dB respectively. Using the amplitude uncertainties we can design a robust power domain beamformer, RPDB. The resulting weights are shown in Fig. 8.15. In contrast to RPDBnon, the robust beamformer subtracts $M2$ from $M1$. The greater SNR at $M1$ means it is used as the basis for the desired source. The weight for $M2$ in the second beamformer is chosen to match the expected power of the interference in $M2$ to that of $M1$, thus when we subtract in the power domain we maximise the resulting SNR. At higher frequencies the $M2$ weighting in the second beamformer reduces to reflect the increased directivity of the

interference source, less interference power is observed at $M1$, so a smaller weight is needed to match the expected power.

The resulting signals were measured with a series of metrics and the results are shown in Table 8.3. The failed suppression of the interference in the RPDBnon causes

Weights	E_{STFT}	SNR	PESQ	STOI
BestM	0.03	-9.54	0.98	0.47
MVDR	0.07	-10.76	0.82	0.51
SNRrob	-0.27	-3.89	1.51	0.65
RPDBnon	0.05	-10.33	0.92	0.50
RPDB	**-3.51**	**6.80**	**1.78**	**0.77**

Table 8.3: Explained Example results

a large performance degradation, which means the performance of the amplitude robust power domain beamformer (RPDB) surpassed all others.

8.7.2 Performance over random geometries

We consider the average performance over 700 random geometries consisting of up to 4 sources and 8 microphones. The sources used the directivity pattern of the head and were constrained to a static random direction. The median performance gains over the best case microphone are shown in Table 8.4. The power domain beamformer

Weights	$\triangle E_{STFT}$ (dB)	$\triangle$SNR (dB)	$\triangle$PESQ	$\triangle$STOI
MVDR	0.00	1.72	0.09	0.15
SNRrob	0.00	2.51	0.16	0.17
RPDBnon	-0.62	6.14	0.28	0.19
RPDB	**-1.72**	**6.25**	**0.31**	**0.22**

Table 8.4: Average results over 700 geometries comparing the performance of each beamformer.

has a small increase in performance over other beamformers by considering the voice directivity pattern. The results will change with different rotation ranges assumed for the sources.

8.8 Conclusions

In this contribution we have designed a novel power domain beamformer in which the outputs from M independent linear beamformers are combined in the power domain. The weights of the individual beamformers and the signs of their contributions in the power domain are chosen to minimise the MSE in the STFT power domain. We have compared the performance of the proposed beamformer against several competing beamformers through simulations. In the absence of uncertainties, the power domain beamformer gives performance gains that are not possible with a time domain beamformer. When uncertainties in the steering vectors are taken into account, the performance remains superior to competing methods.

Chapter 9

Conclusion

9.1 Thesis summary

In this thesis we have presented an algorithm for dereverberation in the STFT domain and methods of designing optimal robust beamformers, in the STFT domain and the power domain, that make use of models of uncertainties in the channel propagation coefficients.

Chapter 4 detailed a novel algorithm for dereverberation in the STFT domain. The approach utilises near by frames to remove reverberation. It overcomes the shortfalls of the time domain algorithms. With knowledge of the impulse response it can successfully remove significant amounts of reverberation at least up to -45 dB DRR.

Beamformers exploit the spatial diversity of the acoustic sources in the environment in order to suppress interference and amplify a desired source location. Conventional beamformers require that the propagation channels are deterministic function of the source and microphone placements. However in reality, there are uncertainties in the phase and magnitude responses of the source-microphone channels and these uncertainties become larger with increasing microphone separation. In these

circumstances the performance of conventional beamformers degrades. Robust approaches to beamforming reduce this degradation through different methods. The robust beamformer designs rely on tuning parameters whose choice is not always well defined or based on the intended application of acoustic channels.

In Chapter 3 we introduced a detailed analysis of speech signals and the corresponding short time Fourier transform (STFT) coefficients. We showed that a Generalised complex Gaussian distribution provided a much better fit to the STFT coefficients of speech than the complex Gaussian distribution that is conventionally used. We used the distribution to form complex kurtosis functions from the second order STFT coefficient statistics and the fourth order statistics. This allowed us to design beamformers in Chapter 8 that use the fourth order statistics.

Realistic acoustic propagation channels were modeled in Chapter 5. The traditional models were extended to incorporate a series of uncertainties in both time and amplitude. The time uncertainties model was validated with experimental data. The amplitude uncertainties model was based on previous measurements of the human head directivity pattern combined with a model of head rotation.

In Chapter 7 we derived a robust STFT domain beamformer, which uses the model of channel propagation uncertainties from Chapter 5. We showed that the proposed STFT domain beamformer is robust to time uncertainties and offers performance gains over competing beamformers under similar scenarios. Incorporating a random head rotation model into each source made it possible to ensure robustness to the direction in which the speaker is facing.

In Chapter 8 we introduced a novel beamformer that extends the traditional SFTF domain beamformer into the power domain. If the microphones in an array are too widely separated, the effects of source movement and sound speed variations destroy the phase correlations that are essential for effective linear beamforming in the STFT-domain. By performing beamforming in the power domain instead it is possible to take advantage of the strong inter-microphone correlations that exist in the power domain even for widely separated microphones. We derived a power domain beamformer that is optimal in the mean squared error sense and showed that it can be decomposed into a set of linear STFT domain beamformers whose outputs are

added or subtracted in the power domain. The design process automatically takes advantage both of the phase correlations between closely spaced microphones at low frequencies and of the power correlations between widely spaced microphones. It is effectively able to perform spectral subtraction on the output of optimally designed linear beamformers to suppress interference and noise. On simulated data the performance of the proposed power domain beamformer exceeds that of an ideal MVDR beamformer. We utilise the uncertain propagation coefficients to ensure robustness to time and amplitude uncertainties, which prevents the performance degrading in uncertain scenarios.

9.2 Future research

Outlined below are several directions in which the research could be taken further forwards. These have been divided into extensions of the acoustic propagation modeling and to the power-domain robust beamformer.

9.2.1 Acoustic propagation modeling

The time uncertainties framework introduced is extensible to any uncertainty contribution that can be modeled with its covariance. We have included two sources of such uncertainty, however, there are many further cases that could be included. For example, in primarily outdoor scenarios, the wind speed can play a factor in the propagation speed. The component of the wind speed along the propagation channel will directly add or subtract from the channel propagation speed.

Occlusions in the propagation channels can affect the amplitude and phase of the channel. Sound diffracts around objects which are similar in size to the wavelength of the sound. Modeling occlusions will enhance the accuracy of the uncertainties. Beamformers combined with visual tracking may be able to locate occlusions in order to factor them into the propagation coefficients.

3D head rotations When modeling the variations in source radiation pattern due to head movement we have only considered the radiation pattern within a horizontal plane and assume that the head rotates only around a vertical axis. In many scenarios this is a reasonable model, however it would be straightforward to extend the model into three dimensions by incorporating the elevation into the direction of the talker. In this case the directivity model can be decomposed into spherical harmonics, which map over a sphere, instead of the circular Fourier basis.

Furthermore, both the time and amplitude uncertainty models would be made more accurate by using data from measurements of the actual talker's head. Knowledge of the translation and rotation movements could be accurately incorporated in the model.

9.2.2 Power domain beamformer

Extending Head Rotation Range The amplitude uncertainties robust beamformer can be extended to larger ranges in head rotation. Current results for larger head rotations offer variable improvements. The expected MSE always gives an improvement, although the expected SNR does not always follow the same trend. The performance gains in the STFT domain beamformer are more significant than those of the power domain equivalent. One option would be to maximise the expected SNR as a function of the power domain weights.

Reverberation model The power domain beamformer can be extended to reverberant environments by use of the propagation coefficients, $\mathbf{D}$. Each source reflects off the surfaces in the room and creates several longer propagation paths, with each separate path being represented by a different propagation coefficient. Reflection coefficients can also be attributed into $\mathbf{D}$ by reducing the magnitude of the propagation coefficient with respect to the number of surfaces it has been reflected off. The resulting signal at the m-th microphone is the sum of all the path propagation coefficients

multiplied by the source coefficients:

$$x_\epsilon[l,k] \triangleq \sum_{b=1}^{B}\sum_{a=1}^{P} d_{\epsilon a}(b)s_a[l,k] + v_\epsilon[l,k], \tag{9.1}$$

where $d_{\epsilon a}(b)$ is the b-th path from source a to microphone ϵ, and we consider B multi-paths in total. Alternatively in vector notation it is denoted as:

$$\mathbf{x} = \sum_{b=1}^{B} \mathbf{D}_b\mathbf{s} + \mathbf{v}.$$

Adaptive RPDB The proposed power domain beamformer can be formulated in an adaptive manner by considering an initial solution and an update term. As the optimal weights use the derivative of the MSE in the STFT domain, we can utilise the same derivative in a gradient descent method. In which case the performance will converge to the optimal weights.

Clipping effects In this case we are clipping all resulting negative powers to 0. However, different clipping algorithms will have different effects on the resulting musical noise, [23, 120, 121, 122]. This is a well studied area of spectral subtraction, and the results are easily applied to the clipping function in the power domain beamformer.

Optimal Amplitude and Log Domain Beamformers The action of power domain beamformer presented in this thesis is similar to that of a speech enhancer that uses spectral subtraction. It has been found in the speech enhancement literature that the perceived quality of enhanced speech is often greater when minimizing errors in the amplitude or log amplitude domains instead of the power domain. It is likely that the ideas presented in this thesis could be extended into these domains.

Appendix A

SNR-Optimal Beamformer

The SNR-optimal beamformer is described in Sec. 6.4. In the following section, the beamformer weights are derived. The SNR is defined as:

$$\mathrm{SNR} = \frac{\left\langle |\tilde{s}|^2 \right\rangle \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}}.$$

It forms a generalised eigenvalue problem. In order to maximise the SNR, we take the derivative and set it to zero.

$$\begin{aligned}
\frac{d}{d\mathbf{w}} \frac{\left\langle |\tilde{s}|^2 \right\rangle \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}} &= 2 \frac{\left\langle |\tilde{s}|^2 \right\rangle \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}} - 2 \frac{\left\langle |\tilde{s}|^2 \right\rangle \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} \mathbf{V} \mathbf{w}}{\left(\mathbf{w}^H \mathbf{V} \mathbf{w}\right)^2} = 0 \\
&= \mathbf{w}^H \mathbf{V} \mathbf{w} \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} - \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} \mathbf{V} \mathbf{w} = 0 \\
\mathbf{w}^H \mathbf{V} \mathbf{w} \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} &= \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} \mathbf{V} \mathbf{w} \\
\left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} &= \frac{\mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}} \mathbf{V} \mathbf{w} \\
\mathbf{V}^{-1} \left\langle |\tilde{s}|^2 \right\rangle \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} &= \frac{\left\langle |\tilde{s}|^2 \right\rangle \mathbf{w}^H \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w}}{\mathbf{w}^H \mathbf{V} \mathbf{w}} \mathbf{w} \\
\mathbf{V}^{-1} \left\langle |\tilde{s}|^2 \right\rangle \left\langle \bar{\mathbf{d}}_1 \bar{\mathbf{d}}_1^H \right\rangle \mathbf{w} &= \lambda_{max} \mathbf{w}.
\end{aligned}$$

Appendix B

Quartic Expectations

In the following sections we derive the quartic expectations of the array data that are required in order to form matrices need for the optimal power domain beamformer weights in Sec. 8.4.

B.1 Quartic in x

The first term, needed to form $\mathbf{Q}$, $\left\langle x_\varepsilon x_\varphi^* x_\rho x_\tau^* \right\rangle$, can be expanded by substituting in the expression for $x_\epsilon = \mathbf{d}_\epsilon^T \mathbf{s} + v_\epsilon$, from (6.1), where $\mathbf{d}_\epsilon^T$ is a row vector: $\mathbf{d}_\epsilon^T = \begin{bmatrix} d_{\epsilon,1} & \dots & d_{\epsilon,P} \end{bmatrix}$.

$$\begin{aligned} \left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle &= \left\langle \left(\mathbf{d}_\epsilon^T \mathbf{s} + v_\epsilon\right) \left(\mathbf{d}_\varphi^T \mathbf{s} + v_\varphi\right)^* \left(\mathbf{d}_\rho^T \mathbf{s} + v_\rho\right) \left(\mathbf{d}_\tau^T \mathbf{s} + v_\tau\right)^* \right\rangle && \text{(B.1)} \\ &= \left\langle \left(\mathbf{d}_\epsilon^T \mathbf{s} \mathbf{d}_\varphi^H \mathbf{s}^* + \mathbf{d}_\epsilon^T \mathbf{s} v_\varphi^* + v_\epsilon \mathbf{d}_\varphi^H \mathbf{s}^* + v_\epsilon v_\varphi^*\right) \right. \\ &\qquad \left. \left(\mathbf{d}_\rho^T \mathbf{s} \mathbf{d}_\tau^H \mathbf{s}^* + \mathbf{d}_\rho^T \mathbf{s} v_\tau^* + v_\rho \mathbf{d}_\tau^H \mathbf{s}^* + v_\rho v_\tau^*\right) \right\rangle && \text{(B.2)} \end{aligned}$$

The noise and speech terms are assumed to be independent and zero-mean. Therefore the expectation of terms which involve an odd number of noise or speech terms will

be 0, $\langle s_\epsilon v \rangle = 0$. Thus we can simplify the above to:

$$\begin{aligned}\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle &= \left\langle \mathbf{d}_\epsilon^T \mathbf{s} \mathbf{d}_\varphi^H \mathbf{s}^* \left(\mathbf{d}_\rho^T \mathbf{s} \mathbf{d}_\tau^H \mathbf{s}^* + v_\rho v_\tau^* \right) \right. \\ &\quad + \mathbf{d}_\epsilon^T \mathbf{s} v_\varphi^* \left(\mathbf{d}_\rho^T \mathbf{s} v_\tau^* + v_\rho \mathbf{d}_\tau^H \mathbf{s}^* \right) \\ &\quad + v_\epsilon \mathbf{d}_\varphi^H \mathbf{s}^* \left(\mathbf{d}_\rho^T \mathbf{s} v_\tau^* + v_\rho \mathbf{d}_\tau^H \mathbf{s}^* \right) \\ &\quad \left. + v_\epsilon v_\varphi^* \left(\mathbf{d}_\rho^T \mathbf{s} \mathbf{d}_\tau^H \mathbf{s}^* + v_\rho v_\tau^* \right) \right\rangle . \end{aligned} \tag{B.3}$$

Both the complex speech coefficients, $\mathbf{s}$, and the noise coefficients, $\mathbf{v}$, have independent real and imaginary parts, $\langle s_\epsilon s_\epsilon \rangle = 0$, $\langle v_\epsilon v_\epsilon \rangle = 0$, thus only paired complex conjugate terms are non-zero when expanded:

$$\begin{aligned}\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle = \left\langle \right. & \mathbf{d}_\epsilon^T \mathbf{s} \mathbf{d}_\varphi^H \mathbf{s}^* \mathbf{d}_\rho^T \mathbf{s} \mathbf{d}_\tau^H \mathbf{s}^* + \mathbf{d}_\epsilon^T \mathbf{s} \mathbf{d}_\varphi^H \mathbf{s}^* v_\rho v_\tau^* \\ & + \mathbf{d}_\epsilon^T \mathbf{s} v_\varphi^* v_\rho \mathbf{d}_\tau^H \mathbf{s}^* + v_\epsilon \mathbf{d}_\varphi^H \mathbf{s}^* \mathbf{d}_\rho^T \mathbf{s} v_\tau^* \\ & \left. + v_\epsilon v_\varphi^* \mathbf{d}_\rho^T \mathbf{s} \mathbf{d}_\tau^H \mathbf{s}^* + v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle . \end{aligned} \tag{B.4}$$

Separating the various expectations gives the following:

$$\begin{aligned}\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle &= \left\langle \left(\left(\mathbf{d}_\tau \mathbf{d}_\rho^H \right) \colon \left(\mathbf{d}_\varphi \mathbf{d}_\epsilon^H \right) \colon {}^T \right) \colon \right\rangle^H \left\langle \left(\mathbf{s}^* \mathbf{s}^T \right) \colon \left(\mathbf{s}^* \mathbf{s}^T \right) \colon {}^T \right\rangle \colon \\ &\quad + \left(\left\langle v_\rho^* v_\tau \right\rangle \left\langle \left(\operatorname{diag} \left(\mathbf{d}_\varphi \mathbf{d}_\epsilon^H \right) \right) \colon \right\rangle \right. \\ &\quad\quad + \left\langle v_\varphi v_\rho^* \right\rangle \left\langle \left(\operatorname{diag} \left(\mathbf{d}_\tau \mathbf{d}_\epsilon^H \right) \right) \colon \right\rangle \\ &\quad\quad + \left\langle v_\epsilon^* v_\varphi \right\rangle \left\langle \left(\operatorname{diag} \left(\mathbf{d}_\tau \mathbf{d}_\rho^H \right) \right) \colon \right\rangle \\ &\quad\quad \left. + \left\langle v_\epsilon^* v_\tau \right\rangle \left\langle \left(\operatorname{diag} \left(\mathbf{d}_\rho^* \mathbf{d}_\varphi^T \right) \right) \colon \right\rangle \right)^H \\ &\quad \left\langle \operatorname{diag} \left(\mathbf{s} \mathbf{s}^H \right) \right\rangle \colon + \left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle . \end{aligned} \tag{B.5}$$

The summations over the quadratic terms can be combined into a single double summation, and $\langle s_a s_b^* \rangle$ can be factored out:

$$\begin{aligned}\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle = &\sum_{a,b,c,d} \left\langle d_{\epsilon,a} d_{\varphi,b}^* d_{\rho,c} d_{\tau,d}^* \right\rangle \langle s_a s_b^* s_c s_d^* \rangle + \left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle \\ &+ \sum_{a,b} \langle s_a s_b^* \rangle \left(\left\langle d_{\epsilon,a} d_{\varphi,b}^* \right\rangle \langle v_\rho v_\tau^* \rangle + \left\langle d_{\epsilon,a} d_{\tau,b}^* \right\rangle \left\langle v_\varphi^* v_\rho \right\rangle \right. \\ &\left. + \left\langle d_{\varphi,a}^* d_{\rho,b} \right\rangle \langle v_\epsilon v_\tau^* \rangle + \left\langle d_{\rho,a} d_{\tau,b}^* \right\rangle \left\langle v_\epsilon v_\varphi^* \right\rangle \right). \end{aligned} \tag{B.6}$$

The sources are assumed to be independent, as such $\langle s_a s_b^* \rangle = 0,\ a \neq b$. Therefore the last line can be simplified:

$$\begin{aligned}\left\langle x_\epsilon x_\varphi^* x_\rho x_\tau^* \right\rangle = &\sum_{a,b,c,d} \left\langle d_{\epsilon,a} d_{\varphi,b}^* d_{\rho,c} d_{\tau,d}^* \right\rangle \langle s_a s_b^* s_c s_d^* \rangle \\ &+ \left\langle v_\epsilon v_\varphi^* v_\rho v_\tau^* \right\rangle \\ &+ \sum_{a} \langle s_a s_a^* \rangle \left(\left\langle d_{\epsilon,a} d_{\varphi,a}^* \right\rangle \langle v_\rho v_\tau^* \rangle + \left\langle d_{\epsilon,a} d_{\tau,a}^* \right\rangle \left\langle v_\varphi^* v_\rho \right\rangle \right. \\ &\left. + \left\langle d_{\varphi,a}^* d_{\rho,a} \right\rangle \langle v_\epsilon v_\tau^* \rangle + \left\langle d_{\rho,a} d_{\tau,a}^* \right\rangle \left\langle v_\epsilon v_\varphi^* \right\rangle \right). \end{aligned} \tag{B.7}$$

B.1.1 Quartic speech term

A similar analysis can be applied to the quartic speech term, which appears as the first term on the right hand side of (B.7):

$$\sum_{a,b,c,d} \left\langle d_{\epsilon,a} d_{\varphi,b}^* d_{\rho,c} d_{\tau,d}^* \right\rangle \langle s_a s_b^* s_c s_d^* \rangle . \tag{B.8}$$

When expanding the above summations, the speech expectation term, $\langle s_a s_b^* s_c s_d^* \rangle$, will only present non-zero terms in a subset of cases. Firstly, when all indices are the same we obtain a quartic in the source powers:

$$\sum_{a} \left\langle d_{\epsilon,a} d^*_{\varphi,a} d_{\rho,a} d^*_{\tau,a} \right\rangle \left\langle |s_a|^4 \right\rangle . \tag{B.9}$$

Secondly, there are many cross terms which result in quadratics. This occurs when there is a set of conjugate pairs in the source indices, $\langle s_a s^*_b s_c s^*_d \rangle$, i.e. $a = b \cap c = d$ or $a = d \cap c = b$. The unique set of non-zero quadratic terms is specified in the summation:

$$\begin{aligned} &\sum_{a=1}^{P-1} \sum_{b=a+1}^{P} \left\langle s_a s^*_a \right\rangle \left\langle s_b s^*_b \right\rangle \\ &\left\langle d_{\epsilon,a} d^*_{\varphi,a} d_{\rho,b} d^*_{\tau,b} + d_{\epsilon,a} d^*_{\varphi,b} d_{\rho,b} d^*_{\tau,a} \right. \\ &\quad \left. + d_{\epsilon,b} d^*_{\varphi,a} d_{\rho,a} d^*_{\tau,b} + d_{\epsilon,b} d^*_{\varphi,b} d_{\rho,a} d^*_{\tau,a} \right\rangle . \end{aligned} \tag{B.10}$$

The general expression can be formed by combining all the above terms:

$$\begin{aligned} &\sum_{a,b,c,d} \left\langle d_{\epsilon,a} d^*_{\varphi,b} d_{\rho,c} d^*_{\tau,d} \right\rangle \left\langle s_a s^*_b s_c s^*_d \right\rangle = \\ &\sum_{a=1}^{P} \left\langle d_{\epsilon,a} d^*_{\varphi,a} d_{\rho,a} d^*_{\tau,a} \right\rangle \left\langle |s_a|^4 \right\rangle \\ &+ \sum_{a=1}^{P-1} \sum_{b=a+1}^{P} \left\langle |s_a|^2 \right\rangle \left\langle |s_b|^2 \right\rangle \\ &\quad \left\langle d_{\epsilon,a} d^*_{\varphi,a} d_{\rho,b} d^*_{\tau,b} + d_{\epsilon,a} d^*_{\varphi,b} d_{\rho,b} d^*_{\tau,a} \right. \\ &\qquad \left. + d_{\epsilon,b} d^*_{\varphi,a} d_{\rho,a} d^*_{\tau,b} + d_{\epsilon,b} d^*_{\varphi,b} d_{\rho,a} d^*_{\tau,a} \right\rangle . \end{aligned} \tag{B.11}$$

B.2 MSE optimisation

The MSE in the STFT domain, $\mathrm{E}_{\mathrm{STFT}}$, is affected by scaling the amplitude of the beamformer output. The metric reduces if there is mismatch between the desired

signal in the reference and the beamformer output. The optimal scaling factor, ς, to minimise the $\mathrm{E}_{\mathrm{STFT}}$ can derived as follows. Given $\mathrm{E}_{\mathrm{STFT}}$ as:

$$\mathrm{E}_{\mathrm{STFT}} = \sum_{\omega=1}^{N_\omega} A_\omega \left(\frac{\mathrm{MSE}_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \right).$$

the optimal value for ς can be found through differentiation.

$$\begin{aligned}
\frac{d\mathrm{E}_{\mathrm{STFT}}}{d\varsigma} &= \frac{d}{d\varsigma} \sum_{\omega=1}^{N_\omega} A_\omega \left(\frac{\mathrm{MSE}_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \right) = 0 \\
&= \frac{d}{d\varsigma} \sum_{\omega=1}^{N_\omega} A_\omega \left(\frac{\varsigma^4 \left\langle \mathrm{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle_\omega - 2\varsigma^2 \left\langle \mathrm{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right)_\omega \tilde{s}_\omega \tilde{s}_\omega^H \right\rangle + \left\langle |\tilde{s}_\omega|^4 \right\rangle}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \right) \\
0 &= \sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left(4\varsigma^3 \left\langle \mathrm{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle_\omega - 4\varsigma \left\langle \mathrm{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right)_\omega \tilde{s}_\omega \tilde{s}_\omega^H \right\rangle \right).
\end{aligned}$$

Removing the case when $\varsigma = 0$, we can divide through by ς:

$$\begin{aligned}
0 &= \sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left(\varsigma^2 \left\langle \mathrm{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle_\omega - \left\langle \mathrm{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right)_\omega \tilde{s}_\omega \tilde{s}_\omega^H \right\rangle \right) \\
&= \varsigma^2 \sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left\langle \mathrm{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle_\omega - \sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left\langle \mathrm{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right)_\omega \tilde{s}_\omega \tilde{s}_\omega^H \right\rangle \\
\varsigma^2 &= \frac{\sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left\langle \mathrm{tr} \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right)_\omega \tilde{s}_\omega \tilde{s}_\omega^H \right\rangle}{\sum_{\omega=1}^{N_\omega} \frac{A_\omega}{\left\langle |\tilde{s}_\omega|^4 \right\rangle} \left\langle \mathrm{tr}^2 \left(\mathbf{F}\mathbf{x}\mathbf{x}^H \right) \right\rangle_\omega}.
\end{aligned}$$

Bibliography

[1] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven (TU/e), Tech. Rep., 2006.

[2] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, Pittsburgh, USA, Sep. 2006, pp. 2614–2617.

[3] J. Allen and L. Radiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[4] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Signal Process.*, vol. 34, no. 5, pp. 1153–1161, 1986.

[5] *Subjective performance evaluation of telephone band and wideband codecs*, International Telecommunications Union (ITU-T) Recommendation P.830, 1998.

[6] *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms*, International Telecommunications Union (ITU-T) Recommendation P.835, Nov. 2003.

[7] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. 153–156.

[8] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, Jan. 1988.

[9] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, Pittsburgh, USA, Sep. 2006, pp. 1447–1450.

[10] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.

[11] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified Bark spectral distortion as an objective speech quality measure," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, May 1998, pp. 541–544.

[12] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[13] *Mapping function for transforming P.862 raw result scores to MOS-LQ*, International Telecommunications Union (ITU-T) Recommendation P.862.1, 2003.

[14] N. Kitawaki and T. Yamada, "Subjective and objective quality assessment for noise reduced speech," in *ETSI Workshop on Speech and Noise in Wideband Communication*, vol. 4, Sophia Antipolis, France, May 2007, pp. 1–2.

[15] "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," International Telecommunications Union (ITU-T), Standard P.863, Jan. 2011.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, Mar. 2010, pp. 4214–4217.

[17] ——, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[18] I. E. C. (IEC), "IEC 61672:2003: Electroacoustics – sound level meters," International Electrotechnical Commission (IEC), Tech. Rep., 2003.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology (NIST), NIST Interagency/Internal Report (NISTIR) 4930, Feb. 1993.

[20] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. E. Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

[21] *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[22] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[23] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[24] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[25] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[26] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[27] D. R. Brillinger, *Time Series: Data Analysis and Theory.* Philadelphia, USA: Society for Industrial and Applied Mathematics, 2001.

[28] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *Image Processing, IEEE Transactions on*, vol. 11, no. 2, pp. 146–158, 2002.

[29] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *J. Acoust. Soc. Am.*, vol. 86, no. 4, pp. 1404–1415, 1989.

[30] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 5, no. 1, pp. 52–56, 1995.

[31] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[32] T. K. Pogány and S. Nadarajah, "On the characteristic function of the generalized normal distribution," *Comptes rendus mathématique*, vol. 348, no. 3, pp. 203–206, 2010.

[33] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sep. 2005, pp. 121–124.

[34] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, 2011.

[35] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1, pp. 117–132, 1998.

[36] A. Sehr, M. Zeller, and W. Kellermann, "Hands-free speech recognition using a reverberation model in the feature domain," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006, pp. 1–5.

[37] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[38] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. Workshop Circuits, Systems and Signal Processing (ProRISC)*, Veldhoven, The Netherlands, Nov. 2004, pp. 250–254.

[39] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Toulouse, France: IEEE, May 2006.

[40] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2003, pp. 92–95.

[41] A. Khong and P. A. Naylor, "Adaptive blind multichannel system identification," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010, pp. 157–187.

[42] Y. Huang, J. Benesty, and J. Chen, "Adaptive blind multichannel identification," in *Springer Handbook of Speech Processing*. Springer, 2008, ch. 13, pp. 259–280.

[43] B. Widrow and E. Walach, "Adaptive signal processing for adaptive control," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 9. IEEE, 1984, pp. 191–194.

[44] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[45] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Applied Signal Processing*, vol. 2007, 2007.

[46] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Apr. 2012.

[47] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[48] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1379–1390, Sep. 2014.

[49] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[50] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," *Proc. REVERB Challenge Workshop*, 2014.

[51] ——, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[52] C. L. Lawson and R. J. Hanson, *Solving least squares problems.* Society for Industrial and Applied Mathematics, 1987.

[53] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, Feb. 1999.

[54] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunications Union (ITU-T), Recommendation P.862, Feb. 2001.

[55] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[56] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, Corpus LDC93S1, 1993.

[57] K. K. Iwai, "Pre-echo detection & reduction," Ph.D. dissertation, Massachusetts Institute of Technology, Massachusetts, USA.

[58] N. Jablon, "Adaptive beamforming with the generalized sidelobe canceller in the presence of array imperfections," *IEEE Trans. Antennas Propag.*, vol. 34, no. 8, pp. 996–1012, Aug. 1986.

[59] D. D. Feldman and L. J. Griffiths, "A projection approach for robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 867–876, 1994.

[60] Y. J. Hong, C.-c. Yeh, and D. R. Ucci, "The effect of a finite-distance signal source on a far-field steering applebaum array-two dimensional array case," *IEEE Trans. Antennas Propag.*, vol. 36, no. 4, pp. 468–475, 1988.

[61] C.-C. Yeh, Y. J. Hong, and D. R. Ucci, "The effect of a finite distance source on an Applebaum array," *IEEE Trans. Antennas Propag.*, vol. 33, no. 9, pp. 1003–1008, 1985.

[62] D. Astély and B. Ottersten, "The effects of local scattering on direction of arrival estimation with MUSIC," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3220–3234, 1999.

[63] J. Goldberg and H. Messer, "Inherent limitations in the localization of a coherently scattered source," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3441–3444, 1998.

[64] O. Besson and P. Stoica, "Decoupled estimation of DOA and angular spread for a spatially distributed source," *IEEE Trans. Signal Process.*, vol. 48, no. 7, pp. 1872–1882, 2000.

[65] H. A. Javed, A. H. Moore, and P. A. Naylor, "Spherical microphone array acoustic rake receivers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 111–115.

[66] R. Stanton and M. Brookes, "Path uncertainty robust beamforming," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lisbon, Portugal, Sep. 2014, pp. 1925–1929.

[67] A. B. Gershman, V. I. Turchin, and V. A. Zverev, "Experimental results of localization of moving underwater signal by adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2249–2257, 1995.

[68] R. Compton Jr, "The effect of random steering vector errors in the Applebaum adaptive array," *IEEE Trans. Aerosp. Electron. Syst.*, no. 4, pp. 392–400, 1982.

[69] L. C. Godara, "The effect of phase-shifter errors on the performance of an antenna-array beamformer," *Oceanic Engineering, IEEE Journal of*, vol. 10, no. 3, pp. 278–284, 1985.

[70] J. Kim and C. Un, "An adaptive array robust to beam pointing error," *IEEE Trans. Signal Process.*, vol. 40, no. 6, pp. 1582–1584, 1992.

[71] P. Annibale, J. Filos, P. A. Naylor, and R. Rabebstein, "TDOA-based speed of sound estimation for air temperature and room geometry inference," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 234–246, Feb. 2013.

[72] M. Anderson, M. McKeag, and G. Trahey, "The impact of sound speed errors on medical ultrasound imaging," *J. Acoust. Soc. Am.*, vol. 107, p. 3540, 2000.

[73] V. Del Grosso and C. Mader, "Speed of sound in pure water," *J. Acoust. Soc. Am.*, vol. 52, p. 1442, 1972.

[74] W. Wagner and A. Pruß, "The IAPWS formulation 1995 for the thermodynamic properties of ordinary water substance for general and scientific use," *J. of Physical and Chemical Reference Data*, vol. 31, p. 387, 2002.

[75] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions*. New York, USA: Wiley, 1994.

[76] T. Funato, S. Aoi, N. Tomita, and K. Tsuchiya, "Smooth enlargement of human standing sway by instability due to weak reaction floor and noise," *Royal Society Open Science*, vol. 3, no. 1, p. 150570, 2016.

[77] *MATLAB and Simulink User's Guide*, The Mathworks Inc., 2008.

[78] F. A. Everest, K. C. Pohlmann, and T. Books, *The master handbook of acoustics*. McGraw-Hill New York, 2001, vol. 4.

[79] W. T. Chu and A. Warnock, *Detailed directivity of sound fields around human talkers*. National Research Council Canada, Dec. 2002.

[80] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 5, no. 78, pp. 1508–1518, 1985.

[81] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, no. 1-2, pp. 207–222, Oct. 1993.

[82] E. Jan, P. Svaizer, and J. L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. Intl. Symp. on Circuits and Systems*, vol. 2. IEEE, 1995, pp. 1460–1463.

[83] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[84] T. Marzetta, "A new interpretation of Capon's maximum likelihood method of frequency-wavenumber spectral estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 2, pp. 445–449, Apr. 1983.

[85] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[86] M. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1378–1393, Dec. 1983.

[87] ——, "A new set of linear constraints for broad-band time domain element space processors," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 320–329, 1986.

[88] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[89] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 397–401, 1988.

[90] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, 2003.

[91] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, 2003.

[92] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1684–1696, 2005.

[93] S. Wu and J. Zhang, "A new robust beamforming method with antennae calibration errors," in *Wireless Communications and Networking Conference, 1999. WCNC. 1999 IEEE.* IEEE, 1999, pp. 869–872.

[94] S. Vorobyov, H. Chen, A. B. Gershman *et al.*, "On the relationship between robust minimum variance beamformers with probabilistic and worst-case distortionless response constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5719–5724, 2008.

[95] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[96] H. L. van Trees, *Optimum Array Processing*, ser. Detection, Estimation and Modulation Theory. Wiley, 2002.

[97] D. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays.* New York, USA: Wiley, 1980.

[98] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[99] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Wiley, 2010, ch. 9, pp. 231–268.

[100] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Am.*, vol. 54, no. 3, pp. 771–785, Sep. 1973.

[101] S. A. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *Signal Processing*, vol. 93, no. 12, pp. 3264–3277, 2013.

[102] L. Chang and C.-C. Yeh, "Performance of DMI and eigenspace-based beamformers," *IEEE Trans. Antennas Propag.*, vol. 40, no. 11, pp. 1336–1347, 1992.

[103] M. Hawkes, A. Nehorai, and P. Stoica, "Performance breakdown of subspace-based methods: Prediction and cure," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6. IEEE, 2001, pp. 4005–4008.

[104] J. K. Thomas, L. L. Scharf, and D. W. Tufts, "The probability of a subspace swap in the SVD," *IEEE Trans. Signal Process.*, vol. 43, no. 3, pp. 730–736, 1995.

[105] A. B. Gershman, *Robustness issues in adaptive beamforming and high-resolution direction finding*. New York: Marcel Dekker, 2003.

[106] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A pracical Approach*. New Jersey, USA: Wiley, 2004.

[107] J. Li and P. Stoica, *Robust adaptive beamforming*. Wiley, 2005, vol. 88.

[108] S. Vorobyov, A. B. Gershman, Z.-Q. Luo, N. Ma *et al.*, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 108–111, 2004.

[109] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust Capon beamformer," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2407–2423, 2004.

[110] A. Beck and Y. C. Eldar, "Doubly constrained robust Capon beamformer with ellipsoidal uncertainty sets," *IEEE Trans. Signal Process.*, vol. 55, no. 2, pp. 753–758, 2007.

[111] A. Hassanien, S. A. Vorobyov, and K. M. Wong, "Robust adaptive beamforming using sequential quadratic programming: An iterative solution to the mismatch problem," *Signal Processing Letters, IEEE*, vol. 15, pp. 733–736, 2008.

[112] A. Khabbazibasmenj, S. A. Vorobyov, and A. Hassanien, "Robust adaptive beamforming based on steering vector estimation with as little as possible prior information," *Signal Processing, IEEE Transactions on*, vol. 60, no. 6, pp. 2974–2987, 2012.

[113] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[114] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1732–1736, Nov. 1962.

[115] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, S. Katagiri, Ed. Artech House, 1998, ch. 16.

[116] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 173–185, 2002.

[117] S. P. Lipshitz, M. Pocock, and J. Vanderkooy, "On the audibility of midrange phase distortion in audio systems," *Journal of the Audio Engineering Society*, vol. 30, no. 9, pp. 580–595, 1982.

[118] M. Brookes, "The matrix reference manual," Imperial College London, Website, 1998-2011. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html

[119] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, pp. 134–139, 1918.

[120] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[121] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Processing Conf*, 1994, pp. 1182–1185.

[122] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.