

PERCEPTUAL MODELLING AND PROCESSING OF REVERBERANT SPEECH

HAMZA A. JAVED

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

Communication & Signal Processing Group
Department of Electrical & Electronic Engineering
Imperial College London
September 2016

Copyright declaration

The copyright of this thesis rests with the author and is made available under a *Creative Commons Attribution Non-Commercial No Derivatives* licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Statement of originality

I declare that this manuscript and the research referred to therein, are the product of my own work under the guidance of my thesis supervisor Dr. Patrick A. Naylor. Any ideas or quotations from the work of others, published or otherwise, are fully acknowledged in accordance with standard referencing practice. I confirm that the material of this thesis has not been accepted for any degree, and has not been concurrently submitted for the award of any other degree.

Abstract

The study of reverberation, broadly defined as the multipath propagation of sound in an enclosed space, has attracted significant interest from speech processing engineers and researchers alike. This is due in large part to the proliferation of speech processing technologies that facilitate distant talking speech input. In such scenarios, understanding the impact of reverberation on speech quality, and mitigating its detrimental impact through dereverberation techniques, are important and practically well motivated tasks. This research concerns both these topics. More specifically, in this work we (1) extend and develop an objective measure for predicting the level of perceived reverberation, (2) conduct an experimental investigation into reverberation perception and (3) propose the use of a spherical microphone array rake receiver to perform speech dereverberation.

In order to assess the level of perceived reverberation in speech, we develop the extended Reverberation Decay Tail (R_{DTX}) measure. Employing an improved perceptual model, the performance of the measure is first evaluated objectively. Later, we propose experimental methodologies and listening test schemes to collect subjective assessments. From the data obtained, the acoustic parameters most strongly correlated with reverberation perception are identified. The insights gained from the experimental investigation are used to further develop and validate the model of perceived reverberation incorporated in the R_{DTX} measure. The final contribution of this work, is the formulation of acoustic rake receivers in the spherical harmonic domain, which exploit signal reflections to perform speech dereverberation. Evaluating the proposed designs using widely adopted objective metrics, as well as the objective measures developed in this work, demonstrates the constructive use of early reflections can lead to substantial dereverberation.

Acknowledgments

This work would not have been possible without the support and guidance of numerous people. First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Patrick Naylor, for the opportunity to undertake this research and for his consistent support. Elements of this work are the direct result of collaborative efforts, to this end I would like to thank Dr. Alastair Moore for the chance to undertake work within the Embodied Audition for Robots (EARs) framework. Special thanks must also be extended to Benjamin Cauchi and Dr. Stefan Goetze for the opportunity to visit and conduct experimental work at Fraunhofer IDMT, Oldenburg. The experience was not only educational, but also extremely enjoyable. I am also thankful to my examiners Anthony Tew and Dr. Javier Barria for their feedback and suggestions, which helped strengthen the quality and scientific rigour of this manuscript. Finally, I would like to thank my family and in particular my parents, Khalid and Uzma Javed. For their support, encouragement and unconditional love, I am eternally grateful.

Contents

Copyright declaration	2
Statement of originality	3
Abstract	4
Acknowledgments	5
Contents	6
List of figures	9
List of tables	11
List of notation	12
Abbreviations	14
Chapter 1. Introduction	16
1.1 Context	16
1.2 Scope and original contributions	18
1.2.1 Research statement	18
1.2.2 Publications	18
1.2.3 Novel contributions	20
1.3 Thesis structure	21
Chapter 2. The Reverberation Decay Tail	23
2.1 Review	23
2.2 Reverberation Decay Tail measure	31
2.2.1 Implementation	36
2.3 Extensions and improvements	37
2.3.1 Perceptual model	37
2.3.2 Decay curve detection	41

2.4	Performance evaluation	42
2.4.1	R_{DTR} versus T_{60}	43
2.4.2	R_{DTR} versus C_{50}	47
2.5	Conclusion	50
Chapter 3.	Perceptual measurements	51
3.1	Review	52
3.1.1	Parametric models of reverberation	52
3.1.2	Perceptual evaluation of reverberation	53
3.2	Considered instrumental measures	55
3.2.1	Weighted Reverberation Decay Tail	55
3.3	Experimental setup	56
3.3.1	Test signals	56
3.3.2	Listening test	60
3.4	Results and analysis	63
3.4.1	Participant scores	63
3.4.2	Participant feedback	67
3.4.3	Evaluation of objective measures	67
3.4.4	Perceptual mapping	69
3.5	Conclusion	71
Chapter 4.	Spherical microphone array acoustic rake receivers	72
4.1	Review	73
4.2	Spherical harmonic domain processing	76
4.2.1	Sound field decomposition	77
4.2.2	Signal model	80
4.2.3	Spherical harmonic domain beamforming	80
4.3	Spherical acoustic rake designs	84
4.3.1	Delay-and-sum rake	85
4.3.2	Pre-echo null rake	87
4.3.3	Weighted pre-echo null rake	89
4.3.4	Rake finger selection	90
4.4	Experimental simulations	90
4.4.1	Setup	90
4.4.2	Performance measures	92
4.5	Evaluation	94
4.5.1	RIR-based performance evaluation	94
4.5.2	Signal-based performance evaluation	101
4.6	Conclusion	108

Chapter 5. Conclusions	110
5.1 Thesis contributions	110
5.2 Future work	112
Bibliography	114
Appendix A. Listening test instructions	130
Appendix B. Impact of the pre-echo null Rake weighting scheme	132
Appendix C. Modelling the Eigenmike	134

List of figures

2.1	Structure of an example room impulse response	24
2.2	Spectrogram of an anechoic and reverberant speech utterance	25
2.3	Block diagram of Bark spectrum transform	31
2.4	Bark spectrogram of an anechoic and reverberant speech utterance	32
2.5	Critical band filters used in R_{DT} calculation	33
2.6	End-point and flat region detection in an arbitrary selected Bark bin	34
2.7	Critical band filters used in R_{DTx} calculation	38
2.8	Impact of Bark spectral resolution on performance	39
2.9	Revised ISO 226:2003 standard equal-loudness curves	40
2.10	Improved flat region detection scheme	42
2.11	RIR generation from a prototype impulse response	44
2.12	R_{DTx} performance evaluation against T_{60}	45
2.13	Performance comparison of R_{DTx} and R_{DT}	46
2.14	R_{DTx} performance evaluation in common types of noise	47
2.15	R_{DTx} performance evaluation against C_{50}	49
3.1	Headphone calibration equipment	61
3.2	MUSHRAR listening test evaluation GUI	62
3.3	Listening test booth	63
3.4	A participant undertaking the listening test	64
3.5	Participant listening test times	65
3.6	Boxplot of mean listener scores for each test condition.	65
3.7	Wilcoxon rank-sum test for all test condition pairs	66
3.8	Correlation coefficients between subjective scores and objective measures	68
3.9	Perceptual mapping function for the wR_{DT} measure	70
4.1	Spherical coordinate system	77
4.2	Magnitude of the spherical harmonics up to third order	79
4.3	PWD beampattern as a function of beamformer order	82

4.4	Three-dimensional directivity pattern of a PWD beamformer	83
4.5	Example directivity pattern of a null steering beamformer	84
4.6	Rake-DS conceptual operation	86
4.7	Rake-DS concept input and output structure	89
4.8	Multipath DOAs for source-array setup in room A	91
4.9	Rake and baseline systems response to a Dirac delta	95
4.10	Omnidirectional response and delay-and-sum rake output structure	97
4.11	Null steering beamformer energy variation	98
4.12	DRR performance of path selective raking.	99
4.13	DRR and DPER performance as a function of the number of rake fingers .	100
4.14	$\Delta w_{\tilde{R}_{DT}}$ scores against the number of reflections considered	102
4.15	ΔWER scores against the number of reflections considered	103
4.16	$\Delta PESQ$ scores against the number of reflections considered	105
4.17	Computational complexity of different designs under consideration	106
B.1	DRR and DPER performance of wRake-PN for different weightings	133

List of tables

2.1	End-point energy and flat region ripple as a function of block size	41
2.2	R_{DTX} performance in presence of common noise types (SNR 20 dB)	48
2.3	R_{DTX} performance in presence of common noise types (SNR 10 dB)	48
2.4	R_{DTX} performance in presence of common noise types (SNR 0 dB)	48
3.1	Selected RIR characteristics	57
4.1	DRR and DPER performance of rake designs in room A	94
4.2	DRR improvements for systems under test in room B	101
4.3	DPER improvements for systems under test in room B	101
C.1	Eigenmike sensor placements and quadrature weights	135

List of notation

General notation

x	Scalar quantity
$\mathbf{x}$	Vector quantity
$\mathbf{X}$	Matrix quantity

$x(n, \mathbf{r})$	Time domain representation of a signal
$X(k, \mathbf{r})$	Frequency domain representation of a signal
$X_B(k_b)$	Bark-frequency domain representation of a signal
X_{lm}	Eigenbeam of order l and degree m

Operators

$x * y$	Linear convolution operator
$[\cdot]^*$	Complex conjugate
$[\cdot]^T$	Vector/matrix transpose
$[\cdot]^{-1}$	Matrix inverse
$[\cdot]^H$	Hermitian transpose
$ \cdot $	Absolute value

Symbols and variables

b_l	Mode strength of order l
B	Bark

c	Speed of sound
C_{50}	Clarity index
δ	Kronecker delta
f	Frequency
h	Room impulse response
i	Imaginary number ($i = \sqrt{-1}$)
k	Wavenumber
k_b	Bark bin
l	Order
L	Array order (maximum spherical harmonic order)
m	Degree
n	Discrete time index
N	Number of listening test trials
$\mathcal{P}_{lm}$	Associated Legendre function of order l and degree m
$\Omega = (\theta, \phi)$	Spherical coordinates (inclination θ , azimuth ϕ)
q	Microphone index
Q	Number of microphones
$\mathcal{Q}$	QAreverb measure
$\mathbf{r} = (r, \Omega)$	Position on a sphere in spherical coordinates (radial r , angular Ω)
s	Desired anechoic speech signal
σ_I^2	Room spectral variance
T_{60}	Reverberation time
v	Ambient noise
Y_{lm}	Spherical harmonic of order l and degree m
Z	Beamformer output

Abbreviations

ASR automatic speech recognition

BSD Bark spectral distortion

DFT Discrete Fourier transform

DNN deep neural network

DOA direction of arrival

DRR direct-to-reverberant ratio

DPER direct-to-pre-echo ratio

ISFT inverse spherical Fourier transform

ITU International Telecommunication Union

MCLP multichannel linear prediction

MOS mean opinion score

MUSHRA multi stimuli test with hidden reference and anchor

MUSHRAR multi stimuli test with hidden reference and anchor for reverberant speech

MINT multiple-input/output inverse theorem

PCEPLR pairwise comparison for the evaluation of the perceived level of reverberation

PEAQ perceptual evaluation of audio quality

PESQ perceptual evaluation of speech quality

PSQM perceptual speech quality measure

PWD planewave decomposition

RIR room impulse response

RSV room spectral variance

SFT spherical Fourier transform

SHD spherical harmonic domain

SMA spherical microphone array

SNIR signal-to-noise-plus-interference ratio

SNR signal-to-noise ratio

SPL sound pressure level

SRMR speech to reverberation modulation energy ratio

STFT short-time Fourier transform

TOA time of arrival

WER word error rate

Chapter 1

Introduction

1.1 Context

The idea of talking to the machines around us, just as we would to a person, may have seemed far-fetched as little as 10 years ago. Yet due to advances in speech signal processing, as well as the availability of more powerful microprocessors, an unprecedented number of computerised devices now possess speech interfaces. Whether it is to capture a speech signal and reproduce it for a human listener in a different location, or to employ speech recognisers that allow users to *speak* with a machine, the incorporation of such speech communication systems in the technologies we use everyday offers promising and exciting applications.

However, despite recent advances, numerous practical challenges remain before such systems can offer reliable performance in real-life settings. The acquisition or capture of a speech signal from a distant talker is a particularly difficult problem. In such scenarios, the signal of interest will often suffer from extensive degradation due to the acoustic environment it propagates in. Factors such as background noise, interference from competing sound sources and reverberation all contribute to deteriorating signal quality. The latter is especially problematic for many speech processing applications. Broadly defined as the multipath propagation of sound in an enclosure, reverberation results in delayed and attenuated copies of a desired signal also being captured by a microphone [1]. Whilst

low levels of reverberation can be desirable, higher levels have been shown to decrease the accuracy of speech recognisers [2], the robustness of localisation algorithms [3], the performance of hearing aid devices [4, 5], as well as the perceived quality and, in extreme cases, even the intelligibility of speech [6].

Speech dereverberation, which aims to remove or at least mitigate the impact of reverberation, is thus a practically well motivated task that has received significant interest from the research community [7]. Of the numerous dereverberation techniques that have been proposed to date, spatial filtering or beamforming-based approaches are perhaps the most practical, given the performance achievable with limited operating assumptions. By utilising the spatial diversity provided by microphone arrays, beamformers can selectively amplify sound from a given direction [8]. Thus by estimating the direction of arrival (DOA) of the desired source, the signal of interest can be amplified relative to the reverberant sound field, thereby achieving substantial dereverberation [9].

However, most conventional beamformers implicitly ignore the contributions of strong early reflections. It has long been known in the field of wireless communications that the multipath propagation of a signal can be exploited for improved signal acquisition, a popular example being the rake receiver design proposed by Price nearly six decades ago [10]. Although speech signal processing techniques are often adapted from those used in wireless communications (owing to the similarity of the problems encountered in both fields), efforts to develop an acoustic equivalent to the rake receiver have been limited. Having said that, a recent theoretical study highlighted the utility of such a technique in the context of signal acquisition in noisy environments [11]. The use of an acoustic rake receiver has, to date, not been investigated for the purposes of performing dereverberation.

Whilst beamformer designs can be implemented on any arbitrary microphone array, the use of spherical microphone arrays for signal enhancement applications has steadily grown [12]. This interest has coincided in part with the commercial availability of spherical arrays, such as the ‘Eigenmike’ from mh acoustics [13]. Moreover, as a consequence of their geometry, spherical arrays possess several intrinsic advantages over the more widely used two-dimensional linear and planar array setups. Namely, they allow a sound field to be

analysed in three-dimensions and offer equal performance in all directions, both of which are useful qualities in the context of beamformer design. Additionally, the captured sound field can be efficiently described in the spherical harmonic domain, which represents an elegant mathematical framework from which to operate [14].

Despite the increased interest in dereverberation research over the last few years, objective measures capable of estimating perceived levels of reverberation, on the other hand, have attracted far less attention. Furthermore, of the existing measures currently used, few have had their reliability as predictors of perceived reverberation experimentally validated. Yet such measures are needed to test the effectiveness of developed dereverberation algorithms quickly and consistently, avoiding the need for time-consuming, labour-intensive and expensive listening tests. Additionally, the knowledge provided by such measures could be utilised to better perform speech processing such as dereverberation, as well as provide insights into the properties of the acoustic space in which the speech signals propagate. It is in this context the contributions of this thesis are made.

1.2 Scope and original contributions

1.2.1 Research statement

The aim of this research is to develop an experimentally validated objective measure capable of predicting the perceived level of reverberation in speech, and to propose a spherical microphone array acoustic rake receiver that exploits energetic signal reflections to perform dereverberation.

1.2.2 Publications

The following publications were produced during the course of this work:

Journal paper

- [J1] H. A. Javed, A. H. Moore and P. A. Naylor, “A method for dereverberation using spherical microphone array acoustic rake receivers,” to be submitted to *IEEE Trans.*

Audio, Speech, Lang. Process.

Conference papers

- [C1] H. A. Javed and P. A. Naylor, “Evaluation of an extended reverberation decay tail metric as a measure of perceived reverberation,” in *Proc. Audio Eng. Soc. (AES) Conf. on Audio Forensics*, London, UK, June 2014.
- [C2] H. A. Javed and P. A. Naylor, “An extended reverberation decay tail metric as a measure of perceived late reverberation,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, Sept. 2015.
- [C3] 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.
- [C4] B. Cauchi, H. A. Javed, T. Gerkmann, S. Doclo, S. Goetze and P. A. Naylor, “Perceptual and instrumental evaluation of the perceived level of reverberation,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.
- [C5] H. A. Javed, A. H. Moore and P. A. Naylor, “Spherical harmonic rake receivers for dereverberation,” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi’an, China, Sept. 2016.
- [C6] H. A. Javed, B. Cauchi, S. Doclo, P. A. Naylor and S. Goetze, “Measuring, modelling and predicting perceived reverberation,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2016.

Given the collaborative nature of the work, the contributions contained in [C4] are not claimed as solely my own. Instead, the results and analysis produced are reviewed, and explicitly referenced when referred to in the text.

1.2.3 Novel contributions

To the best of the author's knowledge, the following aspects of this thesis are original and novel contributions:

- Development of an extended Reverberation Decay Tail (R_{DTX}) measure as a means of predicting the perceived level of reverberation in speech (*presented in Chapter 2, published in [C1, C2]*).
 - The Reverberation Decay Tail (R_{DT}) of [15] is extended from narrowband to wideband operation, incorporates a more realistic perceptual model and an improved decay curve detection scheme.
 - Performance of the extended measure is experimentally assessed on a large set of reverberant speech using objective tests, and the robustness to common types of noise investigated.
- An experimental investigation into reverberation perception in speech captured by a distant microphone, in low to moderately high reverberation conditions, is conducted to further develop and validate the practicality of the proposed R_{DTX} measure (*presented in Chapter 3, published in [C6]*).
 - Listening tests are conducted to determine a human assessor's ability to discern different levels of reverberation in speech, and to identify the acoustic parameters that most influence perception.
 - Reverberation models and measures are evaluated on the test data obtained, and their level of correlation with subjective assessments determined.
 - The development and validation of a weighting scheme that is integrated into the computation of the R_{DT} and R_{DTX} measures, to improve their correlation with subjective assessments.
 - Using the data obtained, a non-linear mapping function is proposed, which maps the output of the weighted R_{DT} to a perceptual score.

- Development of several spherical microphone array rake receiver designs for dereverberation applications (*presented in Chapter 4, published in [C3, C5]*).
 - Formulation of three acoustic rake receiver designs in the spherical harmonic domain, which capture and combine early reflections for improved signal acquisition in reverberant environments.
 - The incorporation of spatial nulls in two of the proposed designs, which are shown to suppress pre-echoes that may otherwise be introduced.
 - Investigation into the impact the choice of reflections to combine has on performance, with a criterion based on reflection energies shown to be most effective.
 - Experimental simulations conducted to illustrate the effectiveness of the proposed designs compared to existing methods, using a range of evaluation metrics.

1.3 Thesis structure

Having provided the context and motivation for this research, as well as summarising the novel contributions made, the remainder of this thesis is organised as follows.

- In **Chapter 2**, the R_{DTX} metric is proposed as a means of predicting the perceived level of reverberation in captured speech. The chapter begins by providing an overview of existing methods used to evaluate reverberant speech, and identifies the need for a measure which correlates with human perception. The R_{DT} of [15] is then reviewed, and extensions and improvements to the method are proposed to enhance its performance as a predictor of perceived reverberation. Finally, the remainder of the chapter presents the results of experimental tests conducted to evaluate the R_{DTX} and validate the extensions proposed.
- In **Chapter 3**, the design of an experimental listening test to assess perceived levels of reverberation in speech is described. The subjective assessments gathered are first analysed to determine a human assessor's ability to perceive different levels

of reverberation, before proceeding to identify the acoustic parameters that most strongly influence perception. These findings are then used to develop a weighting scheme that is incorporated into the R_{DT} and R_{DTx} measures. Finally, the listening test data is used to validate the performance of the R_{DT} variants proposed, and other existing measures, at predicting the perceived level of reverberation.

- In **Chapter 4**, an overview of existing dereverberation techniques is provided and the concept of an acoustic rake receiver is explored. After identifying the potential utility of such a design for dereverberation applications, several rake receiver designs are proposed and formulated in the spherical harmonic domain. The relative merits of the designs are then evaluated through experimental simulations, using a range of performance metrics. The results demonstrate significant levels of dereverberation can be achieved by the designs, when compared to existing beamforming methods.
- The thesis is concluded, and areas for further research are suggested, in **Chapter 5**.

Chapter 2

The Reverberation Decay Tail

2.1 Review

In many real-life settings, speech and audio signals captured by a microphone often have their characteristics altered due to reverberation effects. Reverberation is broadly defined as the multipath propagation of sound in an enclosed space; it is the process in which signal reflections off walls and objects can result in the reception of thousands of delayed and attenuated copies of a signal arriving at a microphone. The quality and intelligibility of a captured speech signal is often degraded as a result, with detrimental effects becoming increasingly pronounced as the distance between speaker and microphone increases [1].

Reverberant speech can be modelled as the convolution between an acoustic channel and the speech signal in question. The channel, or room impulse response (RIR), thus characterises the effect of the room on the sound produced, for a given source-listener or source-microphone placement. For the study of perceptual reverberation, the perceived effects of reverberation can be classified into two distinct phenomena [1], which arise from the structure of a RIR. The first effect, known as colouration, is due to early reflections that are approximately defined as the multipath components of a RIR which arrive within 50 - 80 ms of the direct path. Early reflections of an example RIR are shown in blue in Figure 2.1. Due to the small time difference of arrival between these initial reflections and the direct path, the human auditory system is unable to distinguish them as distinct

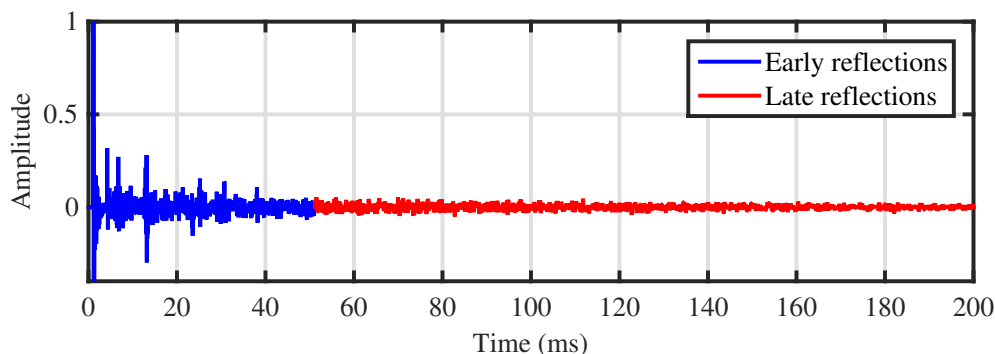


Figure 2.1: Structure of an example room impulse response, between a sound source and microphone, with early and late reflections highlighted.

signals, but rather perceives them as a single sound. Consequently, the result of early reflections is to colour or modify the spectral characteristics of the signal, so that the resulting sound is often described using subjective terms such as ‘thin’, ‘boxy’ and so on.

Late reflections, or the reverberant decay tail of the impulse response, on the other hand, lead to the perception of space and distance. These weaker but more densely packed, noise-like reflections, shown in red in Figure 2.1, result in the perception of the signal ringing on. Such effects are usually associated with highly reverberant enclosures. In contrast to early reflections, which have been shown to have potential to increase speech intelligibility [16], late reverberation is generally responsible for reducing the overall quality and intelligibility of speech, particularly in combination with noise [17] [18].

The physical effects of reverberation therefore manifest themselves both spectrally and temporally on a captured audio signal. This is illustrated by the spectrograms of a non-reverberant speech utterance and the same utterance convolved with an impulse response from a reverberant room, in Figure 2.2. The spectrogram is a visual representation of the short-time Fourier transform (STFT) of a signal, and thus conveys the time-frequency behaviour of the signal in question. We can see that the most notable effect of reverberation is the smearing of speech energy over time. This results in silent intervals, either in the form of pauses between words, or vocal tract closures during the production of consonants, being filled by energy from reflections (for example between 1.3 - 1.4 s). Similarly, reverberation causes noise-like sounds associated with fricatives, such as the voiced sibilance of the /z/ in

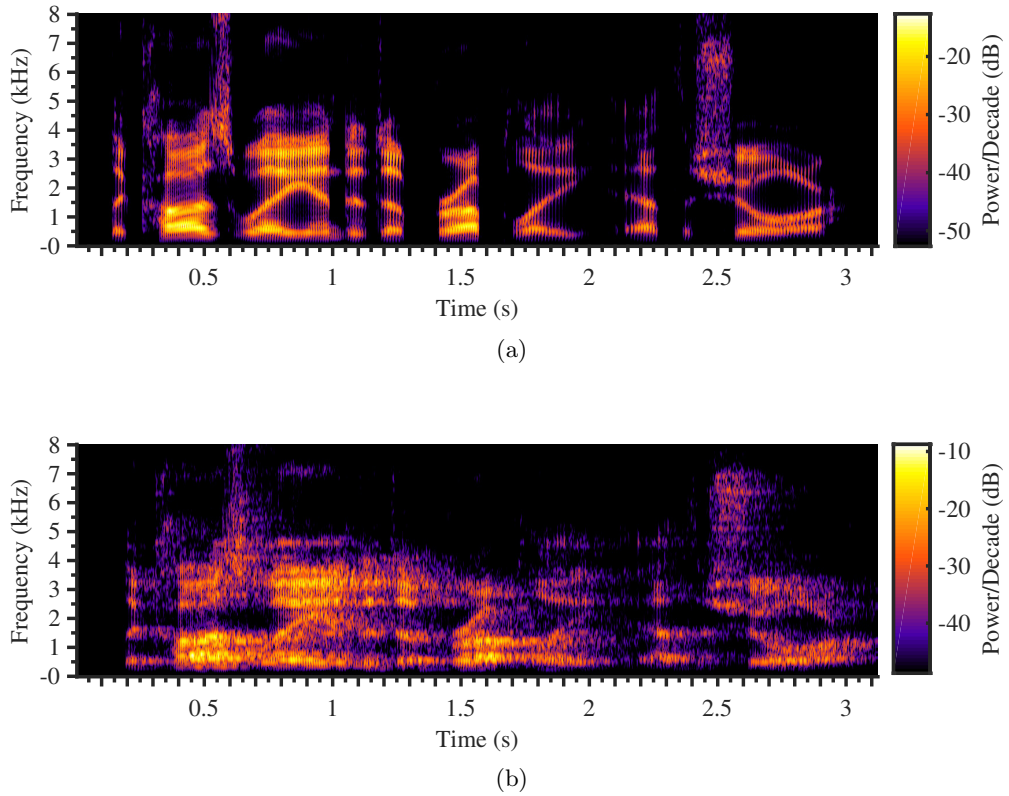


Figure 2.2: The effect of reverberation on the time-frequency characteristics of a speech signal. The spectrogram of (a) the speech utterance ‘*The causeway ended abruptly at the shore*’ spoken by a male talker in an anechoic environment, (b) the same utterance convolved with a RIR with a reverberation time of 1 second. Dark and light colours indicate regions of low and high energies, respectively.

‘causeway’ at 0.6 s, to persist for longer. Rapid frequency changes associated with formant transitions are less distinct and appear flatter in reverberant conditions (1.7 - 2.0 s). Formants correspond to resonances in the vocal tract, and are apparent in voiced vowels, whilst formant transitions represent an important cue in the perceived articulation of an utterance. Considering these effects, which are addressed in more detail in [19], it becomes clear how the presence of reverberation can have a profoundly detrimental impact on speech perception.

Yet despite the impact of reverberation on the time-frequency characteristics of uttered speech, as outlined above, human listeners are remarkably tolerant to its degradative effects. The contrast is particularly stark when compared to the performance of

machines at equivalent speech processing tasks. A poignant example is a human’s ability to recognise speech in reverberant environments, compared to that of an automatic speech recognition (ASR) system. The significant drops in performance suffered by the latter in even moderate levels of reverberation are well documented [2], whilst for a human listener situated in the same environment and with the same levels of reverberation, recognition rates do not significantly decrease below those associated with recognition of anechoic speech [20]. Similarly, for the task of auditory scene analysis, the aim of which is to separate individual audio signals from a mixture, human listeners have shown robust performance compared to computational methods in equivalent levels of reverberation [21]. This is despite the fact that many features thought to be used by humans to discriminate between sound signals from a mixture, such as spatial information or smooth frequency transition cues, are made less distinct by reverberation.

It has been suggested that the ability of human listeners to better tolerate the effects of reverberation, compared to existing speech processing technologies, can be attributed to several auditory mechanisms. One of which is the hearing system’s ability to exploit information on the temporal modulation characteristics of speech [22, 23]. Temporal modulations provide information on the time-varying harmonic structure of the spectral envelopes of a speech signal. It has been observed that temporal modulations occur in speech between 1 and 16 Hz, with a significant component at 4 Hz, which roughly corresponds to the rate at which syllables are spoken [24]. The role of slower temporal modulations in speech perception has been experimentally demonstrated by Drullman in [23], in which it was shown that speech intelligibility was not significantly reduced if envelope modulations between 4 and 16 Hz were preserved. Whilst fast temporal fluctuations in the spectral envelopes of speech are smoothed out by reverberation, slower modulations are comparatively robust to its effects. This indicates that the latter can serve as an important cue for human listeners to combat the effects of reverberation on speech. As we shall see later, numerous objective reverberation and speech quality measures try to take advantage of this fact by working in the modulation domain [24–27].

The binaural nature of the human auditory system (that is, the fact that we use two

ears to hear) has also been shown to play an important role in our ability to mitigate the detrimental impact of reverberation. As far back as the 1950s, Koenig [28] presented compelling evidence to support this assessment. In his work, Koenig compared the perception of room reverberation apparent in the playback of audio recordings captured by a single microphone with its perception in sounds captured by a pair of spatially separated microphones. It was noted that in the latter, the perceived impact of reverberation was greatly reduced. Furthermore, in the work of Moncur and Dirks [29], a perceptual listening test experiment was conducted in which forty-eight participants were asked to evaluate both monaural and binaural recordings of reverberant speech signals. The results from this experiment provided additional evidence highlighting the importance of binaural listening in combatting the effects of reverberation.

Numerous explanations regarding the underlying binaural mechanisms responsible for these results have been put forward by researchers. This includes the argument that the auditory system treats reverberation as uncorrelated noise and performs appropriate noise cancellation using binaural processing [30]. However, in the same work it was shown that the effect of reverberation was less prominent on binaural recordings of real reverberant speech compared to artificially generated reverberant speech files in which phase was randomised, thereby removing spatial cues from the signals. This strongly indicates that reverberation is not just reduced, but rather that localisation information provided by binaural listening is likely actively exploited to mitigate its effect.

In addition to some of the factors outlined above, there is significant evidence to indicate the perceived effects of reverberation are also closely tied to both the characteristics of the source signal itself, as well as the environment it is produced in. Regarding the former, Paulus showed that in equal levels of reverberation, orchestra music was perceived to be less reverberant than speech signals [31]. This is likely due to the time-frequency make up of musical signals compared to speech, but may also indicate that our perception is linked to an internal reference built on past experiences (we expect orchestra music to sound reverberant). Regarding the latter, Watkins has presented compelling evidence to suggest that listeners will use the contextual information that precedes a reverberated test

word to help them identify it [32, 33]. His much cited experiments focused on one particular speech identification task in which listeners attempted to identify the test words ‘sir’ and ‘stir’. Embedded within a phrase, the test words were drawn from a synthetic continuum that was generated from token utterances of the two words. By testing recognition in varying levels of reverberation, Watkins concluded that listeners use information about the temporal envelope of preceding speech to compensate for the effects of reverberation on a particular word. Further evidence to support this conclusion went on to be presented by Beeston and Brown [34–36], who built on Watkins’ work by performing more extensive experimental tests. These findings strongly suggest that our ability to mitigate and perceive reverberation may be both context dependent and adaptive.

The myriad of factors that play a role in how we perceive reverberation, as outlined above, serve to highlight just some of the complexities involved in attempting to measure the perceptual impact of reverberation. It is therefore worth stating at this stage the scope and intended applications of the perceptual measures we propose for this task. Namely, this research is interested in developing an objective measure capable of estimating the perceived level of late reverberation in speech captured by a single microphone. Specifically we are concerned here with its impact on quality, in the form of its perceived presence in the recorded signal, rather than with its impact on intelligibility. As noted above, the effects of reverberation on speech captured using a single microphone are far more pronounced compared with the same signal heard live by a human listener. Given the increasing number of devices that accept speech input spoken from a distance, such a measure is therefore practically well motivated, particularly in speech communication applications such as teleconferencing.

As the perceptual effects caused by late reflections are distinct from other types of signal degradations, it is worth developing measures capable of assessing them specifically. Whilst numerous speech quality metrics have been proposed over the years, even popular and widely adopted measures such as the Bark spectral distortion (BSD) [37], perceptual evaluation of speech quality (PESQ) [38] and Perceptual Objective Listening Quality Assessment (POLQA) [39], are limited to giving indications correlated with overall speech

quality. Originally designed to measure the quality of network transmitted speech, these metrics have been shown to be average to poor predictors of the level of perceived reverberation in speech recordings [40]. Measures capable of assessing the perceptual impact of reverberation specifically, in contrast to general quality metrics on the other hand, are fewer in number, and the use of no one metric is agreed upon by the research community.

Measures that predict the perceived level of reverberation can be broadly categorised as those that require knowledge of the RIR, or those that operate on the reverberant signals. The former extract parameters from the impulse response which mathematically describe its structure. These well known acoustic metrics are primarily based on the rate of energy decay, or on energy ratios of different portions of the RIR. The reverberation time (T_{60}), which is defined as the time it takes for a sound to decay by 60 dB after the sound source has been abruptly terminated, is an example of the former. The clarity-index (C_{50}) or the direct-to-reverberant ratio (DRR) are examples of the latter. The C_{50} metric is defined as the ratio of the early reflection energies to the energy of the rest of the response [41], whilst the DRR metric is the ratio of the direct path energy to the energy of the rest of the response. As the level of perceived reverberation is strongly influenced by the RIR envelope profile, numerous reverberation measures proposed in the literature have attempted to use RIR-based metrics to estimate the perceptual impact of reverberation [31, 42–44]. Whilst such measures, which will be reviewed in more detail in Chapter 3, can give better correlations with perceived reverberation, they are somewhat impractical. This is because knowledge of the RIR (or how it is modified by a dereverberation algorithm) may not always be known or may be difficult to estimate.

On the other hand, signal-based reverberation measures that require only the reverberant, and in some cases reference, signals are potentially more useful. The speech to reverberation modulation energy ratio (SRMR) proposed by Falk [26], is an example of a reverberation measure that does not require knowledge of the RIR. The computation of the SRMR relies on decomposing the input signal using gamma-tone filter banks from which the temporal envelope is extracted. The SRMR value is then computed as the ratio between the energy in the higher and lower modulation frequencies, under the premise that

speech energy is primarily concentrated in the latter, whilst the presence of reverberation introduces higher frequency envelope modulations. However, it was recently observed that the measure suffered from high variability linked to pitch and speech content, limiting its value as a measure of the perceived level of reverberation. In an attempt to avoid these detrimental effects, the SRMR was extended in [27]. By using a narrower range of modulation frequency and by introducing upper and lower energy bound values in each band, the extended measure (denoted by SRMRx in this thesis), limits the influence of pitch and speech content. Whilst the measure has been shown to give improved performance and lower variability relative to the original SRMR [27], further tests are needed to evaluate its performance as a predictor of perceived reverberation.

The Reverberation Decay Tail (R_{DT}), proposed in [15], was specifically designed to evaluate the perceived level of reverberation. An intrusive measure, it uses a reference signal to identify time periods which would be most affected by reverberation, and characterises its effect in the corresponding sections of the reverberant test signal. The measure operates in the perceptually motivated Bark spectral domain, rather than the STFT domain. The former models the peripheral auditory system by generating an indication of perceived loudness from a weighted frequency spectral analysis [37]. It was therefore felt to be better suited to assessing the perceptual impact of late reverberation, as from quality assessment in the field of audio coding, it is known that measures based on more exact models of the human auditory system show higher correlation with subjective data [45]. Furthermore, loudness models have been employed with reasonable success, to detect late reverberation in other works, for example to predict the level of running reverberation in musical signals [46].

However, despite the fact that the R_{DT} was designed specifically to evaluate reverberant speech, validation of the measure has been limited, with experiments usually involving simulated reverberant signals [15]. Moreover, in recent practical studies, it has been shown in some cases to correlate poorly with subjective assessments [47]. Thus the heuristically developed measure may require further improvements and experimental validation, which will be the focus of the remainder of this chapter.

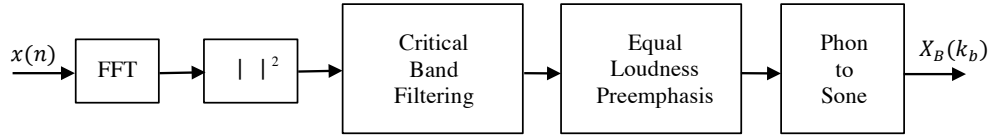


Figure 2.3: Block diagram showing the stages involved in computing the Bark spectrum, $X_B(k_b)$, of an input signal $x(n)$.

2.2 Reverberation Decay Tail measure

The R_{DT} measure operates under the premise that a RIR can be modelled as exponentially decaying, zero-mean, white Gaussian noise [48]. Under this model, the intrusive measure attempts to identify and characterise the energy decays introduced by reverberation, directly from the Bark spectral representations of the reverberant and corresponding reference signal. Computing the measure involves three main steps, namely (i) Bark spectrum calculation; (ii) ‘end-point’ and ‘flat region’ detection; and (iii) parametric decay curve fitting.

The first stage in the calculation of R_{DT} involves partitioning the test and reference signal into non-overlapping frames, and computing the Bark spectra of these frames. Like the STFT, this allows the time-frequency characteristics of the signals to be studied. The Bark spectrum domain is essentially a weighted frequency spectrum that indicates perceived loudness. Working in this domain to detect late reverberation is perceptually well motivated, as the sensation of loudness is a good indicator of the presence of a particular type of distortion [37]. The Bark spectrograms of an anechoic and reverberant speech signal are illustrated in Figure 2.4. As the same speech utterance has been used, the plots can be directly compared and contrasted against the STFT spectrograms in Figure 2.2.

The calculation of the Bark spectrum attempts to model the peripheral auditory system and how it processes incoming sound. In the R_{DT} implementation of [15], the Bark spectra of a clean signal and its reverberant counterpart are calculated using the

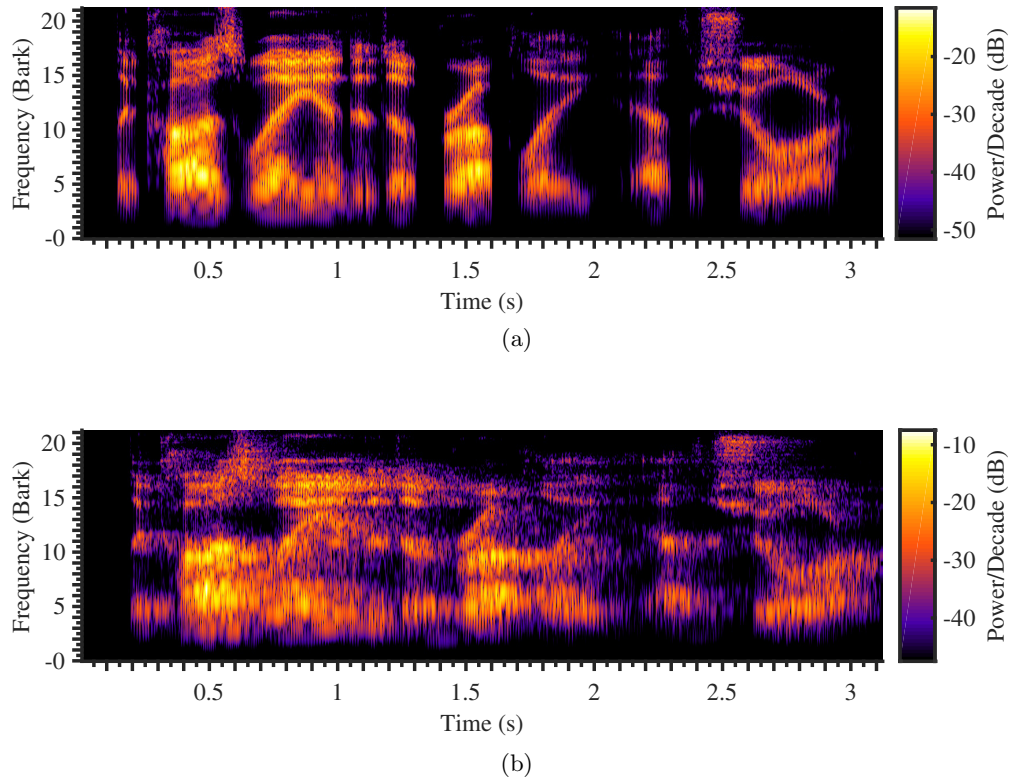


Figure 2.4: The effect of reverberation on the time-Bark frequency characteristics of a speech signal. The Bark spectrogram of (a) the speech utterance ‘*The causeway ended abruptly at the shore*’ spoken by a male talker in an anechoic environment, (b) the same utterance convolved with a RIR with a reverberation time of 1 second. Dark and light colours indicate regions of low and high energies, respectively.

approach described in [37]. The auditory model employed is shown in Figure 2.3, which highlights the main stages of processing undertaken. This involves modelling the frequency resolution of the auditory system via a bank of critical band filters of increasing bandwidth (Figure 2.5). In this way, the role of the cochlea in discriminating frequencies is emulated. Equal loudness curves are then used to model the fact that the ear is not equally sensitive to sounds of different intensities and frequencies. The final stage of processing maps loudness levels measured in phons to sone - a linear loudness scale. The definition of a phon is such that a sound described as being 60 phons loud is perceived to have the same loudness as a 60 dB sound pressure level (SPL) 1 kHz tone, where dB SPL refers to sound pressure relative to the hearing threshold. However, as a general rule of thumb, the power

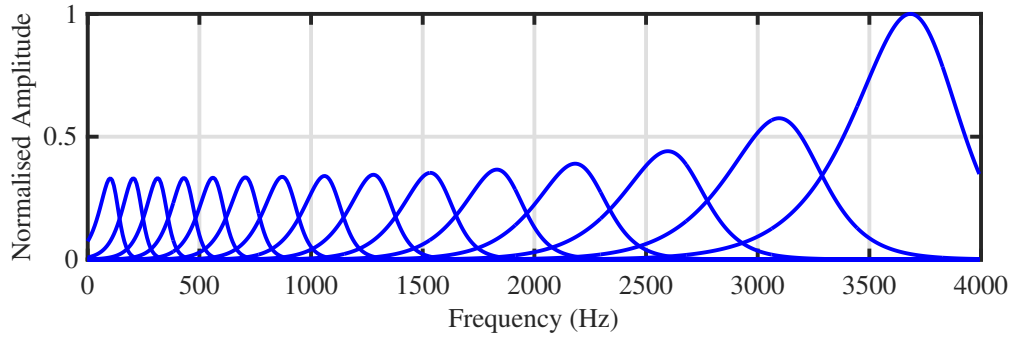


Figure 2.5: Critical band filters, weighted by equal loudness models, used in the calculation of signal Bark spectra by R_{DT} (for visual purposes, only a limited number of filters are displayed).

or intensity of a sound must be increased by 10 times for it to be perceived as sounding twice as loud. Rather than working in the logarithmic phon scale a more intuitive linear loudness scale, measured in units of sones, was introduced. In this scale, 40 phons is defined as 1 sone, 50 phons as 2 sones, 60 phons as 4 sones and so forth. Similarly, 0.5 sones corresponds to 30 phons, 0.25 sones to 20 phons and so on. The upper limit of the scale is 64 sones, which corresponds to 100 phons [49].

End-points and flat regions, defined respectively as points where speech energy drops abruptly and remains approximately flat across several frames, are then identified in the Bark spectrum of the reference signal. This step is motivated by the fact that the effect of the reverberant tail on a speech signal would be most distinctly visible in these regions. As an illustrative example, Figure 2.6a shows the detection of three flat-regions in the Bark spectrum of a reference speech signal. The regions are determined from the energy envelope of the reference signal by gradient and absolute thresholds. These thresholds are defined as percentage values of the global peak in the energy envelope of that bin and were chosen through experimental investigation in [15]. It is worth noting that in this thesis the calculation of the R_{DT} , or any variant of the measure, will operate using these values (discussed in more detail in section 2.2.1).

In the time-aligned and energy-normalised reverberant Bark spectra, exemplified in Figure 2.6b, these corresponding regions exhibit the effect of reverberation; the spreading

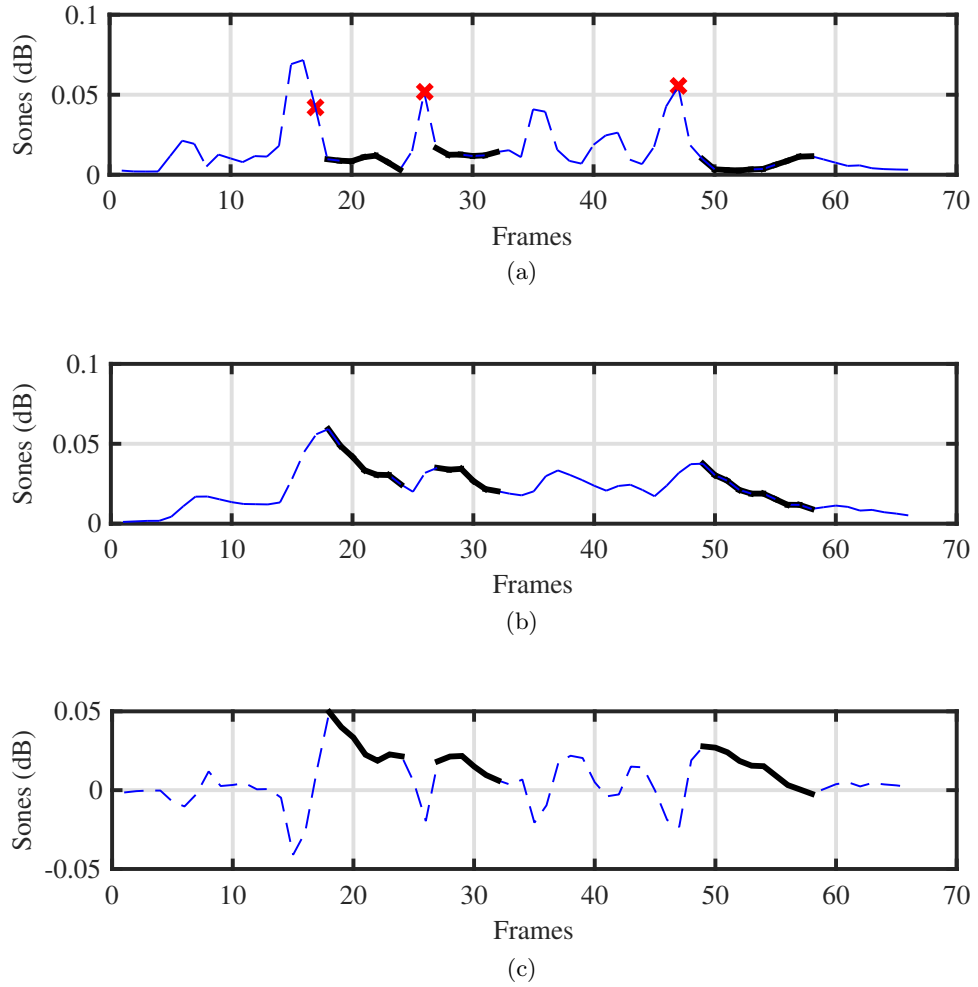


Figure 2.6: End-point and flat region detection in an arbitrarily selected Bark bin of (a) reference speech signal, (b) the reverberant test signal and (c) the Bark spectral difference. Dashed lines indicate loudness, solid black lines indicate detected flat regions and red crosses indicate detected end-points.

of energy across frames. The final step in the calculation of R_{DT} involves characterising the energy decays in these regions in the Bark spectral difference of the reverberant and clean signal, shown in Figure 2.6c. The difference is used rather than the reverberant spectrum, so as to discount any natural decays present in the reference speech signal.

Finally, the identified energy decays in each Bark bin difference are modelled as negative exponentials, that is $Ae^{-\lambda\nu}$, where A and λ denote the energy decay level and rate respectively, and ν denotes the frame number. Using a sum of squares error minimi-

sation criterion, curves are fitted to each decay, and parameters of the exponential model estimated. The reverberant energies are then averaged across the number of detected curves, M_{k_b} , in each Bark bin, k_b , expressed mathematically as

$$\bar{A}_{k_b} = \frac{\sum_{p=1}^{M_{k_b}} A_{k_b}(p)}{M_{k_b}} . \quad (2.1)$$

A weighted average of the reverberation rate of decay, $\bar{\lambda}_{k_b}$, is also computed in each Bark bin,

$$\bar{\lambda}_{k_b} = \frac{\sum_{p=1}^{M_{k_b}} A_{k_b}(p) \lambda_{k_b}(p)}{\bar{A}_{k_b} M_{k_b}} . \quad (2.2)$$

The weighting is introduced to improve the robustness to erroneous estimates that may come about from identified decays with low energy. In addition to computing the aforementioned exponential parameters, the energy of the reference speech signal preceding each flat region, that is the energy of the end-point, is also recorded and averaged. The average direct path energy, $\bar{D}_{k_b}$, is written

$$\bar{D}_{k_b} = \frac{\sum_{p=1}^{M_{k_b}} D_{k_b}(p)}{M_{k_b}} . \quad (2.3)$$

Each parameter is then averaged across the total number of Bark bins, K_b . By characterising the identified reverberant energy decays with the above parameters, the R_{DT} measure is simply defined as the ratio of the average reverberant energy to the average rate of decay, normalised to the average direct path component such that

$$R_{DT} = \frac{K_b \sum_{k_b=1}^{K_b} \bar{A}_{k_b}}{\sum_{k_b=1}^{K_b} \bar{\lambda}_{k_b} \sum_{k_b=1}^{K_b} \bar{D}_{k_b}} . \quad (2.4)$$

The intuitive measure indicates that for large average reverberation tail energies, relative to the direct sound energy, the perceived level of reverberation will be greater. Similarly, the effects of reverberation will be more noticeable for slower average decay rates.

2.2.1 Implementation

The search algorithm and operational parameters used in [15] to identify valid end-points and flat regions is summarised in this section. We first define the function $\Delta\chi_B$, which denotes the difference between the Bark spectrum magnitude of a signal at time frame index ν and $\nu + \epsilon$. Namely,

$$\Delta\chi_B(\nu, \epsilon, k_b) = X_B(\nu, k_b) - X_B(\nu + \epsilon, k_b), \quad (2.5)$$

where ϵ is a positive integer and $X_B(\nu, k_b)$ is the short-time Bark spectrum of the discrete time signal $x(n)$. Non-overlapping frames of 32 ms, using a Hamming window, are employed in the calculation of the short-time Bark spectrum.

The instant of an end-point, ν_0 , is defined as a frame in Bark bin k_b that directly precedes a large drop in energy. The drop is defined as a percentage of the maximum energy in that Bark bin. A frame is therefore defined as an end-point if it satisfies the following inequality,

$$\Delta\chi_B(\nu_0, 1, k_b) > \delta_1 \cdot \max\{X_B(k_b)\}, \quad (2.6)$$

where δ_1 is set at 0.2, that is a threshold of 20%. It is worth noting that end-points can extend beyond the initial frame, thus for each end-point the number of frames, I , is counted for which

$$\Delta\chi_B(\nu_0 + \iota, 1, k_b) > \delta_2 \cdot \max\{X_B(k_b)\} \quad \iota = 0, 1, \dots, I, \quad (2.7)$$

where δ_2 is set at 0.1, that is a threshold of 10%.

A flat region must immediately follow an end-point, and is defined in a Bark bin as a range of frames in which the sone value varies less than δ_{max} whilst also being lower than an absolute threshold δ_t , for a minimum and maximum duration of J_{min} and J_{max} frames respectively. Namely,

$$\Delta\chi_B(\nu_0 + I, j, k_b) < \delta_{max} \cdot \max\{X_B(k_b)\} \quad \text{and} \quad X_B(\nu_0 + I + j, k_b) < \delta_t, \quad (2.8)$$

for $j = 1, 2, \dots, J$. The values of δ_{max} , δ_t , J_{min} and J_{max} are set at 0.1, 0.2, 4 and 10 respectively. Upon identifying valid end-points and flat regions in the reference signal Bark spectrum, using (2.6) - (2.8), the direct path energy $D_{k_b}(p)$ is defined as the energy of the p -th end-point identified in Bark bin k_b , for $p = 1, \dots, M_{k_b}$. The decay curve fittings, from which $A_{k_b}(p)$ and $\lambda_{k_b}(p)$ are estimated, take place in the identified flat regions within each Bark bin, in the Bark spectral difference between the reverberant and reference signal. The fittings are made across flat region lengths of $I_{k_b}(p) + J_{k_b}(p)$. Finally, any fittings that return values of $A_{k_b}(p)$ and $\lambda_{k_b}(p)$ outside the range of $[0, 1]$ and $[0, 100]$ respectively, are deemed outliers and are discarded.

2.3 Extensions and improvements

To improve the performance of the R_{DT} algorithm proposed in [15], several areas that could be extended or further developed were identified. Chief amongst them was that the R_{DT} employed a basic perceptual model in the calculation of the Bark spectrum of a signal and was restricted to operating on narrowband speech (that is speech bandlimited to 4 kHz). It was hypothesised that information from higher frequencies, along with a more realistic perceptual model, could improve the performance of the measure. The measure we propose, known as the extended Reverberation Decay Tail (R_{DTx}), thus extends the R_{DT} to operate on wideband speech (with a bandwidth as high as 12 kHz) and incorporates a more realistic perceptual model. Furthermore, an improved decay curve detection scheme is also presented.

2.3.1 Perceptual model

In the R_{DT} implementation of [15], the Bark spectrum of a signal is obtained using the method described in the seminal BSD paper [37]. As the focus of that work was on developing a quality measure for network transmitted speech, the critical band filters employed were selected specifically for narrowband speech. In order to operate on wideband speech, the R_{DTx} measure we propose uses an increased number of critical band filters, centred at higher frequencies. Furthermore, the increased number of filters allows the effect of the

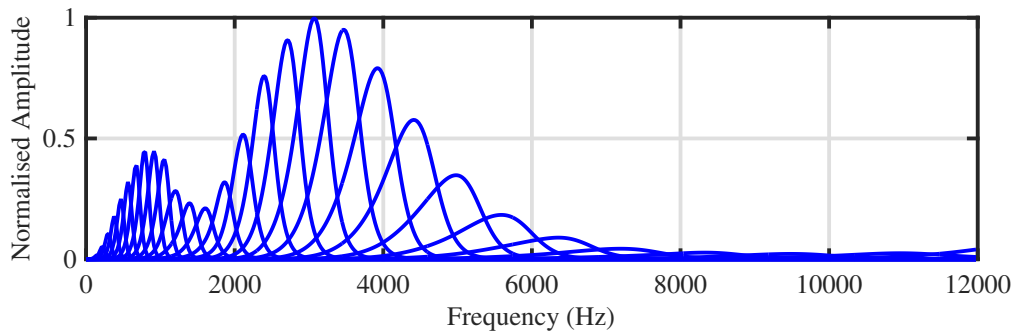
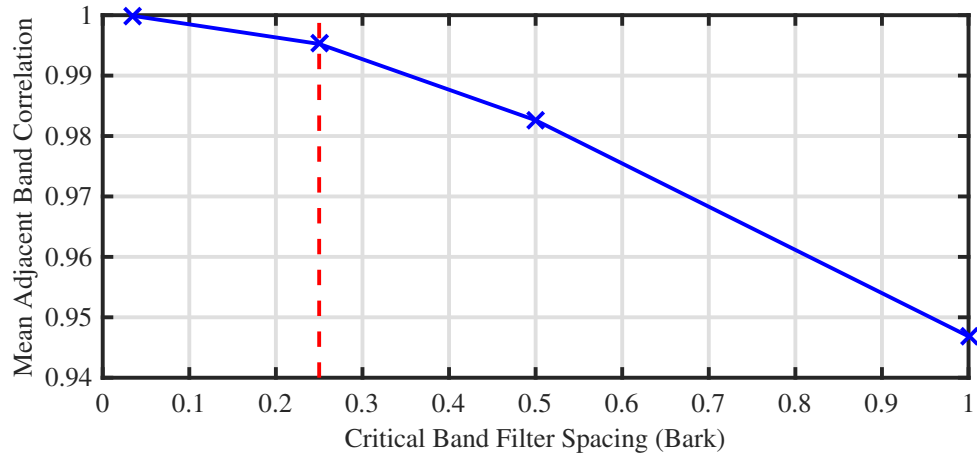


Figure 2.7: Critical band filters, weighted by equal loudness contours, used in the calculation of the Bark spectra of wideband speech signals by R_{DTX} (for visual clarity purposes, only a limited number of filters are displayed).

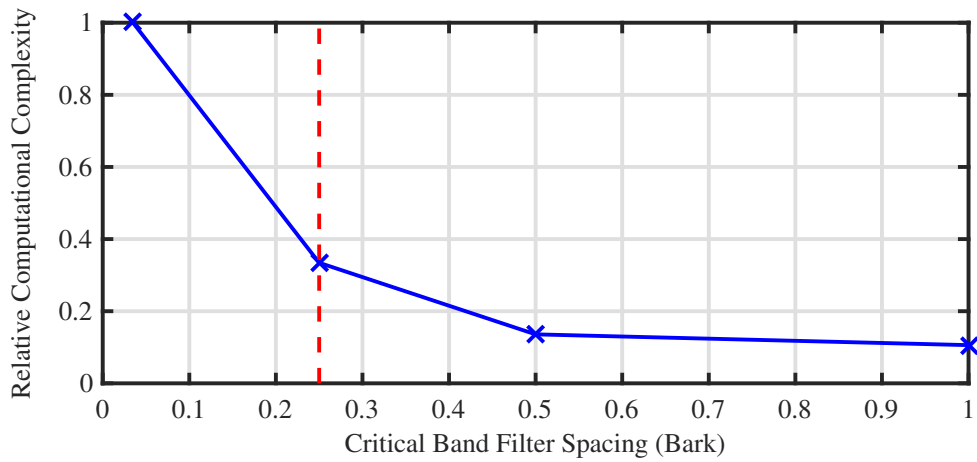
reverberant decay tail to be evaluated in a wider range of frequency bands, when compared to the R_{DT} implementation of [15], which it is believed would lower the variance of the estimated decay parameters.

A trade-off between the Bark spectral resolution (determined by the number of critical band filters used) and computational costs was studied experimentally on the entire TIMIT [50] test set (1344 speech signals). This database is a widely used speech database consisting of male and female speakers from every US dialect region. This investigation demonstrated that the highest Bark spectral resolution, which corresponded to a critical band filter centred per Discrete Fourier transform (DFT) point, was computationally intensive compared to lower resolutions, as illustrated in Figure 2.8. In particular, a critical band spacing of 0.25 Bark resulted in a significant reduction in computational cost, whilst still providing a high number of Bark bins, for which the average correlation between adjacent bins was still similar to that of the highest resolution. The latter fact indicates that no significant information was lost by adopting this resolution. Consequently a design was chosen in which critical band filters were positioned every 0.25 Bark.

Additionally equal loudness weightings, as specified in the revised ISO 226 standard [51] (illustrated in Figure 2.9), were incorporated into the Bark spectrum calculation used by the R_{DTX} . The curves represent loudness as perceived by an average human's hearing. Each curve represents the intensity a sound has to be for it to have the same loudness as a 1 kHz tone of a particular intensity (defined in phons). The curves thus



(a)



(b)

Figure 2.8: The impact of critical band filter spacing on the calculation of the Bark spectrum with respect to the average (a) correlation between adjacent Bark bins and (b) relative computational complexity in the calculation of the signal Bark spectrum. The dashed red line indicates the filter spacing selected in the R_{DTX} design.

illustrate the fact that lower frequencies have to be much higher in intensity to have the same perceived loudness as higher frequencies, where the ear's greater sensitivity around 3 - 4 kHz is due to the resonance in the auditory canal.

Thus by weighting the critical band filters with the 60 phon equal loudness contour, the auditory system's sensitivity to different frequencies can be modelled more accurately

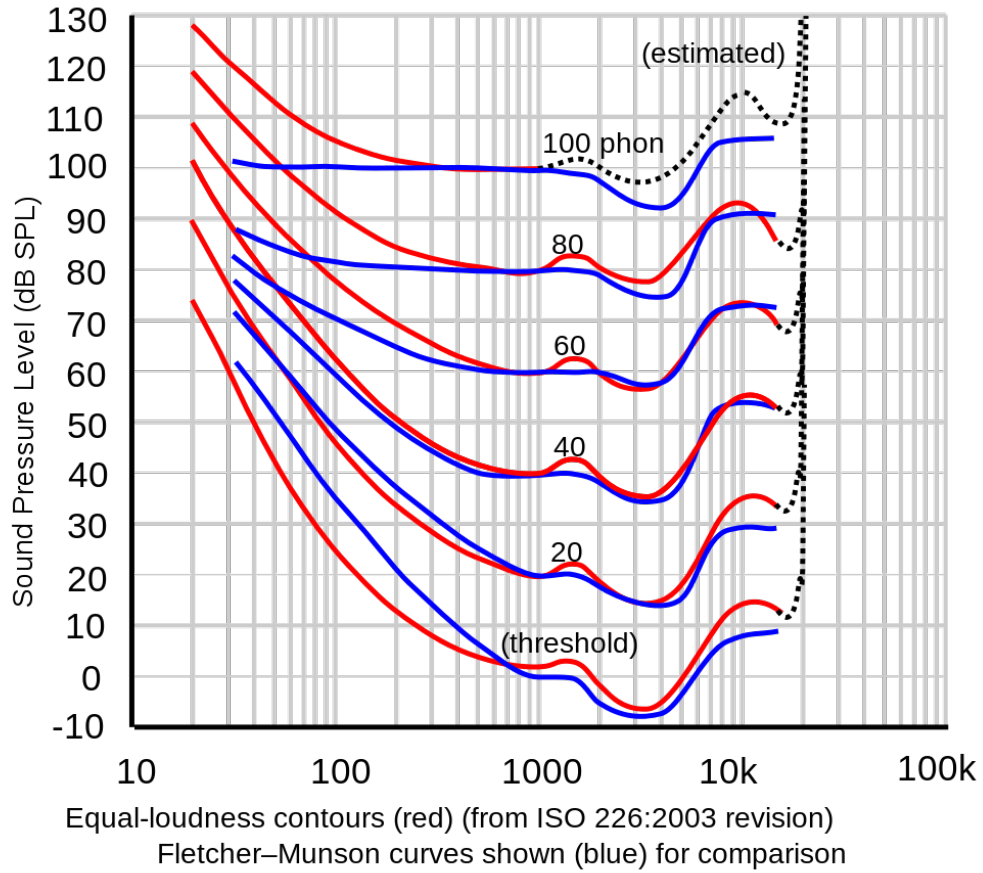


Figure 2.9: The revised ISO 226:2003 standard equal-loudness curves compared with the previous versions (based on the Fletcher-Munson curves).

than with the high pass filter used in the implementation of the R_{DT} . The choice of this contour was motivated by the fact that 60 dB SPL approximately corresponds to the power of conversational speech, spoken by a talker from 1m away. This in turn corresponds to the level of power we would expect from the speech signals under test in our work. It is also worth noting that for our purposes it is the relative weighting applied to each frequency which is important, and that in this respect the 20 - 80 phon contours are similar to each other. The aforementioned modifications are illustrated in Figures 2.5 and 2.7, which show the critical band filters used for the Bark spectrum calculation of the R_{DT} and the extended version developed in this work, respectively. The latter contains critical band filters that are positioned at higher frequencies and which are weighted by a more accurate equal loudness model. Consequently, a higher frequency range and perceptually relevant

Table 2.1: The impact of block size on the average energy of detected end-points, Γ , the level of ripple in detected flat regions, Λ , and their ratio for the speech set under test. A block size of 3 s was found to give the highest ratio of Γ to Λ .

Block size (s)	1.0	2.0	3.0	4.0	5.0
Γ (Sones $\times 10^2$)	3.60	7.45	9.91	1.04	1.05
Λ (Sones ² $\times 10^{-5}$)	0.645	1.03	1.19	1.28	1.35
$\Gamma : \Lambda$ (Sones ⁻¹ $\times 10^3$)	5.59	7.22	8.35	8.06	7.75

Bark spectrum is used in the calculation of R_{DTX} .

2.3.2 Decay curve detection

As described in Section 2.1, the R_{DT} implementation of [15] determines end-points and flat regions by gradient and absolute thresholds, which are in turn defined as fractions of a global peak in each Bark bin energy envelope. Such a detection scheme is susceptible to errors. Most notably, a spurious spike of energy, due perhaps to impulsive noise in even one segment of the captured audio, will result in errors in flat region detection. Moreover, it was observed that even in the absence of erroneous spikes, the scheme of [15] often missed what would be deemed through inspection to be valid flat regions, whilst also occasionally incorrectly identifying other regions as flat.

To overcome these limitations, a detection scheme that uses local rather than global peaks to define the gradient and absolute thresholds is proposed. This is achieved by working on 50% overlapping signal blocks, rather than the entire signal. An ideal end-point and flat region within a Bark bin should exhibit a large drop in energy followed by a completely flat section with minimum ripple. To assess the effect of block size on detecting such segments within a Bark bin, we measured both the average energy of detected end-points and the variance of detected flat regions, where the latter was an indication of the level of ripple present (with a higher variance corresponding to larger ripple). Denoted by the symbols Γ and Λ respectively, the ratio of the average end-point energy and level of ripple was taken, and measured for different block sizes. The results of carrying out this experiment, once again using speech from the TIMIT test set, are shown in Table 2.1.

It was observed that a smaller block size tended to lead to flat regions with less

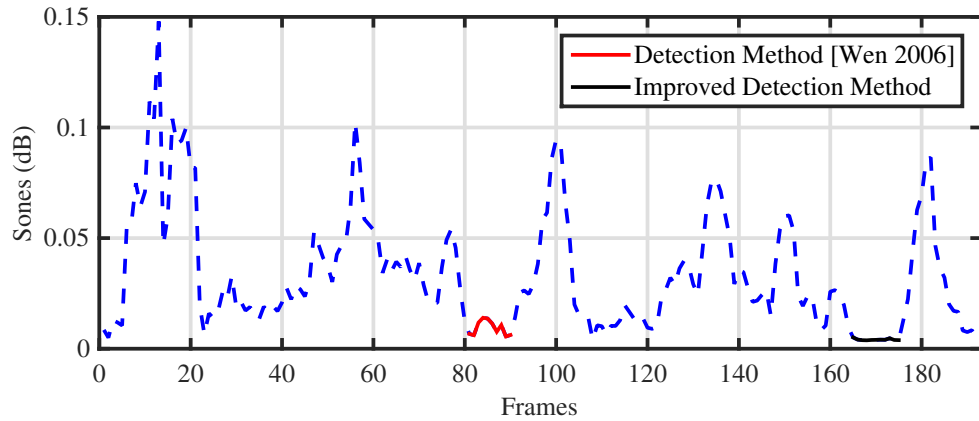


Figure 2.10: Identified flat regions highlighted in the energy envelope of a clean signal, using the method of [15] and the improved scheme proposed for the R_{DTX} . Dashed lines indicate loudness and solid lines indicate detected flat regions in an arbitrarily selected Bark bin.

ripple, but conversely that there was a drop in average end-point energy. Considering this trade-off, a block size of duration 3 s was found to be an effective compromise, after a range from 1 s to 5 s was tested, as this maximised the ratio under test. It is believed that this approach achieved more accurate flat region detection, with an illustrative example case shown in Figure 2.10, in which the method of [15] recognises the frames shown in red as a flat region despite the significant ripple present, whilst the improved detection scheme does not. On the other hand, the improved scheme identifies a valid flat region which was missed by [15], because the energy drop preceding it was not seen as sufficiently large.

2.4 Performance evaluation

In order to assess the effectiveness of the developed R_{DTX} measure, its performance was evaluated in three preliminary experiments, details of which are provided in this section. Whilst a more thorough assessment of the efficacy of the R_{DT} and R_{DTX} measures will be provided in the Chapter 3, the investigations described here provide some initial insight into the measure's effectiveness.

2.4.1 R_{DTX} versus T_{60}

In order to test the effectiveness of the extended measure at determining the level of reverberation in a captured signal, the measure was computed on a large number of reverberant test signals and compared against the T_{60} , an acoustic parameter regarded as being closely linked to perception [42]. A broad range of speech signals were taken from the widely used TIMIT database [50], comprising male and female speakers from every US dialect region. Three sentences spoken by the same speaker were concatenated with a 1 s inter-sentence pause, to produce a 10-15 s audio signal. In total, a speech set consisting of 50 signals was employed for testing purposes. Reverberant speech was simulated by convolving the dry TIMIT signals with simulated RIRs, as next described.

A single monophonic RIR was generated by creating a prototype response using the image method [52, 53], which is a widely used technique to create synthetic RIRs. Exponential decay masks with varying decay rates were then applied to the prototype response to produce a set of RIRs with different T_{60} times. The image method uses geometric rules to model reflections within a rectangular room, of specified dimensions, for a given source and microphone location setup. In addition to room dimensions and relative source-microphone positioning, the reflection coefficients of the walls, floor and ceiling of the room can be specified to produce a RIR with a particular T_{60} . Alternatively, using the RIR generator tool of [53], a T_{60} can instead be specified. The method then derives an average reflection coefficient for all surfaces, which will produce the required T_{60} . This is done using Sabine's formula, which describes the reverberation time of a room as a function of room dimensions and surface absorption coefficients [54].

The prototype RIR used in this experiment was generated using the image method for a room of dimensions 6 x 4 x 3 m, with the microphone placed in the centre of the room and the source placed 3.5 m away, at position [1,1,1]. A series of exponential decay masks were then applied to the prototype RIR to produce a set of 13 RIRs with different T_{60} times, ranging from approximately 0.2 to 1.70 s. As illustrated in Figure 2.11, only the effect of late reflections on the measure were tested in this way; by applying the decay masks exclusively to the reverberation tail of the prototype RIR (i.e. reflections occurring

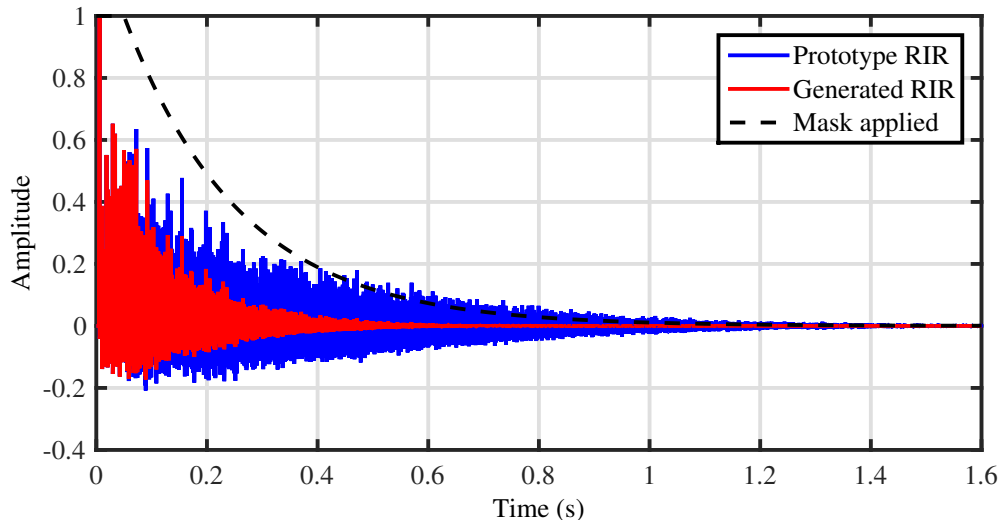


Figure 2.11: RIR generation from a prototype impulse response using an exponential decay mask applied to the reverberant tail of the prototype response.

50 ms after the direct path). Early reflections in the generated responses were, by contrast, not altered by the applied mask and thus allowed the effect of the reverberant tail alone on the R_{DTX} to be investigated in this experiment.

The ground truth T_{60} for each generated RIR was determined using the method of Karjalainen *et al.* [55]. The Karjalainen method calculates the reverberation time of a RIR through the nonlinear optimisation of a model of exponential decay in the presence of a stationary noise floor. By accounting for the presence of noise, it has been shown to outperform competing methods at estimating the T_{60} , particularly when the RIR measurements may be noisy. It has therefore been widely adopted by the research community, having recently been used to determine the ground truth T_{60} times in the Acoustic Characterisation of Environments (ACE) Challenge [56].

The results of the R_{DTX} measure on the reverberant TIMIT speech set described above, can be seen in Figure 2.12. The plot illustrates the relationship between the T_{60} , an acoustic parameter that can be reasonably expected to correlate with human perception, and the R_{DTX} measure. It is immediately clear that there is a positive correlation between the two measures (giving a Pearson correlation coefficient of 0.975). Whilst the average R_{DTX} score increases with T_{60} , the boxplots indicate that the variance does so as well.

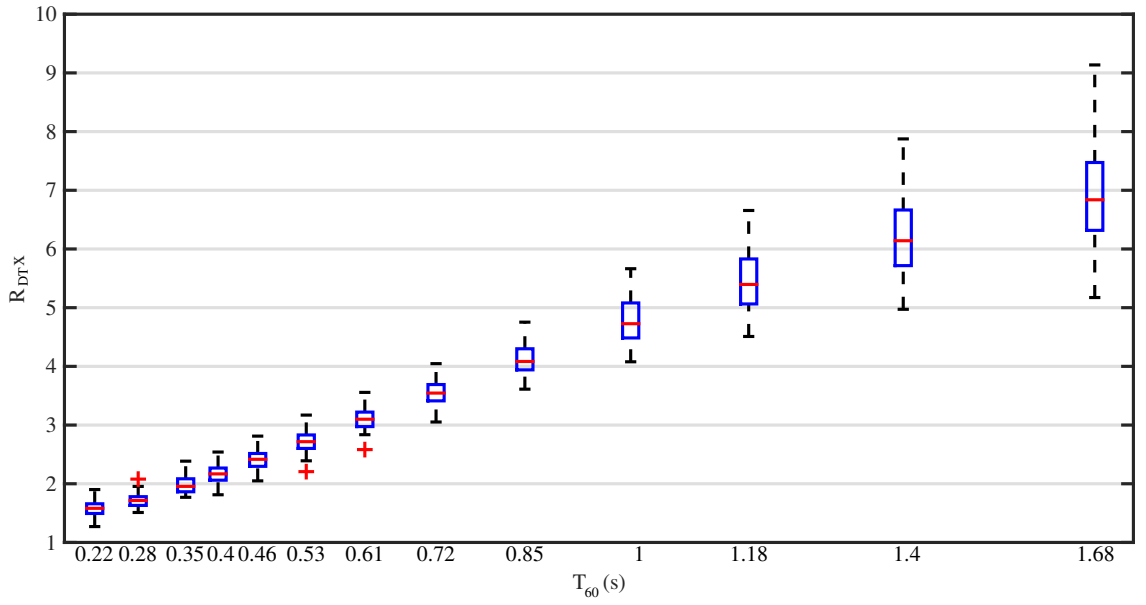


Figure 2.12: Boxplots illustrating R_{DTX} value distribution for 50 TIMIT speech signals that have been convolved with artificial RIR recordings of varying T_{60} values.

This becomes increasingly significant for reverberation times greater than 1 s.

The increasing variance could be attributed to the fact that RIRs with high T_{60} times would mean that flat regions, in which decay fittings take place, are more susceptible to temporal smearing from the frames preceding an end-point. Depending on the spectrum envelope in these preceding frames, that is whether they are low or high in energy, the curve fitting would be affected significantly. This implies that the measured R_{DTX} for high T_{60} times is highly dependent on the speech signal and how it is spoken. However, as most acoustic environments in which speech processing technologies could typically be expected to operate would not exceed reverberation times of 1 s, the ability of the R_{DTX} measure to discriminate between reverberation conditions (indicated by the limited overlap between adjacent boxplots) before this threshold, is a promising result.

Furthermore, comparing the performance of the R_{DTX} measure with the R_{DT} implementation of [15], demonstrates that the former exhibits less variation for each reverberation condition. The results of this comparison are illustrated in Figure 2.13. Whilst the mean values of the two measures are similar, the variation in scores is significantly smaller for the extended measure. In particular, for moderate T_{60} times they are up to

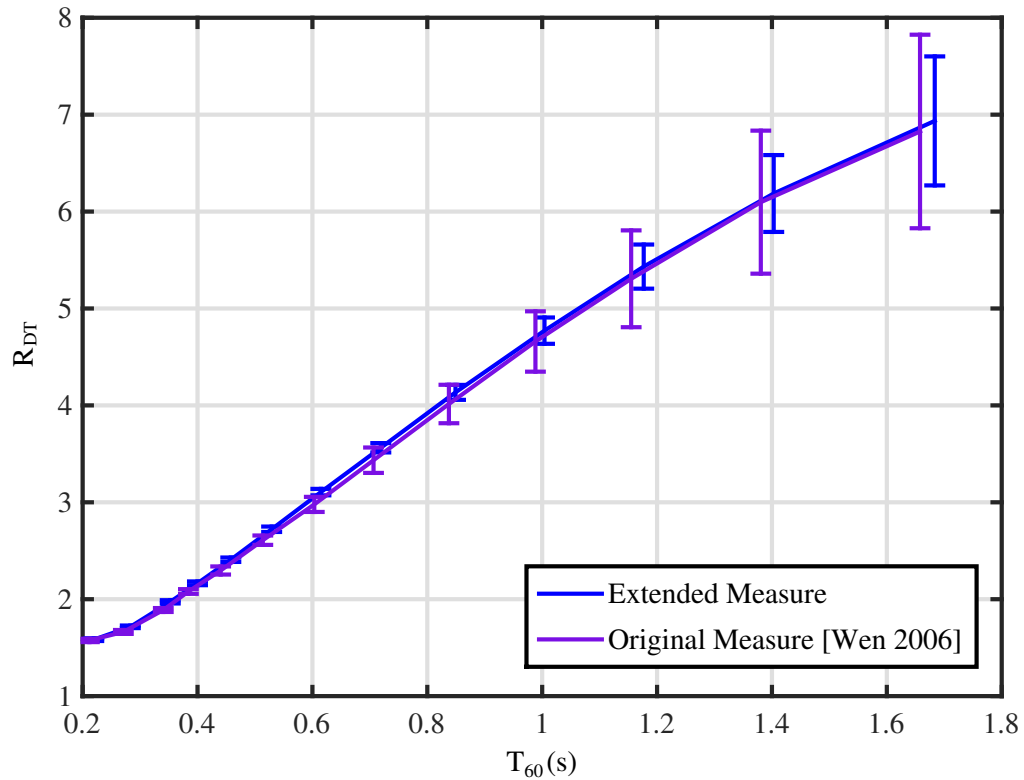


Figure 2.13: R_{DTX} mean and variance of reverberant TIMIT speech signals, compared with the R_{DT} implementation of [15].

50% smaller. This result provides evidence of improved performance, and to an extent validates the development and use of the extended measure over its predecessor.

Finally, the robustness of the R_{DTX} measure to common types of noise was also examined by adding noise from the NOISEX-92 database [57] to the test signals in this experiment. The noise types employed were babble, car and white, with signal-to-noise ratios (SNRs) of 20, 10 and 0 dB. Results testing the robustness of the measure to common types of noise are also presented. As an illustrative example, Figure 2.14 shows the R_{DTX} mean and variance for this experimental speech set in the presence of different types of additive noise, with a SNR of 20 dB. The results for the 10 and 0 dB cases (as well as the 20 dB case) are summarised in Tables 2.1-2.3, these demonstrate that the R_{DTX} mean and variance increase as SNR decreases. The bias introduced is most likely the result of the measure mistaking noise energy for reverberation energy in the flat regions of the test signal. The extent of these deviations from the noise-free case is also dependent on the

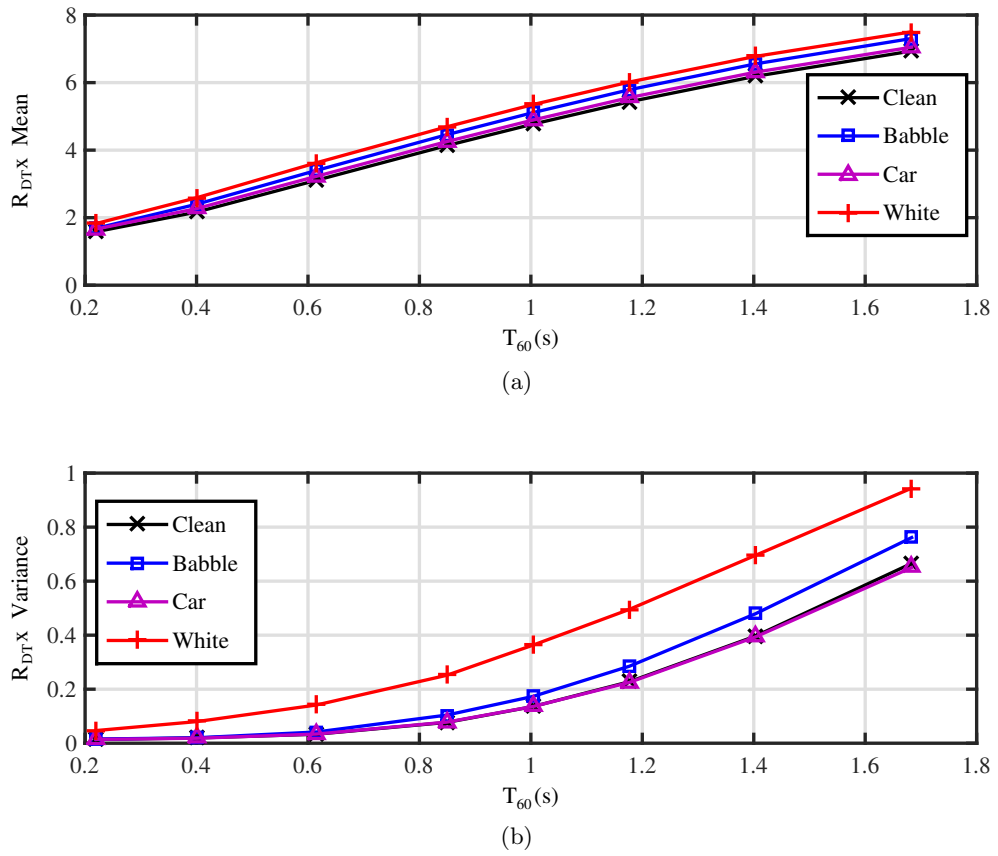


Figure 2.14: R_{DTX} (a) mean and (b) variance of reverberant TIMIT speech signals in the presence of common types of additive noise, SNR 20dB.

type of noise; with car noise causing the least deviation and white noise the most. This is expected if we consider the equal loudness curves employed in the R_{DTX} measure. Low frequency signals, such as car noise, would have their impact weighted down compared to babble noise for example. As white noise covers the entire spectrum, it results in the biggest increase in R_{DTX} values.

2.4.2 R_{DTX} versus C_{50}

In a second experiment, the relationship between R_{DTX} and C_{50} was investigated. This experiment differed from the first primarily in that the reverberant signals used were produced via convolution with real RIR measurements taken from the Aachen Impulse Response database [58]. The database contains impulse responses from a variety of acoustic

Table 2.2: R_{DTX} mean (variance) for reverberant and noisy TIMIT speech, SNR 20 dB.

Noise type	T_{60} (s)		
	0.2	0.6	1.0
Clean	1.58 (0.014)	3.11 (0.033)	4.77 (0.136)
Babble	1.68 (0.015)	3.40 (0.042)	5.11 (0.173)
Car	1.66 (0.014)	3.22 (0.034)	4.89 (0.136)
White	1.82 (0.047)	3.62 (0.141)	5.35 (0.363)

Table 2.3: R_{DTX} mean (variance) for reverberant and noisy TIMIT speech, SNR 10 dB.

Noise type	T_{60} (s)		
	0.2	0.6	1.0
Clean	1.58 (0.014)	3.11 (0.033)	4.77 (0.136)
Babble	1.69 (0.020)	3.51 (0.051)	5.24 (0.188)
Car	1.66 (0.014)	3.22 (0.035)	4.89 (0.135)
White	2.04 (0.080)	4.08 (0.274)	5.91 (0.603)

Table 2.4: R_{DTX} mean (variance) for reverberant and noisy TIMIT speech, SNR 0 dB.

Noise type	T_{60} (s)		
	0.2	0.6	1.0
Clean	1.58 (0.014)	3.11 (0.033)	4.77 (0.136)
Babble	1.71 (0.026)	3.64 (0.060)	5.37 (0.205)
Car	1.62 (0.015)	3.13 (0.033)	4.75 (0.131)
White	2.38 (0.139)	4.76 (0.496)	6.75 (1.022)

environments, ranging from studio booths to cathedral halls, thus allowing a wide range of reverberation conditions to be more realistically simulated. Unlike the first experiment, in which early reflections were kept constant across different reverberation conditions, insights into the ability of the R_{DTX} to measure the impact of the reverberation tail independently of colouration effects can be gained in this experiment. As before, concatenated speech signals from TIMIT were used for testing purposes. This time, however, a set of a 100 signals was considered, along with 13 test RIRs. As the relative positioning of the source in relation to microphone is inherently captured by the C_{50} parameter, it is believed it may contain additional information on perceived reverberation when compared to just the T_{60} .

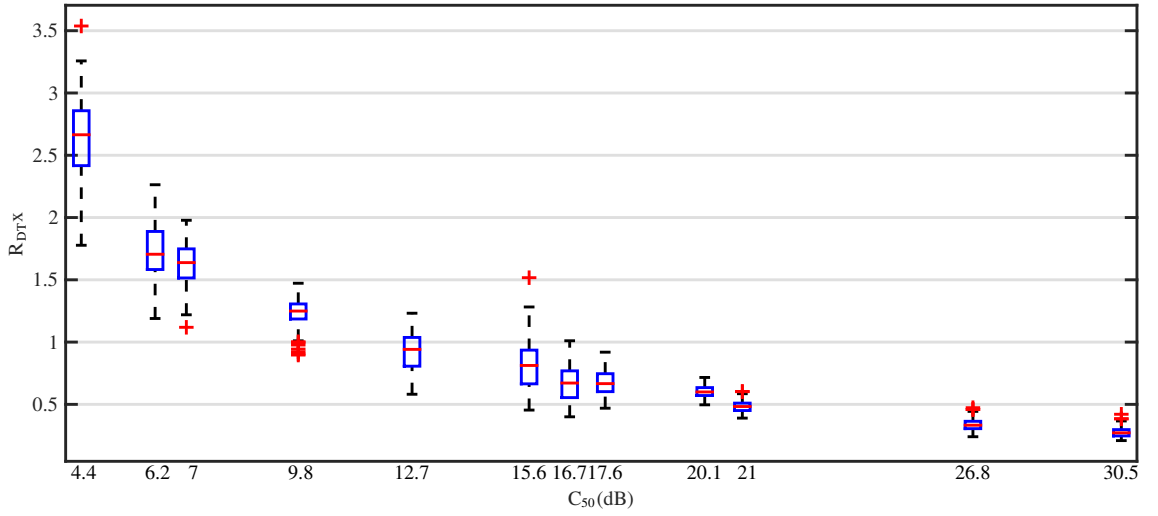


Figure 2.15: Boxplots illustrating R_{DTX} value distribution for 100 TIMIT speech signals that have been convolved with measured RIR recordings of varying C_{50} values.

The C_{50} was chosen in preference to the DRR (a metric that also captures positioning information), as the latter considers all reflections to be detrimental, whilst the aim of the R_{DTX} is to measure the impact of the reverberation tail alone. Thus it was felt that the C_{50} would be the more appropriate metric of the two.

Evaluating the measure on reverberant speech generated using real RIR recordings, the relationship between R_{DTX} and C_{50} , illustrated in Figure 2.15, provides further validation of the measure. Like the results of the R_{DTX} against the T_{60} of artificially generated reverberant signals, a clear trend can be observed. From inspection, an exponential function could be used to map between the two measures. Whilst the mean R_{DTX} score increases with decreasing C_{50} , the boxplots indicate that the variance does so as well. This becomes increasingly significant for small clarity index values. However, for a given speech signal that is convolved with RIRs of decreasing C_{50} , the R_{DTX} tends to decrease monotonically.

2.5 Conclusion

It is well known that in many real-life acoustic environments, reverberation can have a detrimental effect on speech quality and intelligibility. Yet despite this knowledge, developing a measure capable of assessing the perceptual impact of reverberation on captured speech remains an open area of research. In this chapter, the R_{DTX} measure was developed to estimate the level of late reverberation present in speech recordings. The signal-based measure attempts to predict the perceived level of reverberation, by characterising reverberation decays in the perceptual Bark spectrum of a signal. The R_{DTX} , in comparison with the R_{DT} [15], is extended to operate on wideband speech and incorporates an improved perceptual model and decay detection scheme. The performance of the extended measure is experimentally tested on a wide range of reverberant speech, and shown to correlate with objective acoustic parameters related to perception. Furthermore, when compared to the R_{DT} , the extended measure exhibits lower variance across speech signals for a given reverberation level. Thus providing initial evidence of its practical utility, and to an extent validating the extensions proposed.

Chapter 3

Perceptual measurements

In the preceding chapter we presented the R_{DTX} , a heuristically developed instrumental measure for predicting the perceived level of reverberation. The performance of the developed measure was primarily assessed by determining its correlations with acoustic parameters such as the T_{60} and C_{50} , both of which are generally regarded as being related to reverberation perception. However, in order to verify the efficacy and reliability of the R_{DTX} as a predictor of perceived reverberation, the performance of the measure must also be evaluated on subjectively scored test signals.

To this end, in this chapter we present the design and results of an experimental listening test, carried out to obtain subjective assessments of the level of perceived reverberation. The perceptual measurements gathered are first analysed to determine a human assessor's ability to perceive different levels of reverberation in speech produced in a range of acoustic environments and captured by a single microphone. Having ascertained participant sensitivity to the test material, we proceed to identify acoustic parameters that most strongly influence perception. The accuracy of several parametric models, which use a combination of the aforementioned acoustic parameters, at predicting perceived reverberation is then investigated. These findings are incorporated into the development and validation of the R_{DTX} measure. Finally, the listening test data is also used to assess and compare the performance of existing instrumental measures for predicting the perceived level of reverberation.

3.1 Review

3.1.1 Parametric models of reverberation

As we saw in Section 2.1 of the preceding chapter, despite the growing interest in reverberation measures that correlate with human perception, to date there is no consensus within the research community on the use of a single measure. As a result, it is common practice to use quality measures not specifically designed to evaluate reverberation perception. For example, many proposed dereverberation techniques have their performance evaluated by the apparent reductions or improvements they are able to achieve in objective acoustic parameters, such as the T_{60} and DRR, respectively. By describing the acoustic space a sound is produced in, these metrics clearly provide insight into reverberation perception. To date, efforts to model the relationship between acoustic parameters and perceived reverberation have been explored most extensively in the context of architectural acoustics and concert hall designs, e.g. in [59–61]. Whilst valuable insights can be gained from these studies, they usually concern acoustic settings with high levels of reverberation that are not typically encountered in day-to-day scenarios, which in turn may limit their utility for speech processing technologies.

On the other hand, several parametric models proposed in the literature may be more relevant, such as the one proposed by Allen over three decades ago [42]. Perhaps the first attempt to develop such a model, the subjective preference of reverberant speech was reported to be a function of the T_{60} and room spectral variance (RSV), σ_I^2 , and was later validated experimentally in [62]. Referred to as Allen’s score in this work, it is mathematically expressed as

$$\mathcal{A}_s = \mathcal{A}_{max} - \sigma_I^2 T_{60}, \quad (3.1)$$

where $\mathcal{A}_{max}$ is chosen arbitrarily to represent the maximum preference score possible. Although an effective model even by today’s standards, as demonstrated in the study of [43], it was primarily devised for short source-listener distances and thus purported to be less effective for larger distances [44].

More recently, Paulus *et al.* proposed a provisional model derived from listening test

data in which users scored artificially generated reverberant signals [31]. The measure, referred hereafter as Paulus' score, models the impact of changes in T_{60} and DRR on overall perception. Namely,

$$\Delta \mathcal{P}_s = 19.2 \Delta T_{60} - 2.35 \Delta DRR, \quad (3.2)$$

indicating the relatively higher importance of the T_{60} in reverberation perception, which has been observed in other works [63].

In [64], three main acoustic features were identified as playing a key role in predicting perceived reverberation. In the work that followed [43, 44], the authors proposed the QAreverb measure,

$$\mathcal{Q} = -\frac{(T_{60})^\alpha (\sigma_I^2)^\beta}{(DRR)^\gamma}, \quad (3.3)$$

where σ_I^2 denotes the RSV and exponents α , β and γ are introduced to model the relative contribution of the three acoustic parameters. Initial experiments showed that by incorporating the DRR into Allen's score ($\alpha = 1$, $\beta = 1$, $\gamma = 0.3$), the resulting QAreverb measure, produced from these specific exponent weights and referred to as $\mathcal{Q}_m$ in this thesis, correlated best with subjective ratings [43]. However, more extensive experiments presented in [44], demonstrated that the RSV could be discarded altogether, with an optimal weight selection instead of $\alpha = 0.6$, $\beta = 0$ and $\gamma = 0.15$. To distinguish itself from its predecessor, the QAreverb measure computed using these weights is hereafter referred to as $\mathcal{Q}_{mX}$.

3.1.2 Perceptual evaluation of reverberation

When it comes to evaluating the perceived qualities of an audio signal, perceptual measurements obtained from listening tests are usually considered more reliable than those obtained from objective measures [65]. However, listening tests are more often than not labour-intensive, time-consuming and expensive. Furthermore, due to variations in assessors and test stimuli, the resulting data is not always statistically significant. Consequently, extreme care is required in the design of such tests, to ensure conclusions can be reliably

drawn from the data obtained [66]. Carefully obtained perceptual measurements can be used to develop reliable objective measures that correlate with perception. These can in turn be used to evaluate proposed enhancement techniques in situations when it is not practically possible to conduct listening tests.

Currently, the most popular standards for the perceptual measurement of dereverberation systems have been developed for different applications and consequently may not always be the best methods to assess reverberation. The 5-point mean opinion score (MOS) [67] which, for example, was used for the evaluation of single-channel speech dereverberation systems in [68], was originally developed for the evaluation of noise suppression techniques. Similarly the multi stimuli test with hidden reference and anchor (MUSHRA) scheme [69] was used for the evaluation of single-channel and multi-channel speech dereverberation systems, for example in [7, 70], although it was originally developed for the evaluation of audio codecs.

In order to specifically evaluate the the perceived level of reverberation, two alternative schemes were proposed and investigated by the author as part of a collaboration in [71]. The schemes considered were a pairwise comparison for the evaluation of the perceived level of reverberation (PCEPLR), and a multi stimuli test with hidden reference and anchor for reverberant speech. In the latter, modifications to the widely used MUSHRA standard were applied, which aimed at making this test more efficient in the evaluation of the perceived level of reverberation. It differs from the standard MUSHRA scheme in the choice of low and high reverberant speech utterances as reference and anchor signals respectively. Furthermore, the scoring scale (0-100) is inverted such that a more reverberant signal is attributed with a higher score, which is more intuitive to the listener.

However, as different reverberant conditions can be perceived as strongly similar by some assessors, the simultaneous comparison of stimuli can be difficult and may result in the lack of statistical significance sometimes obtained in such tests [72]. The pairwise comparison for the evaluation of the perceived level of reverberation (PCEPLR) scheme that was proposed, aims to facilitate the task of the assessor by asking them to compare only two audio signals at a time, as was similarly applied in the comparison of microphones

for the recording of a singing voice in [73].

Despite the contrasting test styles, the statistical analysis conducted in [71] showed that the two perceptual measures achieve similar results. Consequently, it concludes that the multi stimuli test with hidden reference and anchor for reverberant speech (MUSHRAR) framework is preferable from a practical point of view, as it requires less time from each assessor compared to a pairwise scheme.

3.2 Considered instrumental measures

In addition to analysing a listener’s ability to perceive reverberation, the perceptual measurements gathered from our listening test are used to evaluate several objective measures and models of perceived reverberation. Along with the RIR-based measures described in Section 3.1.1, we also evaluate the ability of several signal-based measures to predict the perceived level of reverberation.

We use the gathered data to evaluate the performance of the SRMR [26] and its extension SRMRx [27], whose operation was reviewed in Section 2.1. As a non-intrusive measure that does not require a reference signal, the SRMR has potentially significant practical value. Also considered is the PESQ score [38], which despite being originally developed for the evaluation of audio codec quality, is often used to evaluate the performances of speech dereverberation algorithms [7, 74]. Furthermore, it has been shown to correlate reasonably well with subjective assessment of reverberation [40].

3.2.1 Weighted Reverberation Decay Tail

The R_{DT} of [15], and the R_{DTx} proposed in this thesis, are also considered. Although a perceptually motivated measure, previous studies have shown the R_{DT} in some cases correlates poorly with subjective assessments of reverberation [47]. Furthermore, a more recent study showed that despite the proposed extensions, the R_{DTx} also showed average correlations with the perceptual measurements obtained in [71]. It must be noted, however, that only a small number of test conditions were considered in this study and the

conclusions drawn may not necessarily generalise.

It is hypothesised that the lack of significant correlation is due to the equal weighting of each component in the calculation of the measure, i.e.

$$R_{DT} = \frac{A_{avg}}{\lambda_{avg} D_{avg}}. \quad (3.4)$$

The R_{DT} and its extension can be thought of as a ratio of the estimated T_{60} and DRR. In order to improve its performance as a measure, we propose incorporating a weighting scheme into the computation of the measure,

$$wR_{DT} = \left(\frac{1}{\lambda_{avg}} \right)^\alpha \cdot \left(\frac{A_{avg}}{D_{avg}} \right)^\gamma \approx \frac{(T_{60})^\alpha}{(DRR)^\gamma}, \quad (3.5)$$

where the introduced exponents α and γ , better model the relative importance of the T_{60} and DRR. Inspired from [44], the wR_{DT} measure takes α and γ values of 0.6 and 0.15, respectively.

3.3 Experimental setup

This section describes the listening test conducted; the reverberant test conditions considered for assessment are summarised and the experimental setup detailed.

3.3.1 Test signals

The audio signals presented for assessment $y_{\iota,j}(n)$, were generated by convolving anechoic speech $s_\iota(n)$ with recorded RIRs $h_j(n)$ of length $\mathcal{L}_h$ and adding ambient noise $v_j(n)$ for each discrete time index n , i.e.

$$y_{\iota,j}(n) = \left(\sum_{\ell=0}^{\mathcal{L}_h-1} s_\iota(n-\ell) h_j(\ell) \right) + v_j(n), \quad (3.6)$$

where subscripts ι and j correspond to the utterance and test condition number respectively. The j -th additive noise term corresponds to recordings made of the ambient sounds within the same acoustic environment, using the same microphone placement as the j -th

Table 3.1: Selected RIR characteristics

RIR Label	A	B ₁	B ₂	C ₁	C ₂	D ₁	D ₂	E ₁	E ₂	F
T ₆₀ [s]	0.17	0.33	0.32	0.38	0.38	0.62	0.62	0.71	0.71	1.16
DRR [dB]	16.44	6.06	2.93	8.65	4.68	12.19	5.41	6.75	2.71	0.65

RIR measurement. The term thus represents the ambient noise associated with a given acoustic environment. For example, noise associated with a lobby will consist of sounds such as footsteps, sliding doors and so on. On the other hand the ambient noise within an empty office would be more discrete by comparison, and may simply consist of a light hum. All signals were sampled at 16 kHz.

For all listeners, a total of $M = 10$ RIRs were used to generate reverberant test signals. The RIRs were selected from either the SMARD corpus [75] ($j = 1$), or from the ACE Challenge [56] corpus ($j > 1$). These databases were chosen for the realistic RIR characteristics available, and for the fact that each RIR is accompanied by a recording of ambient noise, $v_j(n)$, which was recorded in the same room and at the same microphone position, thus enabling realistic reverberant and noisy test signals to be generated.

In order to produce a wide range of reverberant conditions for assessment, RIR recordings obtained from an office, meeting room, lecture theatre and building lobby were selected, all of which are acoustic environments in which speech processing technologies could typically be expected to operate. For each of these test environments, a near and far distance between the source and microphone was considered. The test space, summarised in Table 3.1, is therefore conveniently characterised by the T₆₀ and DRR. The ground truth T₆₀ of a given RIR was once again determined using the Karjalainen method [55] (described in more detail in Section 2.4.1). The DRR was defined as the ratio of the energy of the direct path to the rest of the response, that is

$$\text{DRR} = 10 \log_{10} \left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h^2(n)}{\sum_{n=0}^{n_d-n_0-1} h^2(n) + \sum_{n=n_d+n_0+1}^{\infty} h^2(n)} \right), \quad (3.7)$$

where n_d is the tap corresponding to the direct path (for which $|h(n = n_d)|$ is at its

maximum). Due to finite sampling, the direct path component is defined as an interval, n_0 , either side of the the direct path tap. As was done in [56], from which these RIRs are taken, a value of 128 samples (which corresponds to 8 ms), was used for n_0 . Labels A and F represent the RIRs used to generate a reference and anchor signal, whilst labels B through to E represent the aforementioned test environments, with corresponding subscripts $\{ \}_1$ and $\{ \}_2$ denoting the near and far configurations respectively.

Anechoic speech signals were taken from the ACE Challenge database, which consists of 8 talkers (4 male, 4 female) counting from one to nine in English. The test signals produced from these speech files had durations ranging from 11 s to 17 s. Noise was added to the speech signals so as to produce test signals that had a signal-to-noise ratio of 25 dB. The SNR was computed using the energy of the early reverberant speech, i.e. the speech file convolved with the first 50 ms of the RIR, as target signal and $v_j(n)$ as the noise term (with the speech energy calculated according to [76]). An SNR of 25 dB was chosen after an informal pilot study, consisting of several participants, was conducted. The consensus from this study was that this SNR produced natural sounding test signals, whilst keeping noise levels low enough that they were not deemed distracting. Finally, test signals were normalised to be equal in loudness over all conditions for each speech utterance, and pre-processed to not exceed a level of 65 dB SPL. This was done in software, and validated using the calibrated measuring equipment shown in Figure 3.1.

Numerous loudness equalisation approaches exist, based on loudness models of varying sophistication. The basic procedure in calculating how loud a sound is involves the time-frequency decomposition of that signal, followed by a stage that compensates for the response of the auditory system to individual frequencies, before summation to obtain the total loudness. Currently accepted methods differ primarily in the model of the auditory system employed and, to a lesser degree, in the details of their implementation. Perhaps the most well known method is the approach presented by Zwicker over three decades ago [77]. In the method proposed, spectral decomposition is performed using a filterbank of one-third octave band filters, which roughly approximate the shape and location of the auditory filters. An experimentally derived expression is then used to convert the dB SPL

in individual bands to a corresponding loudness level, before summing the loudness in each band to obtain the total signal loudness.

Variations of Zwicker’s model have gone on to be adopted for use in a number of International Telecommunication Union (ITU) speech and audio quality standards. Alternative implementations include the loudness model used in the perceptual speech quality measure (PSQM) [78] (a pre-cursor to PESQ [38]). Modifications over the model of [77] take the form of a filterbank that more accurately represents the spectral decomposition performed by the human auditory system (filters of increasing bandwidth, which are positioned every 0.321 Bark and are scaled relative to the filter bandwidth). The Zwicker model is also the basis for calculating signal loudness in the ITU perceptual evaluation of audio quality (PEAQ) measure. As before, modifications are made to improve the psychoacoustic model; a non-linear filterbank that more closely resembles the auditory filters is once again employed. Notable changes over the model used in the PSQM measure include the filter spacing used, as well as an offset that is added to individual critical band filter energies to simulate the internal noise generated in the ear.

Although the Zwicker model (and variations of it) continue to be widely used, competing loudness models such as the Moore-Glasberg model [79] and the direct loudness representation that forms the basis of the BSD measure [37], are also popular alternatives. In the former, the equivalent rectangular bandwidth frequency scale is used to perform spectral decomposition, and an alternative expression is used to compute loudness within individual bands. Whilst the loudness model used in the calculation of the Bark spectrum is explained in extensive detail in Chapter 2, the main difference between it and the other models mentioned, is the direct conversion of sound intensities within bands to loudness in sones. Implementation details of the various loudness methods outlined above, including the experimentally derived expressions used to convert the dB SPL sound level in individual bands into loudness levels, are usefully summarised in [80].

The test signals in this experiment were normalised using the loudness model employed within the $R_{\text{D}T\text{x}}$ measure proposed in this thesis, that is through the calculation of the signal Bark spectrum as described in Section 2.3.1. This extended version of the

loudness model used in the BSD measure of [37] was selected as it included modifications believed to more accurately model the peripheral auditory system (e.g. filters weighted by the equal loudness contours). Loudness equalisation of signals was achieved by first calculating the loudness of all signals within a test trial. Then having identified the loudest signal, an iterative process was used in which a scaling factor was applied in turn to all other test signals, until they were within 0.001 sones of the loudest signal. An interval bisection method was used to obtain the appropriate scaling factor for signal normalisation. Although the normalisation process adopted here was computationally intensive, the generation of test signals, which took approximately 5 minutes, could be done offline before participants sat the test. The approach was validated in an informal pilot study conducted before the test, in which several participants confirmed that they were not able to discern any difference in loudness of the signals (which included the judgement of two experienced listeners). Finally, for safety reasons, the measurement equipment shown in Figure 3.1 was used to measure and ensure the maximum sound pressure level of the test signals never exceeded 65 dB SPL. This was achieved by appropriately adjusting the gain of the amplifier driving the headphones, which was then kept fixed on those settings for all participants.

3.3.2 Listening test

To assess the perceived reverberation in the test signals, a MUSHRAR listening test framework, as described in Section 3.1.2, was used. This has previously been shown to be an effective method for perceptual evaluation of reverberation [71]. The evaluation GUI participants were presented with is shown in Figure 3.2.

A total of 28 self-reported normal hearing listeners aged 24 to 48 (average 33.5 years, standard deviation 6.4 years) participated in the listening test. For each participant, $N = 3$ speech utterances were randomly selected from the eight speech files described earlier. All participants undertook the test in a soundproof booth and listened to the diotic signals reproduced using an audio interface and closed-back headphones (Senheiser HDA 200). The booth, a photo of which is shown in Figure 3.3, was located in the House of

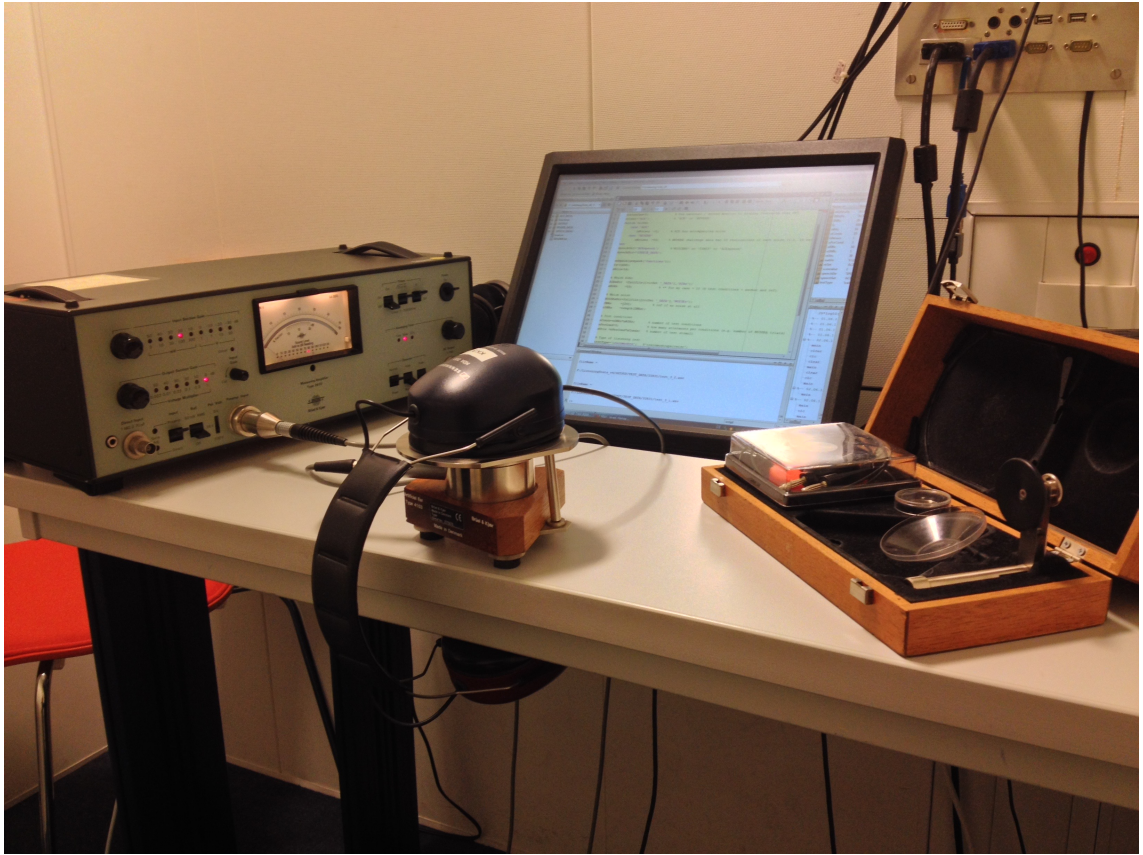


Figure 3.1: Listening test headphone calibration equipment. An ‘Artificial Ear’ (Brüel & Kjær, Type 4153) and measuring amplifier (Brüel & Kjær, Type 2610) were used to ensure none of the test signals that had to be evaluated exceeded 65 dB SPL.

Hearing at Fraunhofer IDMT in Oldenburg. The listening test facilities here are often used by clinicians to run tests for patients with hearing impairments, and as a result exhibit low noise and reverberation, as well as high sound isolation between the booth and adjacent rooms. Within the booth a touch screen monitor and mouse were setup for participants’ use. An illustrative example of a participant taking the test is shown in Figure 3.4. Each session started with spoken instruction read from a script prepared in advance (which can be seen in Appendix A), and a training phase which allowed the participants to familiarise themselves with the sound files they would be asked to assess in the testing phase. Finally, for each speech utterance, the order in which the reverberant test stimuli were presented was randomised to limit potential biases and training effects [65].

In order to avoid influencing participants, no indication was given on how long

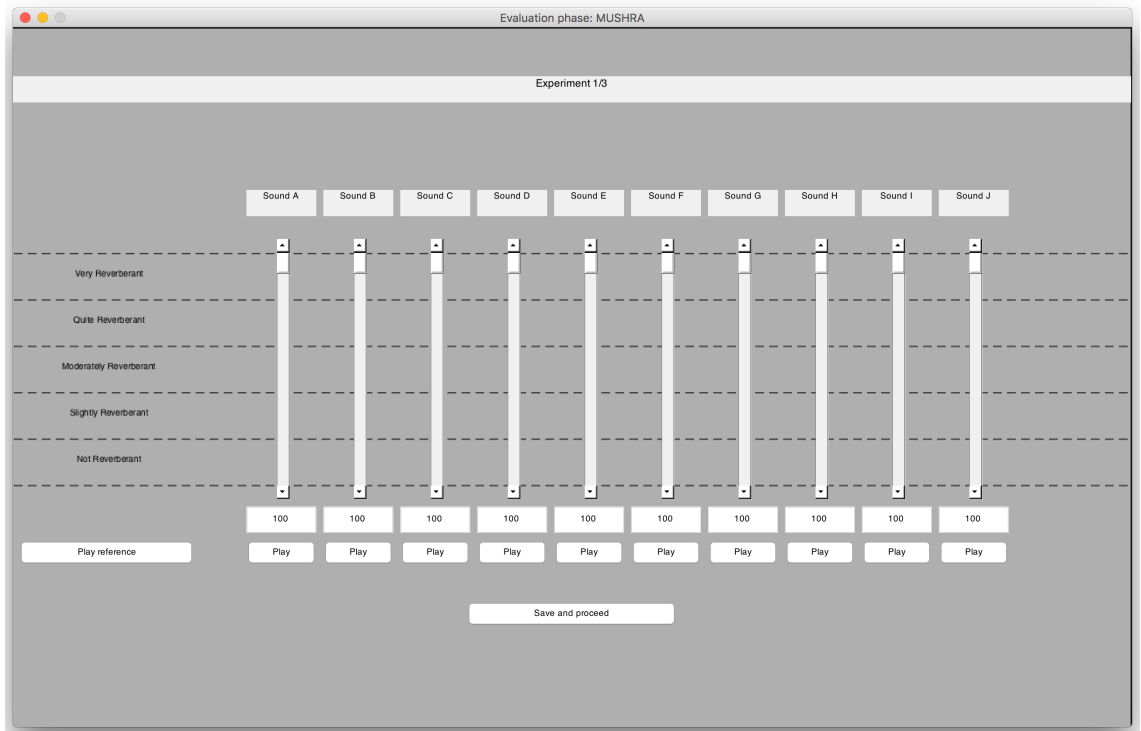


Figure 3.2: MUSHRAR listening test evaluation GUI.

they should spend on the test. Instead they were told that they could take as long as they needed to complete the task, and that they could take breaks at any stage if required. That said, for the purposes of preventing fatigue and unreliable results, the test was designed to take approximately 10-15 minutes to complete (as noted in Section 3.1.2, one of the reasons the MUSHRAR framework was chosen was the simultaneous rating of stimuli makes it less time consuming than alternative test schemes). After completing the test, a post-screening of the raw scores was performed. Scores from participants who reported they did not understand the test once it was underway were removed. Additionally, participants unable to identify the hidden reference or distinguish the anchor as the most reverberant signal were deemed unreliable and thus also excluded. After applying the above criteria four participants were removed from the final analysis, leaving a total of 24 participant scores.



Figure 3.3: Listening booth the subjective tests were conducted in.

3.4 Results and analysis

3.4.1 Participant scores

As can be seen in Figure 3.5, the vast majority of participants completed the test in under 20 minutes, with a median time of 10.75 minutes. It is worth noting that as there was no restriction placed on the time allowed, several participants took considerably longer. Two participants in particular took periodic breaks during the test, giving total test times of 38 and 42 minutes. The mean scores obtained per listener for each test condition (i.e. scores from individual participants are averaged across the three test trials), are depicted in Figure 3.6. Despite several closely spaced medians that indicate the difficulty in discriminating between certain test conditions, the ability to discriminate between other conditions can be observed in the data. Both the anchor and reference have been correctly



Figure 3.4: A participant undertaking the listening test ¹.

identified, suggesting the usefulness of these signals in removing unreliable assessors as part of the post-screening criteria. Scores attributed to the reference and anchor are hereafter excluded from the remaining analyses, unless explicitly stated otherwise.

In order to examine the perceived pairwise differences between conditions, the Wilcoxon rank-sum test is used. As a non-parametric statistical hypothesis test, the test is used to determine if the mean ranks of the populations of two groups of measurements differ. This test was chosen as it does not make any assumptions on the population distribution, unlike other popular approaches such as the paired Student's t-test. As multiple hypotheses are being tested in the form of score comparisons between eight conditions, the Bonferroni correction is applied. The correction compensates for the fact that a null hypothesis is more likely to be incorrectly rejected when multiple comparisons

¹The participant in question has given their permission for this photograph to be used in this thesis.

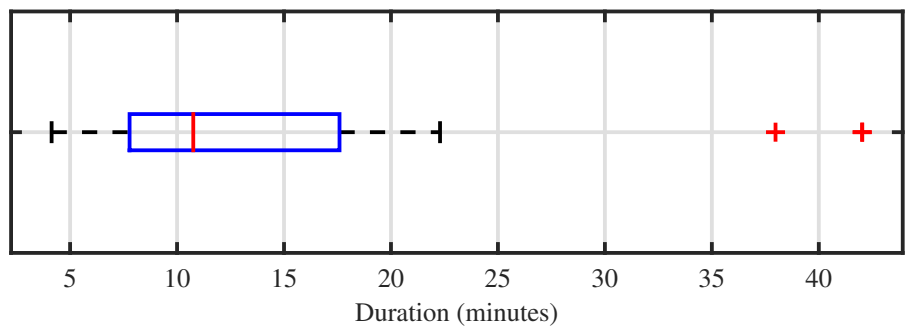


Figure 3.5: Boxplot showing the total time taken by individual participants to complete the listening test.

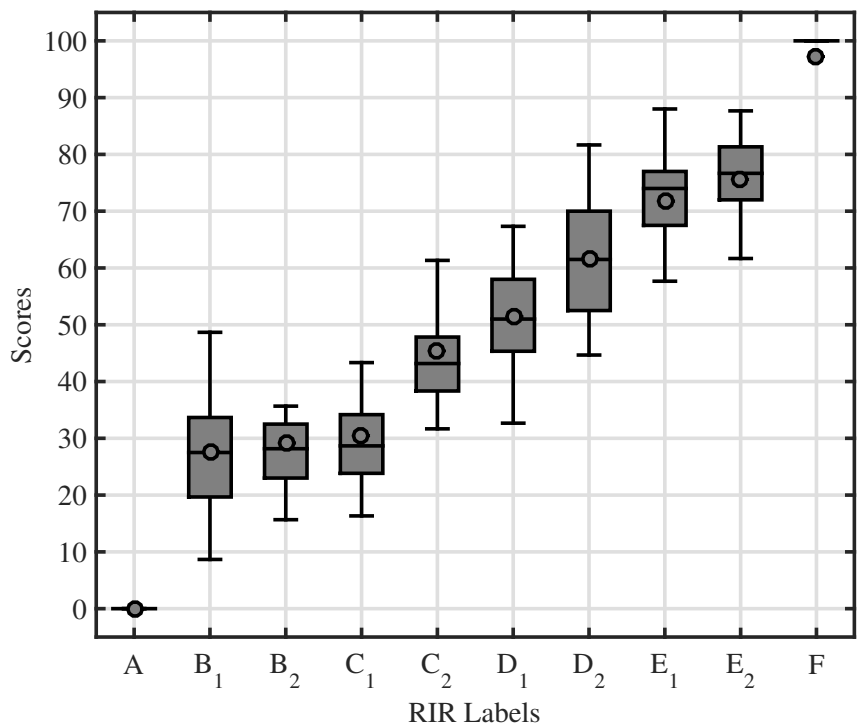


Figure 3.6: Boxplot of mean listener scores for each test condition.

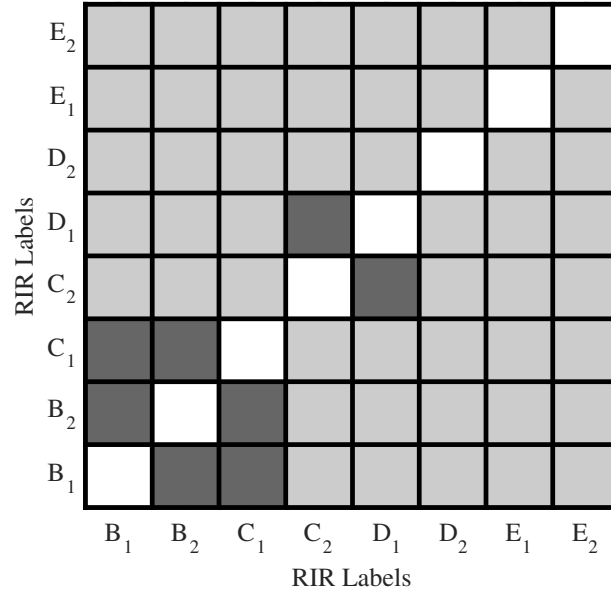


Figure 3.7: Wilcoxon rank-sum test for all pairs except reference and anchor signals. White squares represent unconsidered self-comparisons, light grey squares represent significant differences between pairs and dark grey non-significant differences, where the statistical significance criterion has been set at ($p < 0.05/8$) to account for Bonferroni correction.

are made [81]. This is achieved by scaling down the significance level by the number of multiple comparisons being made (eight in this case).

The results of the Wilcoxon rank-sum test are depicted in Figure 3.7. The plot illustrates that the first three test conditions (B₁, B₂ and C₁), which correspond to rooms with similarly low reverberation times, are not distinguishable to a level of statistical significance by assessors despite their varied DRR values. This supports the notion that the perceived level of reverberation is more influenced by the T₆₀ than the DRR. Interestingly, the only other pair that was not perceived as significantly different were test conditions C₂ and D₁. This case indicates that a speech signal captured in a room with higher T₆₀ can be perceived as equivalently reverberant to a different room with lower T₆₀, if the source-microphone placement is such that the DRR is considerably higher. These results show that to effectively predict the level of perceived reverberation, acoustic parameters such as the T₆₀ and DRR must be considered in combination.

3.4.2 Participant feedback

Upon completing the test, all participants were asked to confirm that, once underway, they had understood the task and were able to complete the test to the best of their ability. A single participant of the 28 who sat the test did not answer in the affirmative, and they were removed from the final analysis as a result. In addition to confirming the above, participants were also invited to share any feedback they had, particularly relating to the difficulty of the test. Great effort was made to ensure participants were not fatigued by a large number of test signals to score. Six participants indicated that the test could be completed relatively quickly and thus additional trials would not have been too cumbersome. The vast majority, however, reported that 10 simultaneous comparisons were manageable and anymore may have been challenging. This largely fits in with the MUSHRA guidelines, which recommends no more than 12 simultaneous comparisons per test trial [69]. Furthermore, as over half the participants had backgrounds in either speech or audio research, interesting observations were also made on the research task at hand. Perhaps the most noteworthy of which was on the possible dependence of the speaker's gender on perceived reverberation. As male and female voices often have different spectral characteristics, the effect of gender-related differences of a speaker's voice on perception (and source signal dependence in general) would be an interesting area of further research.

3.4.3 Evaluation of objective measures

To gain further insight into the parameters most correlated with the perception of reverberation, the Pearson correlation coefficients between participant scores and the considered RIR characteristics is computed and depicted in the top left panel of Figure 3.8. The results demonstrate that the T_{60} gives the highest correlation with perceived reverberation for the chosen speech samples, whilst the DRR and RSV are by contrast less strongly correlated. It should be noted that in a previous experiment by the author [71], the reverse was shown to be true. However, the difference can likely be attributed to the smaller number of test conditions (four), previously used to compute the correlations, from which it is very difficult to obtain a reliable correlation. On the other hand eight conditions

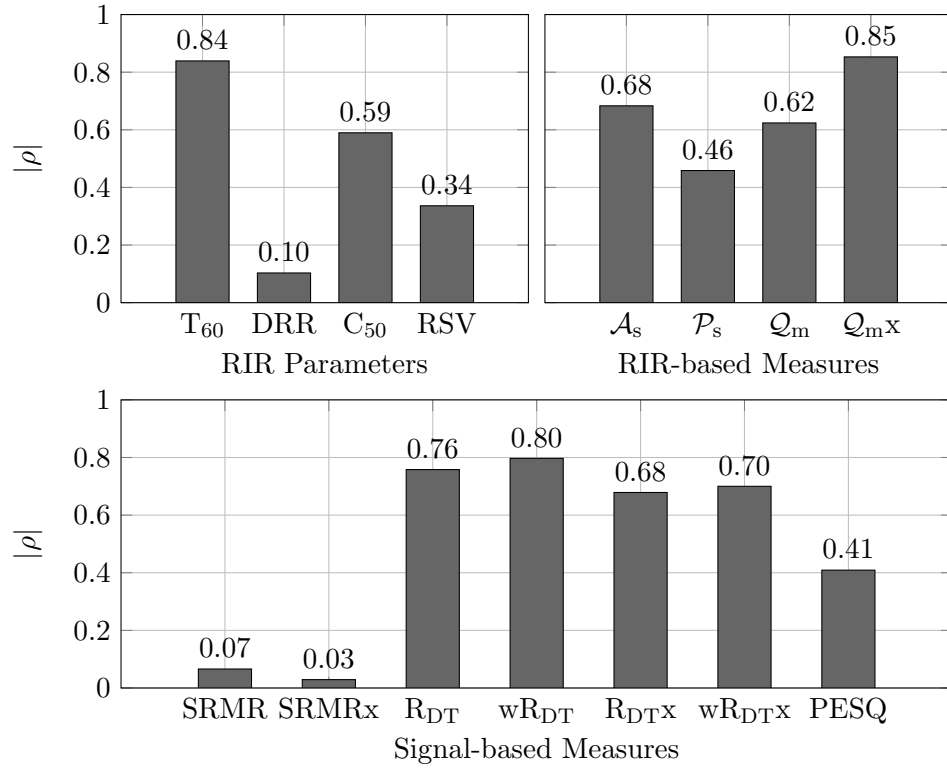


Figure 3.8: Pearson correlation coefficients over all assessed conditions, between mean subjective scores per condition per listener, and the considered RIR characteristics (top left), signal-based instrumental measures (bottom) and RIR-based instrumental measures (top right), where the Allen, Paulus’ and two QAreverb scores are denoted by $\mathcal{A}_s$, $\mathcal{P}_s$, $\mathcal{Q}_m$ and $\mathcal{Q}_{mx}$ respectively.

were tested in this experiment. Moreover the relative importance of the T_{60} on perception compared to the DRR, largely corroborates the findings of other researchers [64].

The performance of reverberation measures, detailed in Section 3.1.1, are also evaluated to determine how well they correlate with subjective assessment. We can see from the top right panel of Figure 3.8 that of the RIR-based measures, Paulus’ score performs comparatively poorly ($|\rho_{\mathcal{P}_s}| = 0.46$). This could be due to the fact that it was modelled from data obtained in a relatively small listening test, in which users were asked to score artificially generated reverberant signals. By contrast, the other measures exhibited stronger correlations. In particular, the QAreverb metric, with weights proposed in [44] and denoted here as $\mathcal{Q}_{mx}$, gave the strongest correlation of all ($|\rho_{\mathcal{Q}_{mx}}| = 0.85$).

Correlations between signal-based reverberation measures and participant scores

are shown in the bottom panel of Figure 3.8. The R_{DT} measure, weighted or unweighted, actually performs better than both R_{DTX} and wR_{DTX} . Perhaps indicating that narrow-band frequency operation contains more cues related to perceived reverberation. However, the weighting proposed does improve correlation in the case of both R_{DT} compared to wR_{DT} and R_{DTX} compared to wR_{DTX} . Therefore, of all the considered signal-based measures, wR_{DT} appears to be the best predictor of the perceived level of reverberation with a Pearson correlation coefficient of 0.80. The wR_{DT} is only outperformed by Q_{mx} , whose computation requires knowledge of the T_{60} and of the DRR.

On the other hand, the SRMR and SRMRx measures show nearly no correlation with the subjective scores from this experimental study. Whilst the performance of the SRMR could be attributed to the high inter- and intra-speaker variability the measure has been known to exhibit [27], the SRMRx had achieved comparatively higher correlations in a previous, albeit smaller, study [71]. Further analysis showed that for the data obtained, the SRMRx displayed higher correlation with the DRR compared to the T_{60} . Given the larger influence the latter has on the perceived level, this may explain the poor performance achieved. It is interesting to note that as part of an ACE challenge entry, variants of the SRMR were used to estimate different acoustic parameters [82]. Therefore, much like the weighting scheme introduced for the R_{DT} improved its performance, adjusting the operating parameters of the SRMR could similarly be used to develop a variant that is better correlated with perception. Finally, relatively weak correlations are obtained from the PESQ measure, with a Pearson correlation coefficient of 0.41 highlighting its limitations as a predictor of the level of perceived reverberation.

3.4.4 Perceptual mapping

Whilst improved correlation with subjective assessments is a welcome development, the output of the wR_{DT} measure considered on its own may not be particularly meaningful. The utility of the wR_{DT} could therefore be improved by applying a non-linear mapping function to the output of the measure, so as to essentially convert it into a score, as is commonly done with other objective measures [38, 83]. To this end, we propose a non-

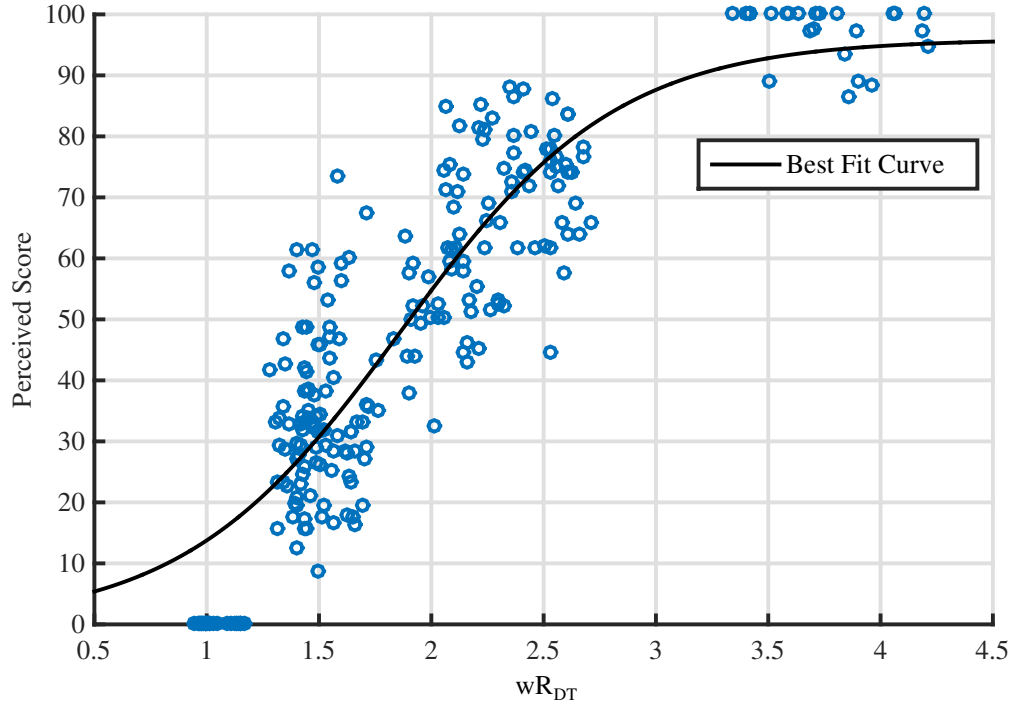


Figure 3.9: Scatter plot of mean scores per listener (average taken across the three trials) for each reverberant test condition, against wR_{DT} . Fitted curve, ζ , modelled as $\zeta = c_0 \cdot (1 + \exp(c_1 \cdot wR_{DT} + c_2))^{-1}$, with $(c_0, c_1, c_2) = (95.9, -2.1, 3.86)$

linear mapping based on the logistic function or sigmoid curve, similar to the one used in PESQ. The mapping takes the form:

$$w\tilde{R}_{DT} = \frac{c_0}{1 + e^{(c_1 \cdot wR_{DT} + c_2)}}, \quad (3.8)$$

where the parameters c_0 , c_1 and c_2 are selected to ensure, in the minimum mean square error sense, the best fit curve between wR_{DT} and the subjective assessments obtained from the listening test across all test conditions. This is illustrated in Figure 3.9. The mapped output, which we denote as $w\tilde{R}_{DT}$, thus produces scores approximately in the range of 0-100, where the two extremes roughly represent the perceived level of the reference and anchor test conditions presented in this test. As the mapping function is approximately linear in the test signal range (conditions B_1 to E_2), the proposed mapping does not significantly impact the measures correlation with the listening test data; the Pearson

correlation coefficient between $\tilde{wR}_{DT}$ and subjective scores, to two significant figures, stands at 0.80.

3.5 Conclusion

The results of an experimental listening test to assess the perceived level of reverberation in speech captured by a single microphone were presented. The relation between the subjective scores and RIR characteristics showed that the T_{60} has a larger influence than the DRR on reverberation perception, but that both need to be considered in combination to predict the perceived level of reverberation. Of the considered instrumental measures, the QAreverb of [44] showed the highest correlation with participant scores, with a Pearson correlation coefficient of 0.85. However, the proposed wR_{DT} measure also performed well with a correlation coefficient of 0.80, giving comparable performance. Modified to weight individual features more heavily, the signal-based measure has the advantage of not requiring knowledge of the RIR or of how this RIR would be altered by a dereverberation algorithm. Combined with the non-linear mapping introduced, which converts the output of the measure to a perceptual score roughly in the range of 0-100, the proposed $\tilde{wR}_{DT}$ measure is a potentially useful evaluation tool for dereverberation research.

Chapter 4

Spherical microphone array acoustic rake receivers

Having discussed the detrimental impact of reverberation on speech processing applications in the preceding chapters, we now address the task of dereverberation. To this end several spatial filter designs, which we refer to as acoustic rake receivers, are proposed and developed in this chapter. Motivated by the rake receiver concept in wireless communications, the designs exploit strong early reflections along with the direct path of propagation, in order to perform dereverberation.

We begin this chapter by providing an overview of existing dereverberation methods, highlighting the relative merits of one over another. In particular we review efforts in the literature to develop acoustic equivalents to the rake receiver, which we broadly classify as techniques that utilise early reflections, rather than suppressing or ignoring them, for improved signal acquisition. Having established the limited work in this area, we proceed to propose several acoustic rake receiver designs that are implemented on a spherical microphone array. The choice of a spherical array is discussed and the relevant prerequisite background theory that underpins their operation is summarised.

The proposed designs are then validated through experimental simulation, using a range of evaluation metrics; namely reverberation measures based on the acoustic impulse response, the wR_{DT} metric proposed in the preceding chapter and scores obtained from a

baseline ASR system. The conducted simulations demonstrate that by selectively incorporating additional reflections in our spatial filter designs, significant levels of reverberation suppression are possible.

4.1 Review

Dereverberation techniques to date have tended to focus on either spatial filtering approaches [8] or methods that estimate and suppress the reverberant signal component according to a model of the reverberation process. Of the latter methods, single channel spectral subtraction [84, 85] is perhaps the most straightforward. It employs a statistical model of late reverberation, which is then subtracted from the reverberant signal, in an effort to arrive at a dereverberated signal. However, as is the case with spectral subtraction in classical de-noising contexts, the approach tends to suffer from musical noise.

At the other extreme, acoustic channel equalisation methods can theoretically give perfect dereverberation. Such approaches use knowledge of the room impulse response between a source and microphone to construct inverse filters that undo the effects of the channel. Single-channel equalisation methods, as shown by Neely in [86], are however limited in use as the conditions for inverting an individual RIR are rarely met in practice. Multichannel approaches, such as the well known multiple-input/output inverse theorem (MINT), on the other hand, can provide exact inverse filters with less stringent conditions that can be more easily met [87].

However, despite their promise, equalisation methods still require estimating the RIRs between sources and microphones with a degree of accuracy that is usually not possible in practice. It is therefore desirable to develop techniques which are robust to channel estimation errors. Whilst strides have been made to develop such methods, primarily through channel shortening techniques that relax the conditions required of the inverse filters [88–90], more research is required before such systems have practical applications.

Of the existing techniques, the system of Delcroix *et al.* [91] performed particularly well in a recent dereverberation challenge [7]. Composed of several cascaded modules, in-

cluding a deep neural network (DNN)-based ASR back-end, the system contained a speech dereverberation front-end based on multichannel linear prediction (MCLP) [92, 93]. By exploiting the fact that reverberation results in long-term time correlation of the reverberant speech, the signal-based MCLP method predicts the late components of reverberation through iterative estimates of a subband autoregressive process. The estimates are then subtracted from the reverberant speech, also in the STFT domain, to arrive at a dereverberated signal. Although capable of achieving significant reverberation reduction, MCLP-based methods can be computationally intensive. Moreover, in multichannel contexts they rely on similar assumptions to multichannel estimation techniques, namely that the RIRs do not share common zeroes. Although the latter is not particularly problematic in practical applications, implementation of MCLP methods can be computationally prohibitive, particularly when operating in the time domain. Whilst the systems of [91, 93] have proposed subband approaches that naturally lend themselves to efficient parallel implementation, it has been noted in [94] that this comes at the cost of losing information on the spectral structure of speech. Practical implementations of MCLP-based approaches thus remain an active area of research [94–96].

In contrast to the above methods, spatial filtering or beamforming-based approaches only require estimates of the direction of arrival of the desired source [8]. The spatial diversity provided by a microphone array is exploited by a beamformer to selectively amplify sounds originating from a particular ‘look’ direction, whilst attenuating sound from all other directions. Given the diffuse nature of late reverberation, in which reflections arrive with equal probability from all directions, beamformers can thus achieve substantial levels of dereverberation [9]. However, by focusing only on the direct path, most conventional beamformers implicitly ignore the contributions of strong early reflections.

A rake receiver, on the other hand, aims to exploit multipath propagation for improved signal acquisition. Essentially, this is achieved by using strong signal reflections to reinforce the direct path component. It can therefore be thought of as collecting the reflections (much like a garden rake collects leaves). Although a well established and popular technique in wireless communications [10], developing equivalent designs in the

acoustic domain has attracted less interest from researchers. Whilst the term acoustic rake receiver was first used in [97], the concept can be traced back to earlier work by Flanagan *et al.* [98]. Using knowledge of the image source locations up to third order, independent beams (‘fingers’) were directed at each of the reflections and summed, having first time aligned the image source signals with the direct path signal. In subsequent work [99], it was shown that this operation is conceptually similar to applying a matched filter to each channel of the array, given prior knowledge of the RIRs. In an effort to enhance the practical applicability of [98], the effect of wall-impedances on signal reflections was modelled and compensated for in the simulations conducted in [100].

Although not referred to as such by the authors, an acoustic rake receiver was practically demonstrated in [101]. Implemented on a spherical microphone array, the design collected early reflections by using a steered response power map to identify dominant DOAs. Beams were then pointed in these directions before being summed coherently, with interbeam delays determined using a modified adaptive eigenvalue decomposition method. More recently, the concept of an acoustic rake receiver was explored by Dokmanic *et al.* in [11, 102]. Alternative formulations were presented in great mathematical detail and shown to optimise different performance metrics in experimental simulations of two-dimensional rooms. However, similarly to [101], the designs focused primarily on utilising early reflections to improve the overall signal-to-noise-plus-interference ratio (SNIR) in the presence of white noise and/or in scenarios where there was a single static interfering source (a recording of an opera singer). Assuming exact knowledge of reflection DOAs and the noise and interferer statistics, several of the proposed signal-based designs were shown to lead to improvements in SNIR as the number of reflections considered increased. A particularly illustrative example highlighted by the authors was when an interferer might be occluding the desired speech source, a situation the proposed rake designs combatted by generating a spatial filter that would only use the source reflections. The use of an acoustic rake as a means of dereverberation, by contrast, has not yet been explored.

The acoustic rake receivers proposed in this work are implemented on a spherical microphone array (SMA), that is, an array made up of microphones placed in a spherical

configuration. Whilst the designs can be generalised to any arbitrary array, SMAs possess several intrinsic advantages over the more conventional linear and planar array setups. Consequently, interest in the design, operation and application of SMAs, has steadily grown within the research community over the last 15 years. Meyer and Elko were among the first researchers to investigate their use, presenting a beamformer design based on a rigid array in [103]. Soon after, Abhayapala and Ward demonstrated an open sphere design to record the higher order harmonics of a sound field [104], information they believed could facilitate the accurate reproduction of a three-dimensional sound field.

Since these seminal works, further research into the design properties of SMAs and their operation has been conducted, primarily by Rafaely [14, 105–107]. Additionally, spherical arrays have been used for a wide range of speech and audio processing applications. These include, but are not limited to: sound field analysis [108–110], beamforming [111–114], speech enhancement [115, 116], source localisation [117–119] and robot audition [120]. For the purposes of the rake designs proposed in this work, the choice of spherical microphone arrays is motivated primarily by their ability to produce three-dimensional beampatterns independent of look direction, and for the elegant mathematical framework they provide for processing three-dimensional sound fields. The brief literature survey presented is therefore deemed sufficient. However, for a more detailed overview, the interested reader is referred to two recently published books on the topic [12, 121].

4.2 Spherical harmonic domain processing

While the field of spherical microphone array processing is relatively new, the mathematical framework employed is, by comparison, well established. In this section the theory that underpins SMA operation is briefly reviewed, with a focus on the decomposition of a sound field for processing in the spherical harmonic domain (SHD). The signal model and beamformers considered in this work are then formulated and presented in the SHD.

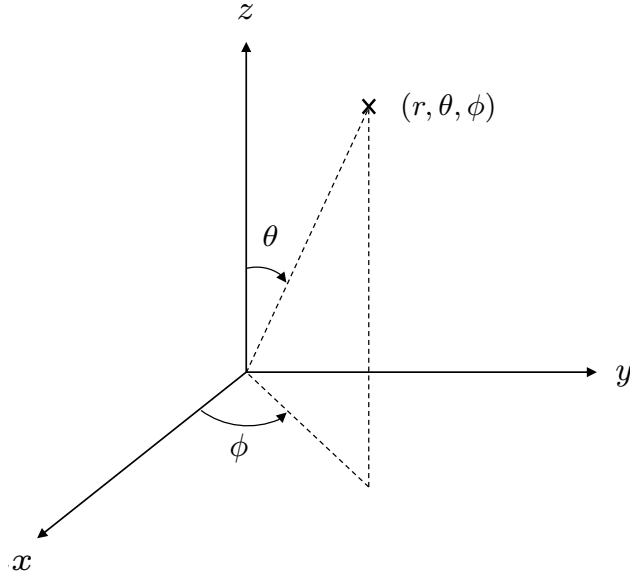


Figure 4.1: The spherical coordinate system as defined and used in this thesis. The radial distance r is the distance between the observation point and the origin of the coordinate system. The inclination angle θ is measured from the positive z -axis, and the azimuth angle ϕ is measured in the xy -plane from the positive x -axis.

4.2.1 Sound field decomposition

A sound pressure field impinging a sphere, $P(k, \mathbf{r})$, can be conveniently expressed as a weighted sum of spherical harmonic basis functions,

$$P(k, \mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l P_{lm}(k, r) Y_{lm}(\theta, \phi), \quad (4.1)$$

where k is the wavenumber, $\mathbf{r}=(r, \theta, \phi)$ the position on the sphere in spherical coordinates (radial distance, inclination and azimuth angles respectively) and Y_{lm}^* the complex conjugate of a spherical harmonic, of order l and degree m . The coefficients P_{lm} , which can be considered analogous to the Fourier series coefficients in one-dimension, are the respective spherical harmonic weights required to reproduce the sound field. Generally referred to as eigenbeams, they can be calculated via the spherical Fourier transform (SFT)

$$P_{lm}(k, r) = \int_0^{2\pi} \int_0^\pi P(k, \mathbf{r}) Y_{lm}^*(\theta, \phi) \sin(\theta) d\theta d\phi, \quad (4.2)$$

It is worth noting that the spherical harmonic expansion expressed in (4.1), is sometimes also referred to as the inverse spherical Fourier transform (ISFT).

Constituting a set of solutions to the wave equation in spherical coordinates, the spherical harmonic of order l and degree m (satisfying $|m| \leq l$) is defined as [122, eqn 6.20]

$$Y_{lm}(\theta, \phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} \mathcal{P}_{lm}(\cos(\theta)) e^{im\phi}, \quad (4.3)$$

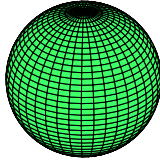
where $\mathcal{P}_{lm}$ is the associated Legendre function, and i is the imaginary number ($i = \sqrt{-1}$). By definition a complex function, the magnitude of several spherical harmonics, from zeroth to third order, are plotted for illustrative purposes in Figure 4.2. It can be seen that the zeroth order harmonic is omnidirectional, whilst higher orders possess more sophisticated directivity patterns.

In practice, it is not possible to measure a sound field across the entire surface of a sphere. Instead the field must be spatially sampled using a finite number of microphones, Q , placed on a spherical array. In such a scenario the SFT from (4.2) can be approximated to a discrete summation,

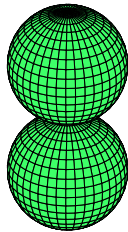
$$P_{lm}(k, r) \approx \sum_{q=1}^Q \Theta_q(k) P(k, \mathbf{r}_q) Y_{lm}^*(\theta_q, \phi_q), \quad (4.4)$$

where $\mathbf{r}_q = (r_q, \theta_q, \phi_q)$, represents the position of the q -th microphone in spherical coordinates, and the quadrature weights $\Theta_q(k)$ are chosen to ensure the expression is a good approximation for a given spatial sampling configuration [14]. The number of microphones also limits the maximum sound field order, L , that can be decomposed without significant spatial aliasing. A minimum of $(L+1)^2$ microphones are required to sample a sound field of order L , with the exact number dependent on the sampling scheme used. In practice, captured sound signals are not order-limited and are instead represented by an infinite series of spherical harmonics. However, eigenbeam magnitudes rapidly decay for $l > kr$. Practically, this means a wideband signal captured by a spherical array will suffer negligible aliasing, if it is first sufficiently band limited. Namely, the upper operating frequency f_{max} must satisfy $f_{max} < \frac{Lc}{2\pi r}$, where c corresponds to the speed of sound [106].

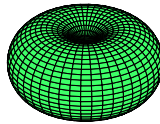
$$|Y_{00}(\theta, \phi)|$$



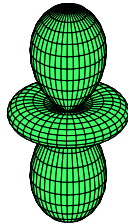
$$|Y_{10}(\theta, \phi)|$$



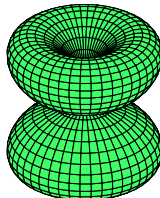
$$|Y_{11}(\theta, \phi)|$$



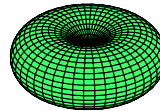
$$|Y_{20}(\theta, \phi)|$$



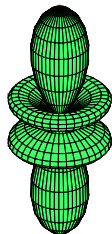
$$|Y_{21}(\theta, \phi)|$$



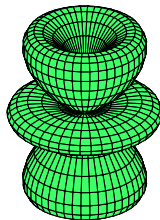
$$|Y_{22}(\theta, \phi)|$$



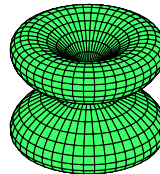
$$|Y_{30}(\theta, \phi)|$$



$$|Y_{31}(\theta, \phi)|$$



$$|Y_{32}(\theta, \phi)|$$



$$|Y_{33}(\theta, \phi)|$$

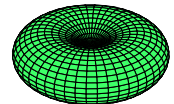


Figure 4.2: Magnitude of the spherical harmonics for order $l = 0, 1, \dots, 3$, and degree $m \leq l$. The magnitude plots for $m < 0$ are omitted as they are identical to those for $m > 0$.

4.2.2 Signal model

A sound signal produced in a reverberant environment and captured by a spherical microphone array, can be expressed as a weighted sum of planewaves arriving at the array from different directions. A unit-amplitude planewave, A , arriving from the direction $\Omega_u = (\theta_u, \phi_u)$, can be mathematically described in the spherical harmonic domain as

$$A_{u,lm}(k, r) = b_l(k, r) Y_{lm}^*(\Omega_u), \quad (4.5)$$

where the mode strength, b_l , is a function that describes the effect of the array configuration (e.g. the size and type of array) on the captured signal.

It follows that the sound pressure field can be expressed in the spherical harmonic domain, as a summation of planewaves, each with its own respective amplitude and phase [115]. Namely,

$$P_{lm}(k, r) = \sum_{u=0}^{\mathcal{U}} \psi_u(k) A_{u,lm}(k, r) S(k), \quad (4.6)$$

where $S(k)$ is the desired source signal, and the term $\psi_u(k)$ represents the amplitude and phase of the u -th planewave. Although modelling a reverberant signal accurately requires $\mathcal{U}$ to be large (or even infinite), in our work we consider a signal model composed of only the direct path and a maximum of six additional reflections ($\mathcal{U} \leq 6$). Although in practice this will result in model mismatch, if the reflections considered are sufficiently strong, the simplified formulation is suitable for our intended application of dereverberation.

4.2.3 Spherical harmonic domain beamforming

The output of a spherical harmonic domain beamformer can be expressed as a weighted sum of eigenbeams,

$$Z(k, r) = \sum_{l=0}^L \sum_{m=-l}^l W_{lm}^*(k) P_{lm}(k, r), \quad (4.7)$$

where W_{lm} represents the beamformer weights in the SHD. The beamformer output can be more conveniently expressed using vector notation,

$$Z(k, r) = \mathbf{w}^H \mathbf{p}, \quad (4.8)$$

where $(\cdot)^H$ denotes the Hermitian transpose,

$$\mathbf{w} = [W_{00}, W_{1(-1)}, W_{10}, W_{11}, \dots, W_{LL}]^T \quad (4.9)$$

and

$$\mathbf{p} = [P_{00}, P_{1(-1)}, P_{10}, P_{11}, \dots, P_{LL}]^T \quad (4.10)$$

are the $(L+1)^2 \times 1$ vectors of the spherical harmonic array weights and sound pressure fields respectively. Note that the dependencies on wavenumber and array radius are omitted for brevity.

As in the spatial domain, beamformers can be designed in the spherical harmonic domain by selecting array weights to produce a desired directivity pattern, where the weight selection is either signal dependent or independent. In the latter, the beamformer weights depend only on the DOA of the source of interest. In the former statistical properties of the desired signal, and/or the noise field, are also taken into account to achieve statistically optimal noise reduction and minimal speech distortion. Array weights can also be chosen to maximise or minimise additional design criteria such as robustness, sidelobe level, white noise gain and so on [112].

The planewave decomposition (PWD) beamformer is a well known signal independent beamformer in which a maximum directivity beam is steered at a specified look direction Ω_u . The beamformer weights are chosen by compensating for the array mode strength:

$$W_{u,lm}^*(k) = \frac{Y_{lm}(\Omega_u)}{b_l(kr)}. \quad (4.11)$$

The spatial response, or beampattern, of the PWD beamformer becomes increasingly directive as the array order increases. A cross-section of the beampatterns of a zeroth-,

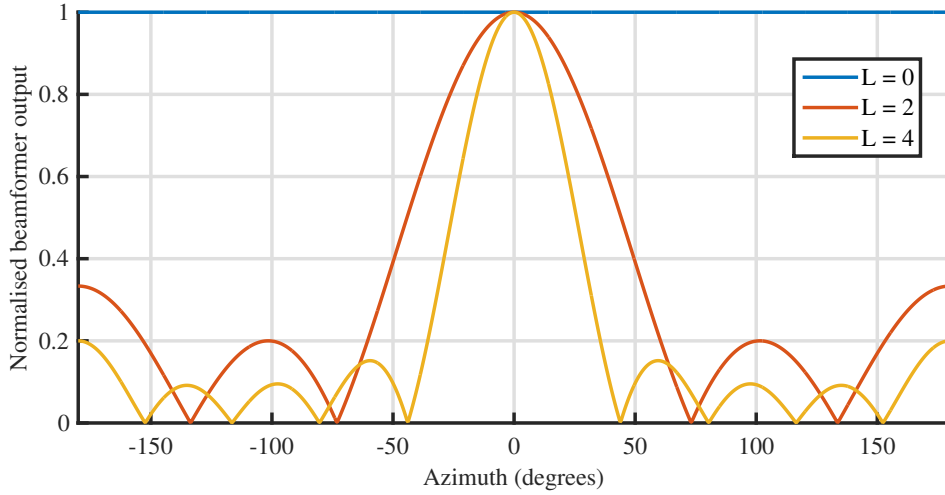


Figure 4.3: Cross-section of the normalised beampattern of a PWD beamformer with a look direction $\Omega_u = (90^\circ, 0^\circ)$, as a function of the beamformer order L and ϕ , for $\theta = 90^\circ$.

second- and fourth-order PWD beamformer are illustrated in Figure 4.3. Pointing in the same arbitrary direction, the plots show the increased directivity as the order increases. The amount of ripple, or oscillation, also increases with the order but it is worth noting that these oscillations are of smaller amplitude. Although rotationally symmetric about the look direction, three-dimensional beampatterns of the second- and fourth-order PWD beamformer have been included for illustrative purposes (Figure 4.4). As the array order approaches infinity, the beamformer output tends towards a Dirac delta function. At which point, the sound field is decomposed into a single planewave originating from Ω_u , making clear where the beamformer derives its name.

However, as described in 4.2.1, the order which a SMA can operate at is constrained by spatial aliasing considerations and the number of microphone elements that make up the array. For the purposes of our proposed rake designs, which aim to isolate and capture individual multipaths, the sidelobe levels of a finite-order PWD could amplify unwanted reflections. The realisation of a beamformer capable of pointing nulls is therefore an important objective. As a result we consider the approach of [123], in which weights are chosen to satisfy a distortionless response as well as multiple null constraints in specified

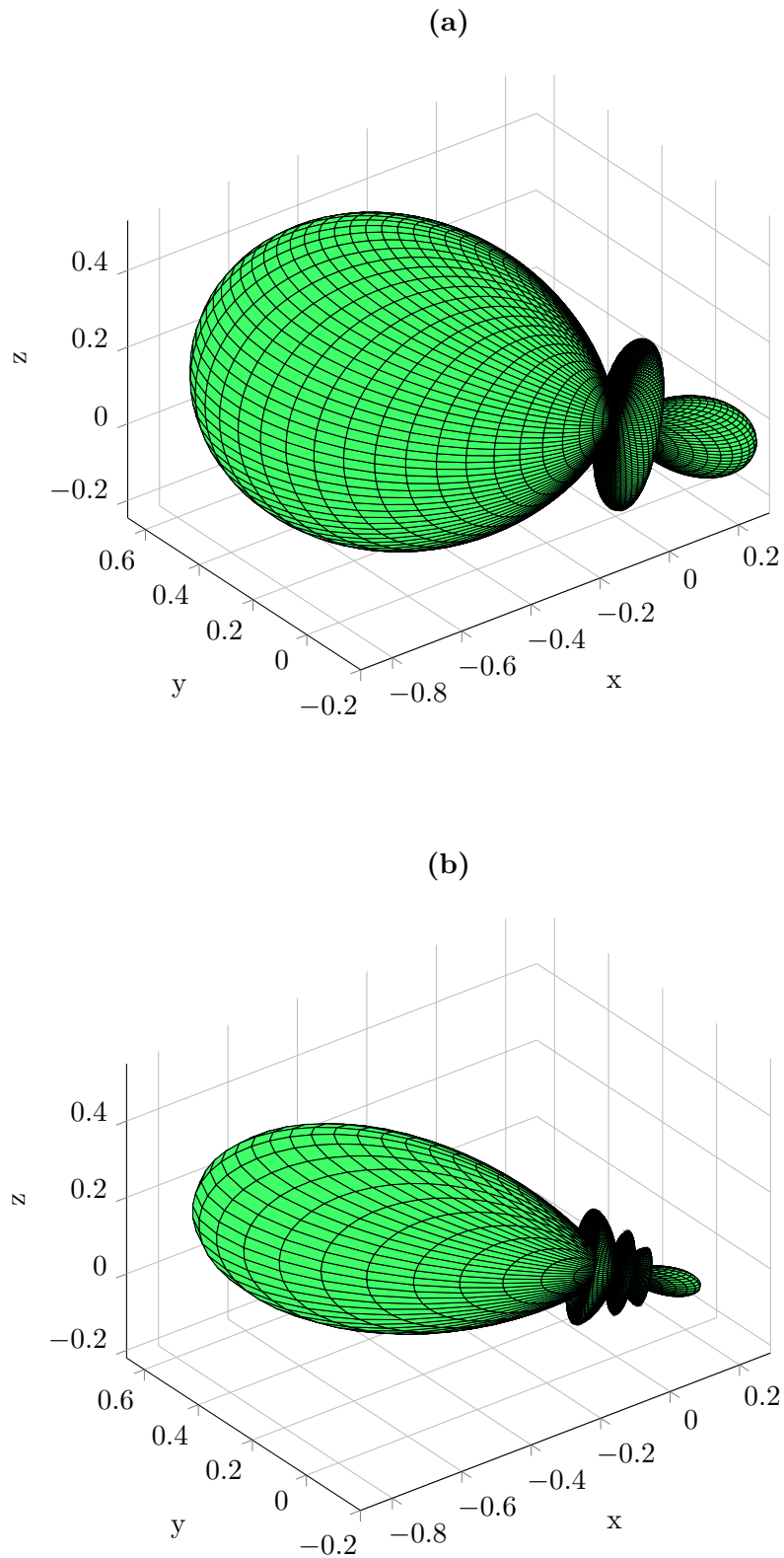


Figure 4.4: Three-dimensional beampatterns of (a) second-order and (b) fourth-order PWD beamformer steered to an arbitrarily selected look direction.

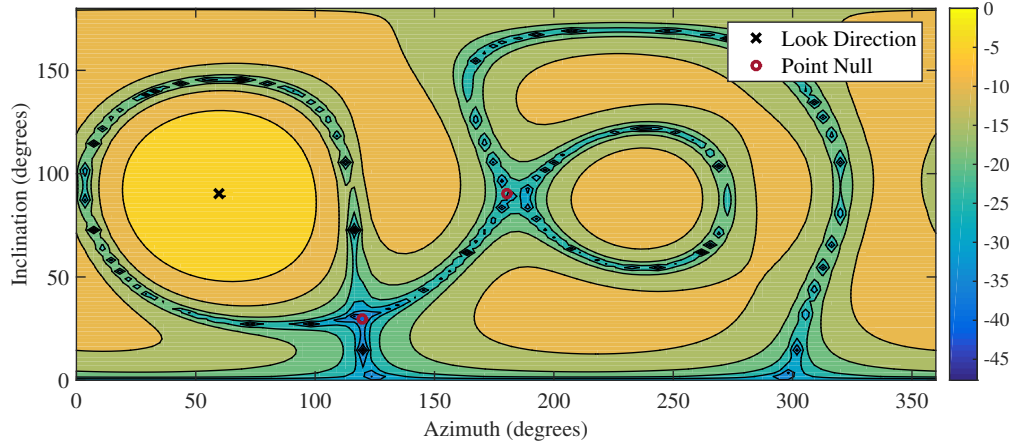


Figure 4.5: An example beampattern of a null steering beamformer, consisting of a single beam, and two nulls.

look directions. The constraints can be expressed as

$$W_{u,lm}^*(k) A_{l,nm}(kr) = \delta_{\iota u}, \quad (4.12)$$

where $\delta_{(\cdot)}$ denotes the Kronecker delta and $A_{u,lm}$ is the array response to a planewave from direction Ω_u , as described in (4.5). By setting $\iota = u$, where $u = (0, \dots, \mathcal{U})$, a distortionless response constraint in the u -th direction is established, along with $\mathcal{U}$ number of null constraints in the remaining specified directions. A least squares solution can then be formulated to select array weights which satisfy the above constraints. The resulting beampattern produced for an example case, in which a beam and two nulls are steered in arbitrary directions, is shown in Figure 4.5.

4.3 Spherical acoustic rake designs

In the context of signal acquisition in reverberant environments, an acoustic rake receiver that captures and combines individual multipaths constructively, introduces dual benefits. The first is reinforcement of the direct path component. The second is the attenuation of the reverberant decay tail, for which suppression under their summation is due to the incoherent nature of late reverberation. Formulated in the spherical harmonic domain,

the rake designs proposed in this section operate under the assumption that the directions and time of arrival (TOA) of the direct path and a number of strong early reflections is known.

In practice, steered response power maps can be used to estimate direction of arrival information of strong reflections [101, 124], whilst relative time delays can be computed using established techniques [125], a popular example being correlation based approaches [126]. Alternatively, in some scenarios knowledge of room geometry can be exploited in combination with source localisation methods, to determine image positions and time delays using geometric rules [52]. In this work, we assume oracle knowledge of this information (which we can determine exactly through the experimental simulations conducted).

4.3.1 Delay-and-sum rake

The first rake receiver design proposed is the spherical delay-and-sum rake, or Rake-DS. In this intuitive implementation a series of beamformers, or rake fingers, are constructed to point a single PWD beam in the direction of individual sources - the direct path (true source) and a chosen number of early reflections (image sources). This is conceptually illustrated in Figure 4.6, in which a plan view of the Rake-DS collecting the direct path and two additional multipaths is demonstrated in a room for a given source-array setup. Knowledge of the relative time lags between multipaths is then used to apply the appropriate phase correction, $\varphi(\tau_u)$, so that the rake fingers are time aligned and combine together constructively (an idealised illustration of this operation is shown in Figure 4.7). The Rake-DS output is expressed as

$$\begin{aligned} Z_{DS}(k, r) &= \sum_{u=0}^{\mathcal{U}} Z_u(k, r) \cdot e^{\varphi(\tau_u)} \\ &= \sum_{u=0}^{\mathcal{U}} \mathbf{w}_{u,\text{pwd}}^H \mathbf{P} \cdot e^{\varphi(\tau_u)}, \end{aligned} \tag{4.13}$$

where $\mathbf{w}_{u,\text{pwd}}$ denotes the PWD array weight vector that points a beam towards Ω_u , the direction of the u -th multipath. The weights are computed according to (4.11). Individual

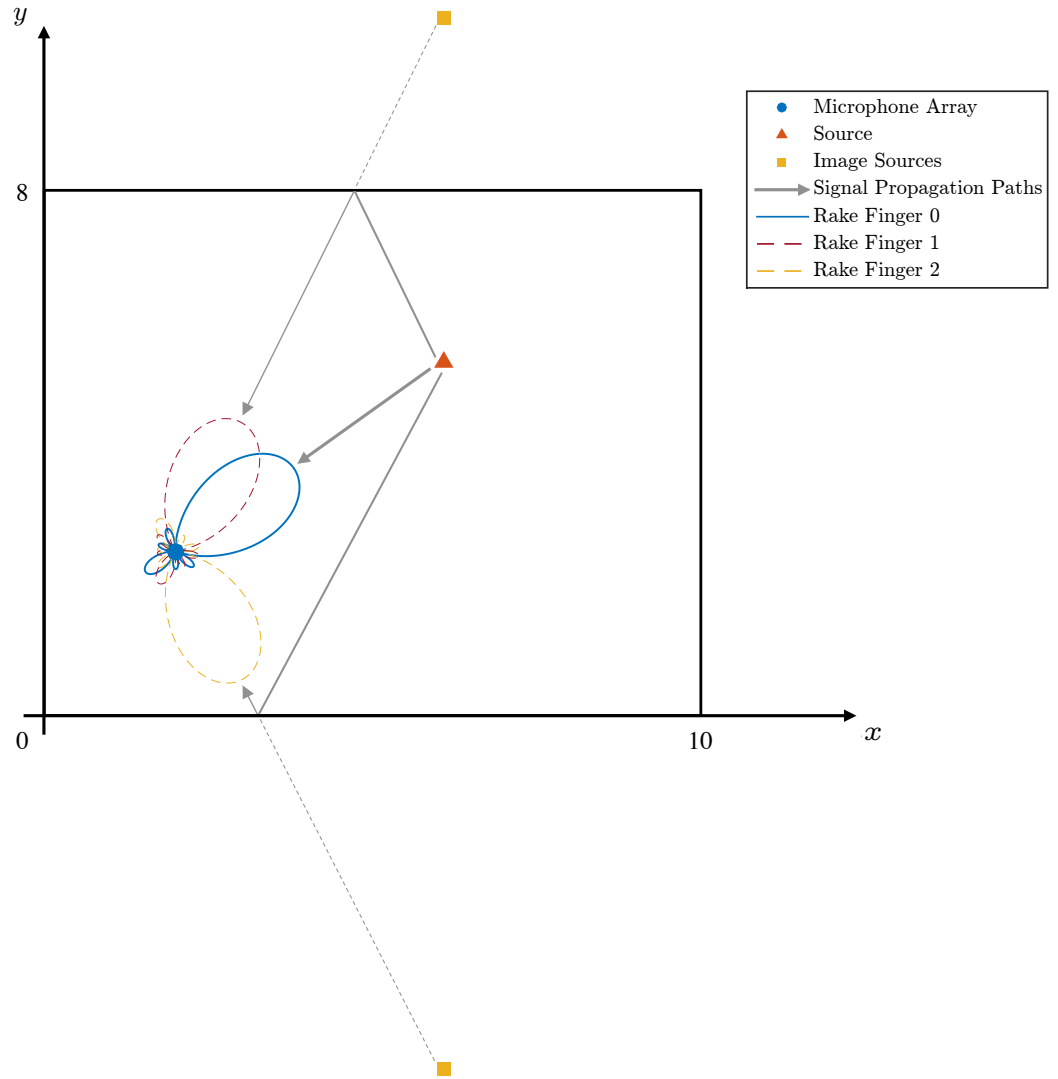


Figure 4.6: Conceptual operation of a Rake-DS design, operating in room B for an arbitrarily selected source-array placement. Rake fingers consist of PWD beamformers, one pointed in the direction of the direct path (source) and two pointed towards first-order reflections (image sources).

rake fingers are multiplied by complex exponentials to time align them with the $\mathcal{U}$ -th path, in order to combine them constructively. This results in the output of rake fingers being delayed in time, a practical step chosen because it allows for real-time operation (the reverse, that is advancing the outputs in time, by contrast is not practically feasible in real-time). The phase correction, $\varphi(\cdot)$, is a function of the TOA of the u -th multipath, τ_u . It is described as

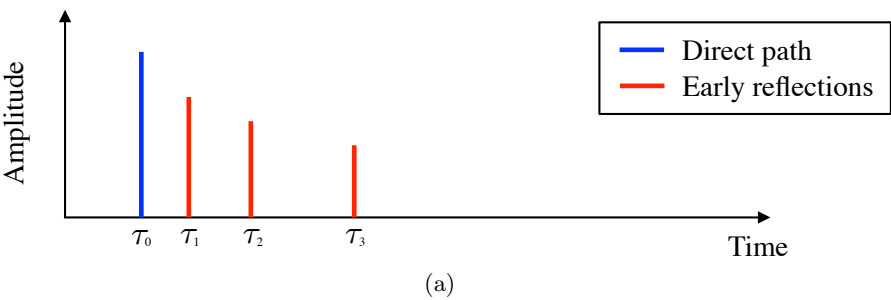
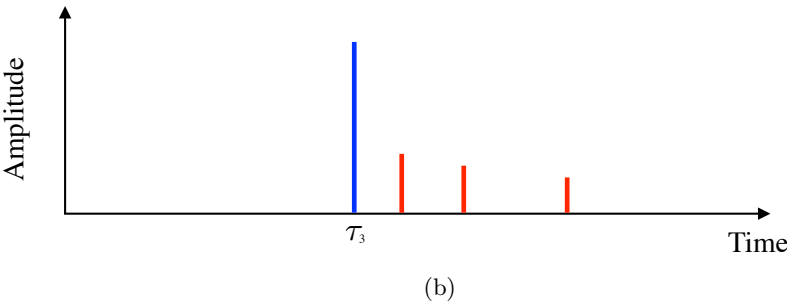
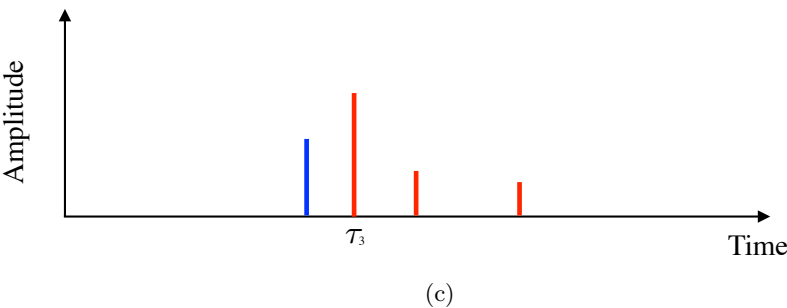
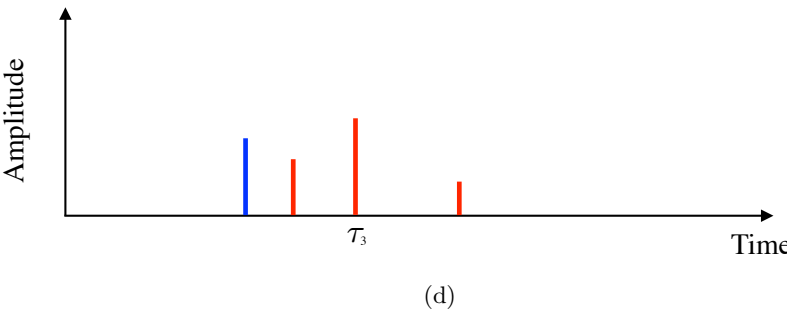
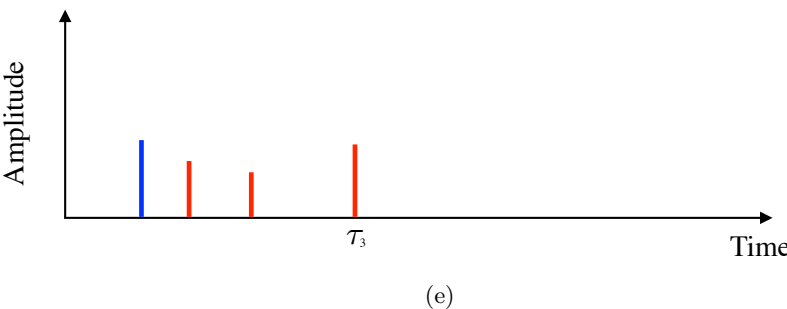
$$\varphi(\tau_u) = i k c (\tau_{\mathcal{U}} - \tau_u). \quad (4.14)$$

4.3.2 Pre-echo null rake

The second rake design proposed is the spherical pre-echo null rake, or Rake-PN. This design differs from the Rake-DS in that the rake fingers designed to point beams at individual early reflections, also steer nulls in the direction of any preceding multipaths, up to and including the direct path. The incorporation of nulls in this design is motivated purely by perceptual concerns; to avoid audible pre-echoes that would be introduced in the Rake-DS design (Figure 4.7f). The Rake-PN can be expressed as

$$Z_{PN}(k, r) = \sum_{u=0}^{\mathcal{U}} \mathbf{w}_{u,\text{null}}^H \mathbf{P} \cdot e^{\varphi(\tau_u)}, \quad (4.15)$$

where $\mathbf{w}_{u,\text{null}}$ denotes the null steering beamformer array weights, which point a beam towards the u -th multipath, and nulls at the $0, \dots, (u-1)$ -th paths. The weights are the least square solution to (4.12).



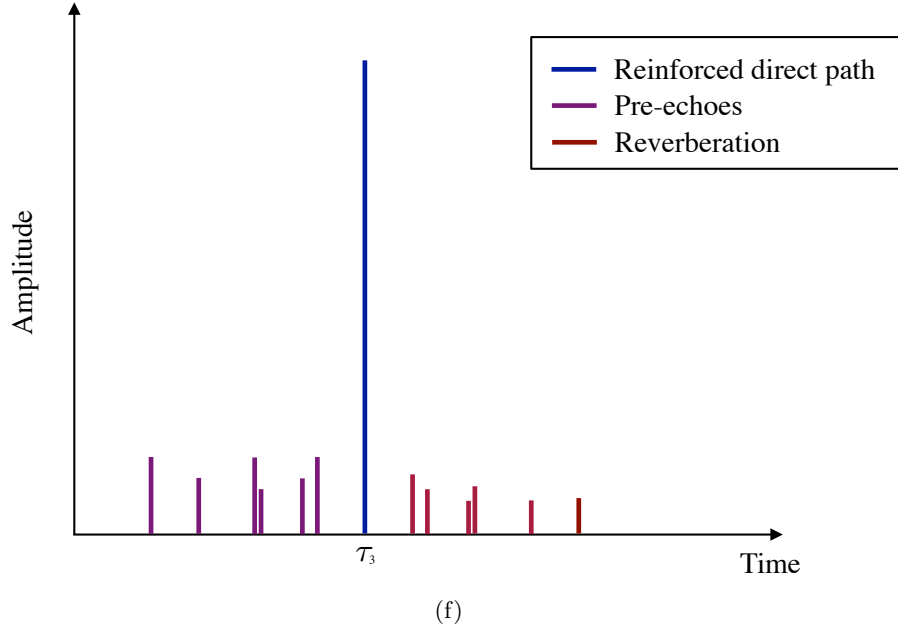


Figure 4.7: An idealised diagram showing the output of the delay-and-sum rake and its individual fingers, in response to a specific input RIR (a) consisting of a direct path signal and three early reflections. Assuming the multipath DOAs are known and that it is possible to produce a perfectly directive beamformer: (b) is the zero-th rake finger output, which amplifies the direct path and is appropriately delayed in time. The first to third rake fingers apply the same relative gain to their respective multipaths (c)-(e). The rake fingers are then combined in (f), leading to higher DRR as well as the introduction of pre-echoes.

4.3.3 Weighted pre-echo null rake

The final rake design proposed is the spherical weighted pre-echo null rake, or wRake-PN. The design differs from the Rake-PN in that the contribution of individual rake fingers is controlled. In order to satisfy the multiple null-constraints, high magnitude beamformer weights may be required to synthesise the desired beampatterns; for example, when a beam and a null are in close proximity. Weighting rake fingers based on the inverse of the beamformer energies, $\|\mathbf{w}_{u,\text{null}}^{\mathbf{H}}\|^2$, required to produce them, gives

$$Z_{wPN}(k, r) = \sum_{u=0}^U \frac{1}{\|\mathbf{w}_{u,\text{null}}^{\mathbf{H}}\|^2} \mathbf{w}_{u,\text{null}}^{\mathbf{H}} \mathbf{P} \cdot e^{\varphi(\tau_u)}, \quad (4.16)$$

the wRake-PN thus de-emphasises fingers which would otherwise actually amplify the reverberant sound field. It is worth noting that numerous weighting scheme exponents can be considered. In this thesis body, wRake-PN denotes an exponent of 2, however in Appendix B we investigate the impact of choosing higher and lower exponents on overall performance.

4.3.4 Rake finger selection

For the rake designs described, we also investigate the impact that the choice of reflections to rake has on dereverberation performance. Specifically, we evaluate performance as a function of the number and combination of first-order reflections that are raked with the direct path. Additionally, we also determine whether, as the number of reflections raked increases, it would be beneficial to select reflections based on their energy rather than if they were first-order reflections. Note that in the context of this work, first-order reflections refer to signal multipaths that arrive at the microphone via a single reflective surface. On the other hand, the energy of a reflection corresponds to its amplitude, and is a function of the reflection order as well as the propagation distance and absorption coefficients of the surfaces it reflects off. Unless stated otherwise, of the schemes tested, first-order reflections are selected based on their TOA, whilst for the energy based selection, the strongest reflections are selected, irrespective of their order or TOA. Asterisks are appended to a rake design name to indicate the selection criterion used. A single asterisk (*) represents selection based on first-order reflections, whilst a double asterisk (**) represents one based on reflection energy.

4.4 Experimental simulations

4.4.1 Setup

The dereverberation performances of the acoustic rake receivers proposed in this work were evaluated in two simulated rooms, hereafter referred to as ‘room A’ and ‘room B’, respectively. The former had dimensions of 5 x 4 x 3 m, the size of a typical office, with a

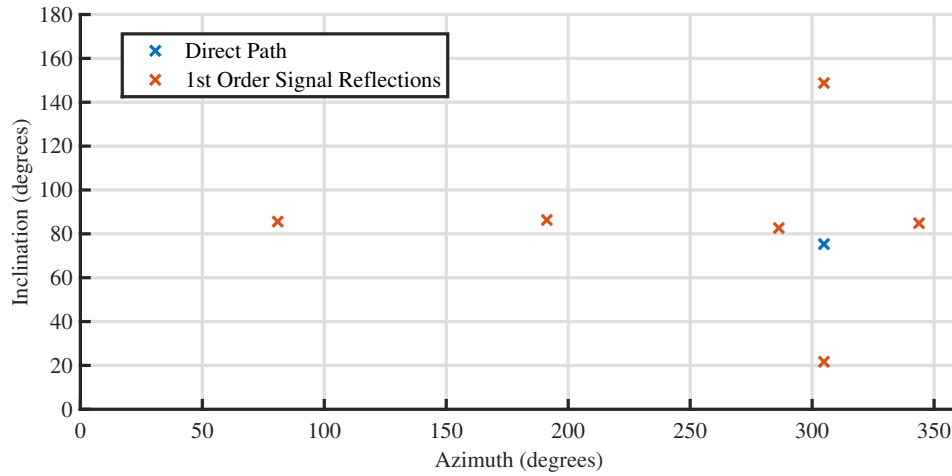


Figure 4.8: The DOAs of the direct path and all six first-order reflections for the source-array setup in room A.

T_{60} of 500 ms. Room B represented a larger, more reverberant enclosure; it was 10 x 8 x 5 m, with a T_{60} of 900 ms. The spherical microphone array impulse response generator (SMIRgen) [127,128] was used to simulate the RIRs within these rooms, between a source and a rigid 32 element SMA of radius 4.2 cm. This model was based on the commercially available ‘Eigenmike em32’ from mh acoustics [13]. Details for simulating the Eigenmike can be found in Appendix C).

In the first of two experiments, the array was placed in the centre of room A, at a height of 1 m, whilst the source was positioned at an azimuth of 305° and an inclination of 75°, at a distance of 1.5 m from the array. The choice of source and microphone array positions was selected in order to achieve a low DRR, as well as distinctly localisable early reflections that would allow the effectiveness of the Rake-DS and Rake-PN methods to be assessed. The DOAs for all six first-order reflections produced for this source-array setup are shown in Figure 4.8, where the closest DOAs are 19.9 ° apart. In the second experiment, the array was randomly assigned a position within the room, with the only constraint being that the distance between the array baffle and any given room boundary never dropped below 10 cm. The position of the source within the room was then constrained to be exactly 5 m away from the source, but was otherwise randomly positioned. Using this

approach, a total of 100 source-array test cases were generated. These configurations were selected to assess the generality of the methods, particularly that of the wRake-PN, in high levels of reverberation.

The spherical array beamformers implemented are of third order ($L = 3$) and a sampling frequency of 8 kHz was chosen. To avoid spatial aliasing, captured signals are bandlimited such that the frequency of operation is restricted to 300-3400 Hz, i.e. narrowband operation. The maximum number of reflections considered in the rake designs is six ($\mathcal{U} \leq 6$). Baseline comparisons are made with the response of a virtual omnidirectional microphone placed at the centre of the spherical array and of a regular PWD beamformer. As an optimal signal independent beamformer, in the maximum directivity sense, the PWD beamformer (4.11) is deemed an appropriate baseline with which to compare the signal independent rake designs proposed. Comparisons are also made with a null-steering beamformer [123] which, like the rakes proposed, also utilises knowledge of multipath DOAs. Based in part on the method presented in [115] (with the exclusion of the noise reduction stage), such a beamformer has been shown to perform dereverberation by pointing a beam at the source and nulls at specified reflections.

4.4.2 Performance measures

To assess dereverberation performance, we consider the well known DRR metric. It should be noted that in the definition used here, the reverberant component is defined exclusively as the taps following the direct path (rather than the entire signal minus the direct path, as has been presented elsewhere, e.g. (3.7) in this work, the definition of which is taken from [56]). This is in order to draw a distinction between the reverberant tail and pre-echoes. The ratio is therefore expressed as

$$\text{DRR} = 10 \log_{10} \left(\frac{\sum_{n=n_{\mathcal{U}}-n_0}^{n_{\mathcal{U}}+n_0} z^2(n)}{\sum_{n=n_{\mathcal{U}}+n_0+1}^{\infty} z^2(n)} \right), \quad (4.17)$$

where $z(n)$ is the effective impulse response in the time domain after raking, and $n_{\mathcal{U}}$ is the sample index corresponding to the reinforced direct path, that is, the time of arrival of the $\mathcal{U}$ -th reflection being raked. Due to finite sampling, the direct path component is defined as an interval, n_0 , either side of the reinforced direct path tap. In this work we set $n_0 = 10$, which corresponds to 1.25 ms. This value was chosen as empirical analysis showed that if the true direct path occurred between two sampling points, it was an appropriately sized interval to capture over 99% of the direct path energy that would have leaked through to adjacent taps (leakage that can be modelled using a sinc function).

As rake designs can introduce perceptually damaging pre-echoes (Figure 4.8f), the relative pre-echo energy is also considered an important evaluation criterion. We define the direct-to-pre-echo ratio (DPER) as the energy ratio of the rake output reinforced direct path, to the total energy of the taps preceding it:

$$\text{DPER} = 10 \log_{10} \left(\frac{\sum_{n=n_{\mathcal{U}}-n_0}^{n_{\mathcal{U}}+n_0} z^2(n)}{\sum_{n=0}^{n_{\mathcal{U}}-n_0-1} z^2(n)} \right). \quad (4.18)$$

This is in contrast to the DRR, in which the direct path energy is compared to the taps which follow.

In addition to the RIR-based performance measures described above, the level of perceived reverberation suppression possible is also evaluated using the $\widetilde{\text{wR}}_{\text{DT}}$ measure presented in 3.4.4. Average scores obtained from 50 TIMIT test speech signals [50] over all 100 source-array arrangements in room B are computed for the dereverberated signals, resulting in a total of 5000 test signals. The practical utility of the designs as the front-end of an ASR system is also investigated using the the open-source Kaldi speech recognition toolkit [129]. The average word error rate (WER) scores are computed for the same test speech signals and represent an alternative way to evaluate the effectiveness of the different systems under test at performing dereverberation. The baseline Kaldi ASR system employed is that developed in [130]. It is a context-dependent Gaussian Mixture Model - Hidden Markov Model phone recogniser, that is trained on the TIMIT speech corpus following the ‘s5’ recipe of Kaldi [129]. The ASR feature vector used includes mel-frequency

Table 4.1: DRR and DPER performance of rake receiver designs in room A, compared to virtual and PWD baselines.

	DRR (dB)	DPER (dB)
Omni-mic	-7.06	-
PWD	1.88	-
Rake-DS*	5.52	9.21
Rake-PN*	3.98	16.56

cepstral coefficients with delta and delta-delta features.

The effect of the systems on overall speech quality is evaluated using PESQ [38]. As the PESQ measure has been shown to be effective at predicting the presence of artefacts such as musical noise in speech signals [131], the results of this evaluation can also serve to indicate if the rake designs introduce perceptually damaging artefacts. Finally, It is worth noting that whilst on occasion the absolute values of the performance measures will be stated, results of the different rake designs and null-steering beamformer will primarily be presented as increases or decreases obtained, relative to the performance achieved by regular PWD beamforming (i.e. ΔDRR , ΔDPER and so on).

4.5 Evaluation

In this section the performances of the various rake receiver designs are evaluated and compared against one another and alternative systems.

4.5.1 RIR-based performance evaluation

Initially, we consider the performance of the Rake-DS* and Rake-PN* for a single test case, as a proof of concept. The output of the rake designs for a Dirac delta input in room A are illustrated in Figure 4.9. The plots help illustrate that in low DRR cases, rake designs that incorporate all six first-order reflections produced in a shoebox room, can lead to substantial reverberation reduction over regular PWD beamforming. Of the two rake designs, it can clearly be seen that, although the Rake-DS* design results in higher direct

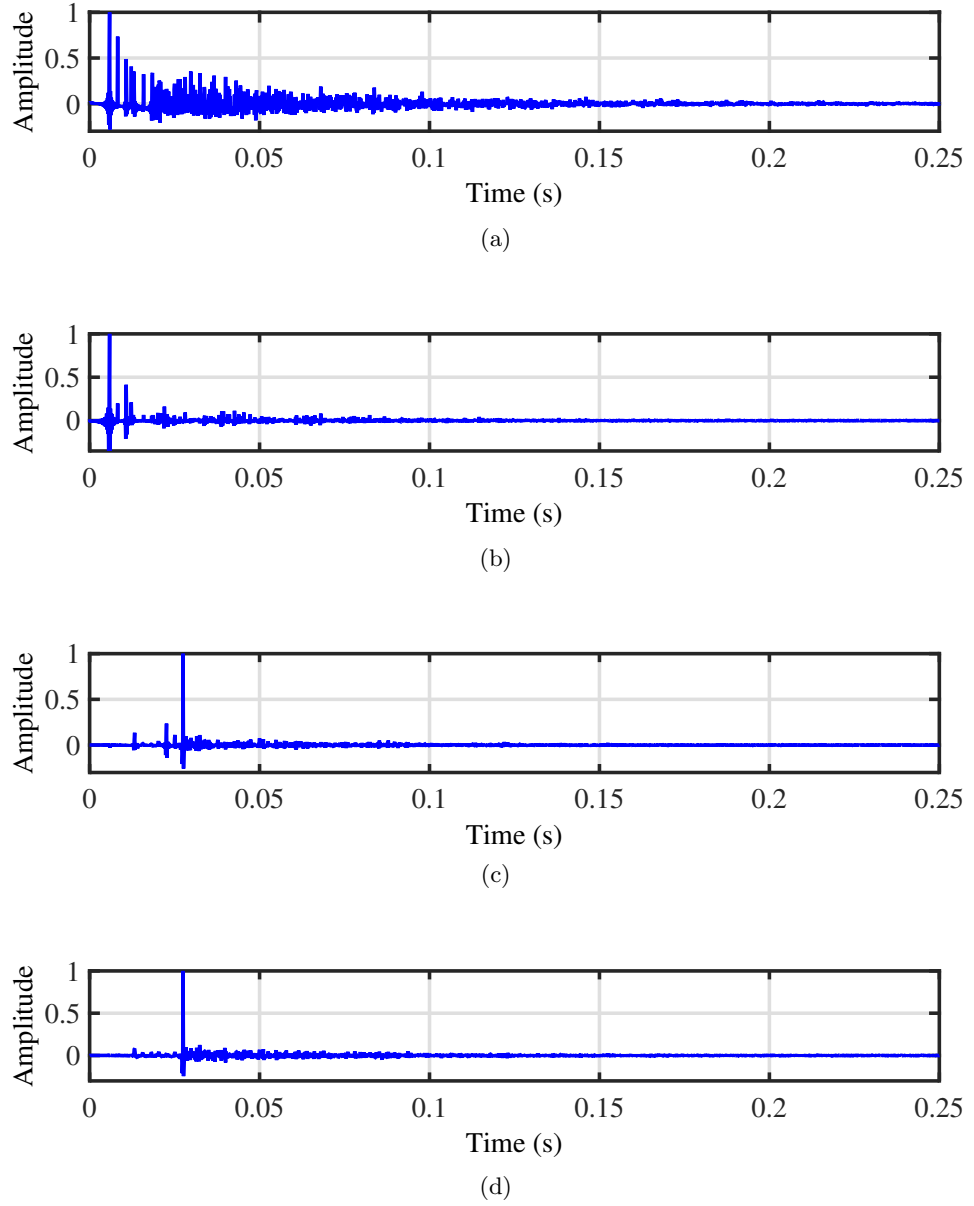


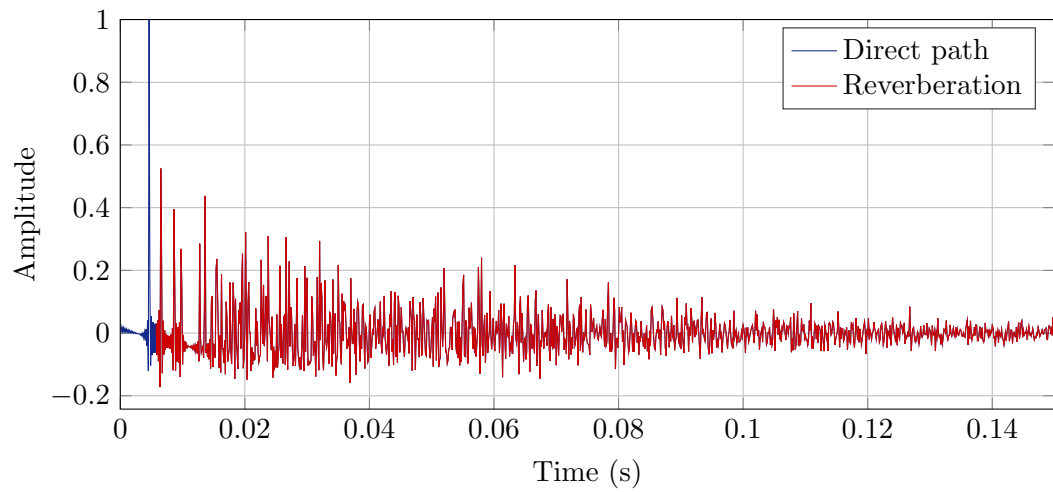
Figure 4.9: Response of a SMA (a) virtual omnidirectional microphone, (b) PWD beamformer, (c) Rake-DS and (d) Rake-PN, to a Dirac delta input in room A ($5 \times 4 \times 3$ m, $T_{60} = 500$ ms).

path energy over the reverberant decay tail, the level of pre-echo energy is high. This is explicitly highlighted in Figure 4.10b. By incorporating null steering, the Rake-PN* on the other hand results in slightly smaller DRR improvements but does not suffer from as much pre-echo (Figure 4.9d). These results are numerically summarised in Table 4.1.

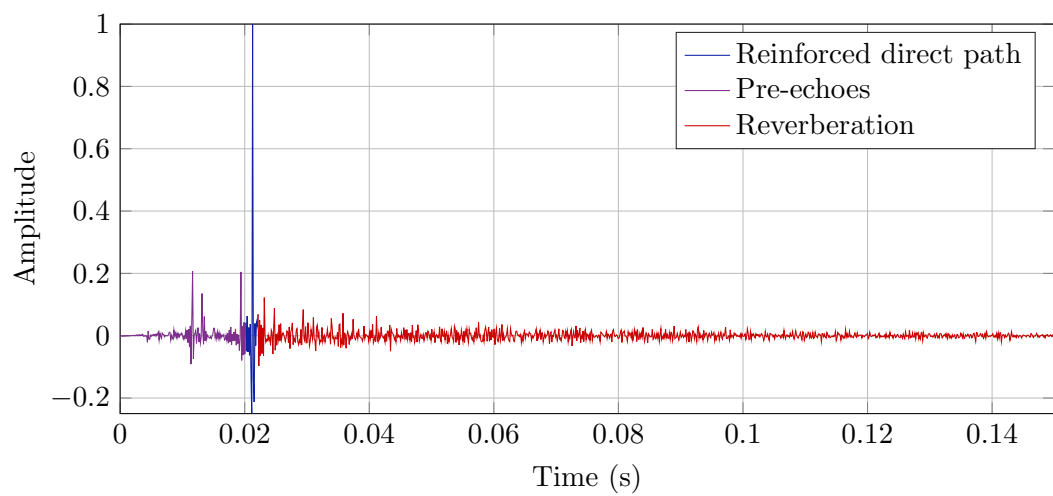
Further investigation showed that the performance of the rake receivers is highly dependent on the number and combination of multipaths being raked. Figure 4.12 shows the distribution of DRR scores for all possible rake combinations for both rake designs. The plots make clear that whilst on average an increasing number of rake fingers results in DRR improvement over regular PWD beamforming, there is a large variation. Depending on the multipaths selected, the rake designs can lead to lower DRR scores than using a PWD beamformer. This is especially true for the Rake-PN. It is hypothesised that this can be attributed to the difficulty in synthesising beampatterns that require a null constraint close to the direction of a beam.

In order to test their generality, the performance of the designs is evaluated for a wider range of source-array setups in room B. From Table 4.2, it is confirmed that by incorporating an increasing number of early reflections, the Rake-DS designs can achieve substantial increases in DRR over regular PWD beamforming. Furthermore, of the two formulations tested, our results demonstrate higher levels of reverberation suppression are possible if the rake finger selection criterion is based on the energy of the reflections, irrespective of their order, with an average improvement of approximately 0.6 dB when six reflections are raked. The larger improvements in DRR achieved in room B compared to the test case in room A indicates that the raking of additional reflections is particularly useful in highly reverberant, low DRR situations. However as was expected, significant pre-echoes are introduced by the delay-and-sum designs (summarised in Table 4.3).

By incorporating null steering, the Rake-PN designs can suppress the level of pre-echo at the output. It is clear from Table 4.2, that an unweighted Rake-PN design offers inconsistent performance; exhibiting high variability across the test cases. This can be explained by the difficulty in synthesising beampatterns that point beams and nulls in similar directions, which is required in some of the source-array configurations. To satisfy



(a)



(b)

Figure 4.10: Output structure, within the moderately reverberant room A, of a spherical microphone array (a) virtual omnidirectional response and (b) delay-and-sum rake. The latter results in substantial DRR gain, but also introduces pre-echoes.

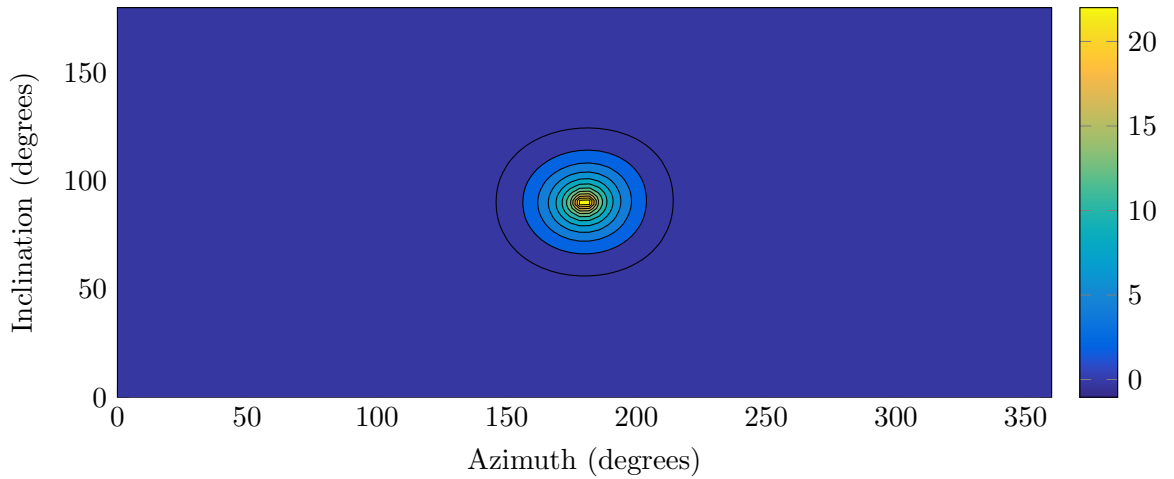
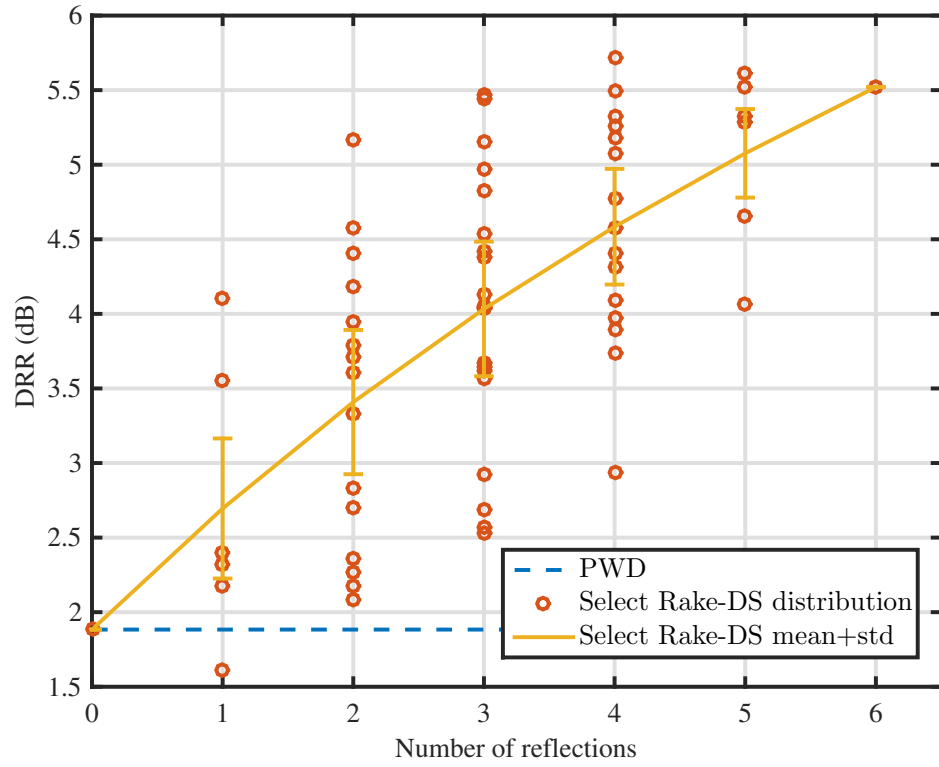


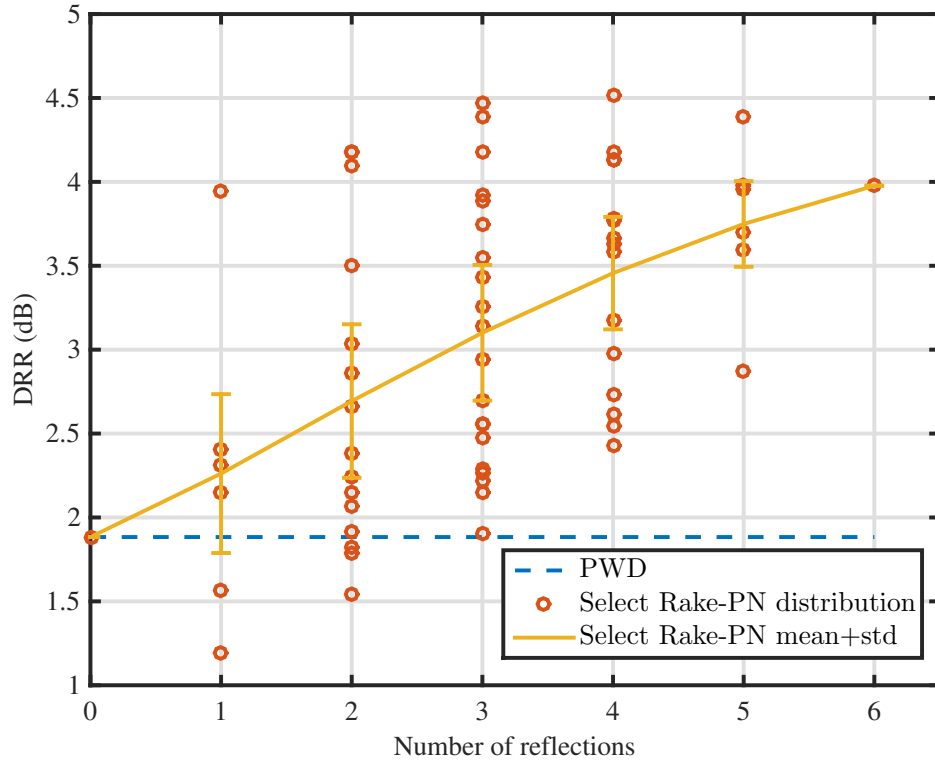
Figure 4.11: Beamformer energies, in decibels, of a third order null steering beamformer with a fixed look direction (180° , 90°) as the direction of a single point null is varied in azimuth and inclination. As the direction of the point null approaches that of the look direction, the beamformer energy increases.

this constraint, beamformer weights with high magnitudes are needed as illustrated in Figure 4.11, which leads to the reverberant sound field being amplified. By weighting rake fingers based on the beamformer energies, the wRake-PN effectively minimises the influence of problematic rake fingers. The wRake-PN thus proves to be a more robust design that can offer DRR improvements, whilst not suffering from the same level of pre-echo as the Rake-DS, as illustrated in Figure 4.13. Further investigation into different weighting exponents is presented in Appendix B.

In contrast to the rake designs, the null-steering beamformer was unable to exploit knowledge of reflection DOAs to the same effect. Whilst it does not suffer from pre-echoes by virtue of its design, pointing nulls at additional reflections actually has a detrimental effect on DRR as can be seen in Table 4.2. This is not altogether surprising, considering in [115] it was able to achieve more significant reverberation suppression by using knowledge of a much larger number of reflection DOAs (50). Furthermore, it was implemented on an eight order ($L = 8$), 162 element SMA. These operating differences could likely account for the poor performance, whilst highlighting its lack of applicability as a technique when only a handful of nulls can be steered at reflections.

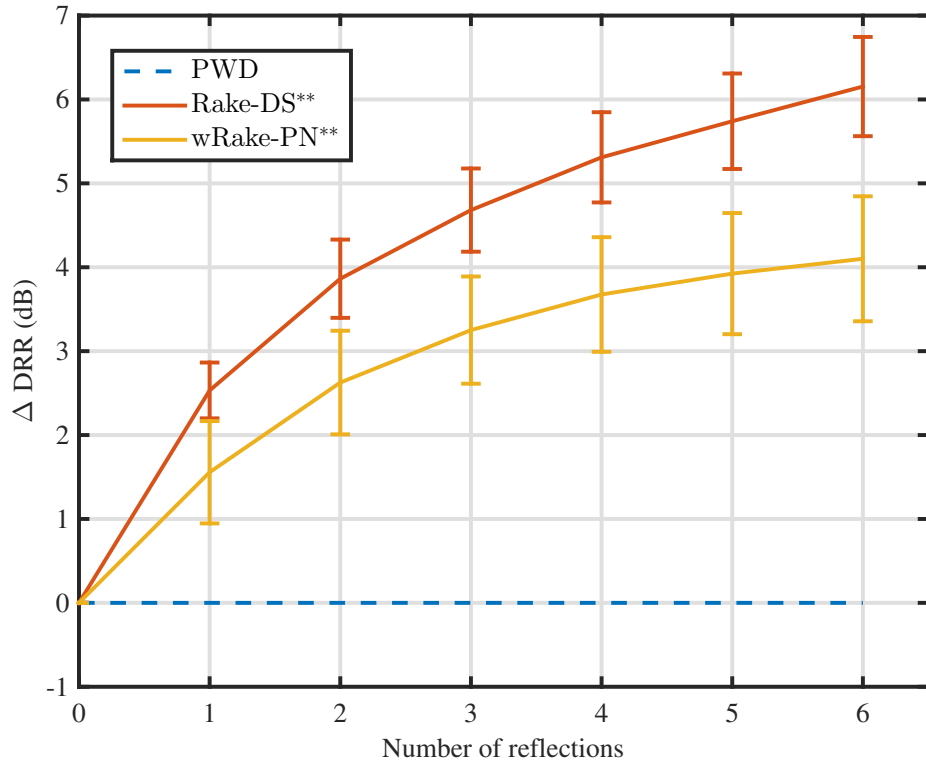


(a)

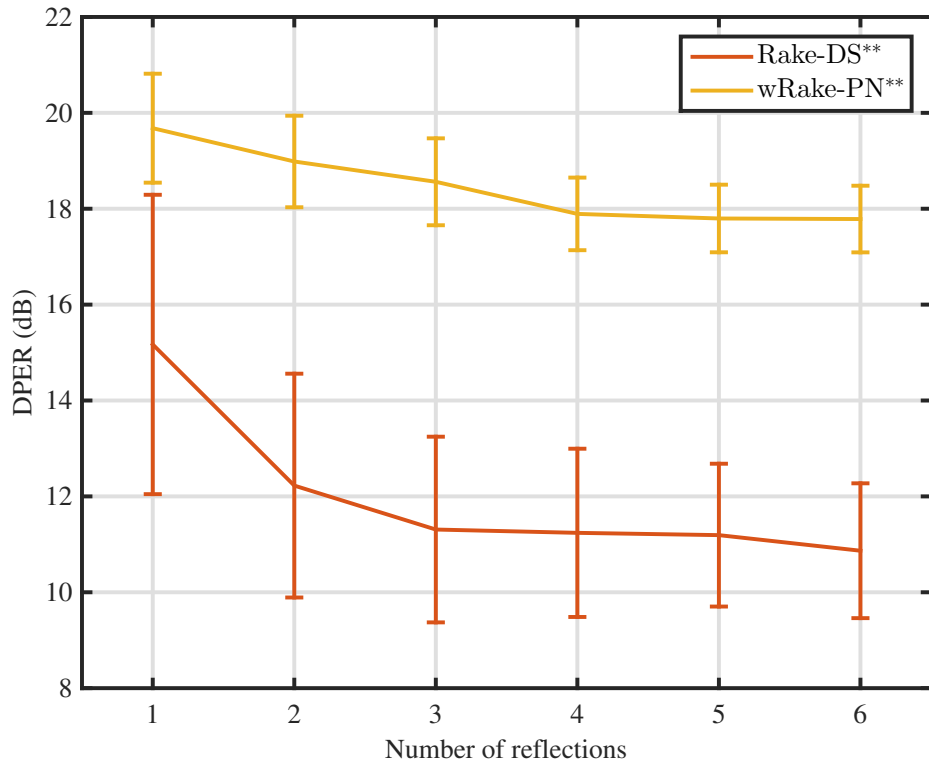


(b)

Figure 4.12: DRR performance of (a) Rake-DS and (b) Rake-PN as a function of the number and combination of first-order reflections raked in room A (note the different vertical scales). Error-bars represents mean and ± 0.5 standard deviation.



(a)



(b)

Figure 4.13: Performance of different rake designs with respect to (a) relative improvements in DRR compared to a PWD beamformer baseline and (b) absolute DPER, as the number of reflections raked varies in room B. Error-bars represent mean and ± 0.5 standard deviation.

Table 4.2: ΔDRR relative to PWD beamforming in room B, as number of reflections considered is varied for different designs. Mean (standard deviation) calculated over all source-array configurations.

Designs	ΔDRR [dB] — Reflections raked		
	1	3	6
Rake-DS*	2.11 (0.90)	3.46 (1.28)	5.30 (1.17)
Rake-DS**	2.53 (0.67)	4.68 (0.99)	6.15 (1.18)
Rake-PN**	-0.87 (5.51)	-0.80 (6.00)	-1.88 (6.68)
wRake-PN**	1.56 (1.22)	3.25 (1.28)	4.10 (1.49)
Null-steerer**	-4.76 (7.42)	-8.58 (7.28)	-10.3 (6.09)

Table 4.3: DPER performance in room B, as number of reflections considered is varied for different designs. Mean (standard deviation) calculated over all source-array configurations.

Designs	DPER [dB] — Reflections raked		
	1	3	6
Rake-DS*	13.7 (5.20)	9.99 (4.10)	10.1 (1.37)
Rake-DS**	15.2 (6.24)	11.3 (3.87)	10.9 (2.81)
Rake-PN**	19.1 (1.74)	16.6 (4.95)	11.8 (6.21)
wRake-PN**	19.7 (2.27)	18.6 (1.81)	17.8 (1.39)
Null-steerer**	—	—	—

4.5.2 Signal-based performance evaluation

Having evaluated the performance of the proposed rake designs using the objective metrics in the previous section, in this section we assess their practical utility. This is achieved by using selected signal-based measures to evaluate the impact of the designs on reverberant test speech.

$w\mathbf{R}_{\text{DT}}$ performance

We initially assess the perceived amount of reverberation suppression possible by using the $w\tilde{\mathbf{R}}_{\text{DT}}$ measure that was presented in Chapter 3 and shown to correlate with perceived reverberation. For the proposed measure, smaller scores correspond to a lower level of perceived reverberation. The changes in $w\tilde{\mathbf{R}}_{\text{DT}}$ for the different designs, relative to the PWD beamformer, are shown in Figure 4.14.

The plots illustrate that whilst the incorporation of additional reflections in the rake

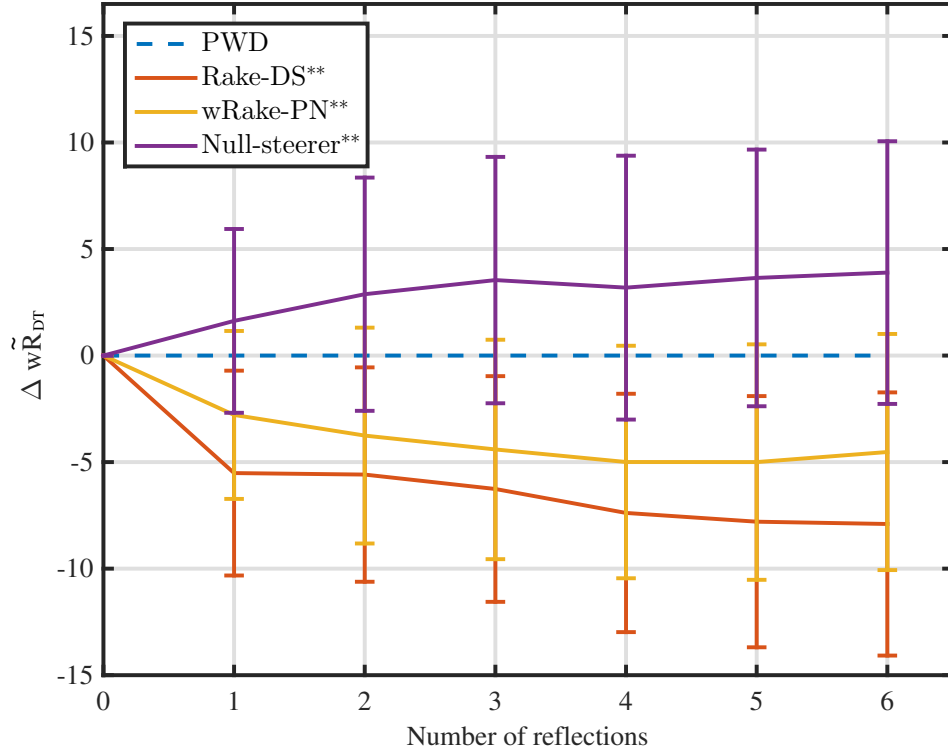


Figure 4.14: Relative improvements in $\tilde{wR}_{DT}$ of rake and null-steerer designs operating in room B, as number of reflections considered varies. Error-bars represent mean and ± 0.5 standard deviation.

designs does, on average, lead to reductions in $\tilde{wR}_{DT}$ scores, there is variability across the signals tested. Moreover, for the wRake-PN design for example, a distinct trend cannot be observed as the number of reflections considered increases. Furthermore, the large overlap in variances suggests that the incorporation of additional reflections does not lead to significant changes in performance of the considered designs. The results obtained could be attributed to the fact that the rake designs primarily reinforce the direct path signal, and do not affect the rate at which the signal decays. As seen in Section 3.4.3, it is the latter property which more significantly impacts the perception of reverberation, whilst modest improvements in DRR are less likely to influence perception. As before, the null-steerer was unable to achieve the same level of performance and in fact leads to average increases in $\tilde{wR}_{DT}$. This is consistent with the performance it has exhibited in previous experiments, in which it was shown to decrease the overall DRR.

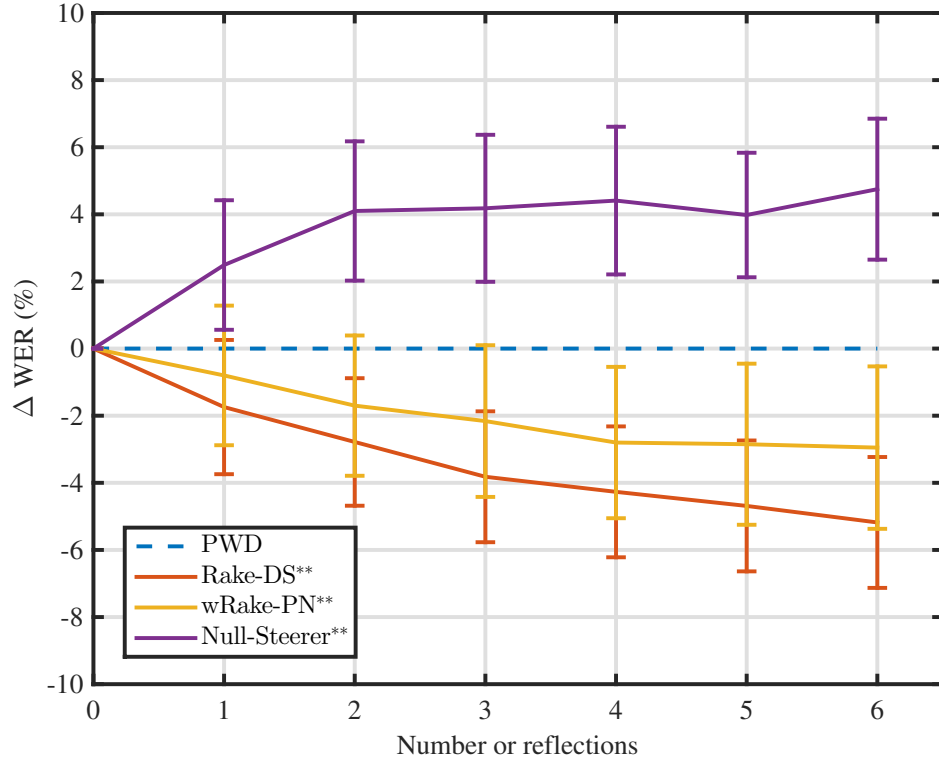


Figure 4.15: Relative improvements in ASR scores, in the form of WER, of rake and null-steerer designs operating in room B, as number of reflections considered varies. Error-bars represent mean and ± 0.5 standard deviation.

For reference, the mean absolute $w\tilde{R}_{DT}$ score for the reverberant test signals, represented by the virtual omnidirectional response, stands at 71.6 with a standard deviation of 9.95. By contrast, applying the PWD beamformer leads to an average $w\tilde{R}_{DT}$ score of 69.9, a relatively modest reduction. Keeping this in mind, we can see that the application of the Rake-DS in a highly reverberant and low DRR environment such as room B, can lead to average $w\tilde{R}_{DT}$ reductions of 9.6 points when six additional reflections are exploited, a reduction which is comparatively more substantial.

ASR performance

As reverberation is known to severely degrade the performance of ASR, we also evaluate the practical effectiveness of the proposed designs at performing dereverberation, by assessing the reduction in WER they can achieve for a baseline speech recognition system.

The performances of the different designs, relative to the PWD beamformer, are shown in Figure 4.15. The plots illustrate that application of the null-steerer to the test data does not result in improvements in recognition, but rather worsens the performance compared to the PWD baseline. By contrast, the rake designs, for the most part result in an average reduction in WER. Whilst initially displaying incremental improvements as the number of reflections raked increases, the reduction in WER for the wRake-PN tends to plateau around either side of about 2.85%.

The Rake-DS, on the other hand, exhibits consistent improvements as the number of reflections raked increases, reaching an average reduction of about 5.18% when the maximum six reflections considered are raked. This seems to imply that the pre-echoes introduced by the Rake-DS design do not have a detrimental effect on ASR performance. For reference, the absolute WER of the anechoic TIMIT speech set under test is 34.7%, with a standard deviation of 9.6 %. The introduction of reverberation to the test signals, represented by the virtual omnidirectional response, by contrast have a WER more than double that at 81.1%, with a standard deviation of 6.3%. Applying the PWD beamformer improves the ASR performance, through an average reduction in WER of 3.84%. Keeping this in mind, we can see that the application of the Rake-DS in a highly reverberant and low DRR environment such as room B, can lead to impressive average WER reductions of 9.1%, for the test ASR system under consideration.

PESQ performance

The impact of the designs on overall perceived quality was also assessed, in the form of improvements achieved in PESQ. The results are displayed in Figure 4.16. As before, the plots illustrate that the null-steerer does not seem to offer improvements in performance over the baseline PWD beamformer. By contrast, the rake designs offer consistent and substantial improvements as the number of reflections considered in the designs increases. Interestingly, the Rake-DS once again exhibits superior performance over the wRake-PN. This would seem to imply that the pre-echoes introduced by the Rake-DS design are not deemed to be perceptually damaging by the PESQ measure.

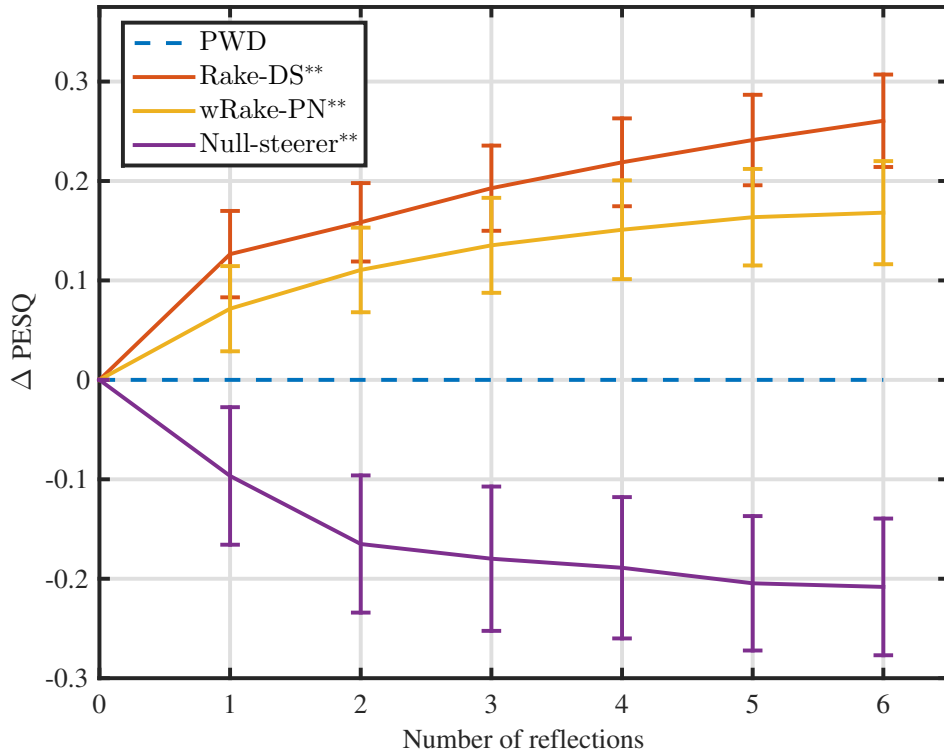


Figure 4.16: Relative improvements in PESQ scores of rake and null-steerer designs operating in room B, as number of reflections considered varies. Error-bars represent mean and ± 0.5 standard deviation.

It is worth noting that having been designed for evaluating speech codecs, PESQ may not be sensitive to pre-echoes in the same way that the human hearing system is. This could perhaps be due to the relative low energies of the pre-echoes and how closely spaced they are to the reinforced direct-path, which ends up masking their effect. Informal listening tests seem to also indicate the impact of pre-echoes is not significant in the test cases considered, though further study may still be required. Additionally, no other artefacts were observed, indicating that the designs considered did not introduce perceptually damaging distortions. However, it may be important to note that in all the test cases considered, beamformer look directions exactly corresponded to multipath DOAs and that errors in this estimation may introduce perceptible distortions as the generated beampatterns are constrained in the look directions.

For reference, the mean absolute PESQ score for the reverberant test signals, rep-

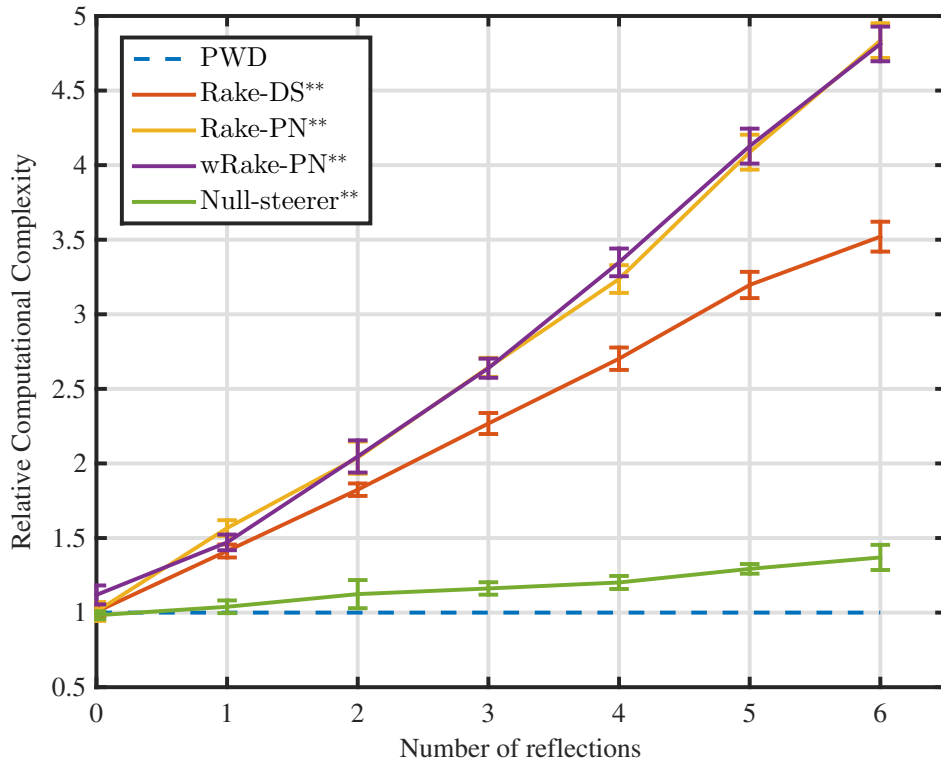


Figure 4.17: Relative computational complexity, measured as factor of PWD beamformer complexity, of the different designs under consideration as number of reflections considered varies.

resented by the virtual omnidirectional response, stands at 1.99 with a standard deviation of 0.19. Applying the PWD beamformer improves the overall speech quality, through an average improvement in PESQ scores of 0.12 points. Keeping this in mind, we can see that the application of the Rake-DS in a highly reverberant and low DRR environment such as room B, can lead to impressive average PESQ improvements as high as 0.35 points when six additional reflections are exploited.

Computational complexity

In addition to evaluating the dereverberation performance of the various designs presented, another performance metric of practical importance is the computational complexity. To this end, we assess the computational complexity of the algorithms under test, in the form of processing times of each design relative to the PWD beamformer. Relative performance

has been chosen as it is less machine dependent than absolute measures, providing greater insight into how computational intensive an operation is. The PWD has been shown to be an efficient implementation that is often used for practical beamforming in realtime [103], thus it provides an appropriate baseline from which to compare against.

The computational cost was measured as the average processing time for each design, starting from the inputting of both the number and the corresponding DOAs of the multipaths to incorporate, through to generation of the output signal. This entailed the synthesis, and for the rake designs phase shifting and combination, of individual beam patterns. It was assumed that the required multipath DOAs were known beforehand and thus their estimation was not factored into the computation (it is thus worth noting that the complexities of the designs based on the reflections to rake, first-order or most energetic, are practically identical).

The results of this analysis are shown in Figure 4.17. We can see that the relative computational cost of all the designs under test increases with the number of reflections considered. However, the null-steerer is computationally the least expensive by a significant margin, with the incorporation of additional nulls not requiring significant processing time. The rake designs by contrast exhibit a roughly linear increase with the number of reflections considered. This behaviour is expected, and similarly it is not surprising that the rake designs that incorporate nulls require yet larger processing times, given the matrix inversion required to synthesize the required beam patterns. It should be noted that the rake designs in this work were implemented in a sequential fashion. Substantial improvements could therefore be made if parallel processing were considered, which given the generation of individual rake fingers, naturally lends itself to such a task. Moreover, in [103], it was shown that the PWD beamformer can be steered in any look direction by multiplying it with a complex exponential. This would require a single PWD beam pattern to be generated, rather than having to resynthesise individual fingers. These possibilities could both facilitate practical, real-time operation of the rake designs.

4.6 Conclusion

In this chapter, three acoustic rake receiver designs formulated in the SHD were presented. Inspired from the widely used rake receiver concept in wireless communications, the proposed designs take advantage of signal multipaths to perform dereverberation. Experimental simulations show that by capturing and constructively combining individual reflections with the direct path, significant reverberation suppression is possible over regular PWD beamforming. Furthermore, by incorporating null steering into their design, the Rake-PN and wRake-PN are able to substantially reduce the level of pre-echoes that are otherwise introduced. The latter design in particular offers robust performance across a wide range of test cases, whilst still maintaining improvements in DRR. The experimental investigation conducted also demonstrates that the performance of the rake designs is dependent on the number and combinations of multipaths raked. A rake finger selection criterion based on reflection energy rather than reflection order was shown to give better performance.

Finally, the practical utility and the relative merits of the different designs were assessed using a select set of evaluation metrics. The $w\tilde{R}_{DT}$ measure, proposed in this thesis, was used to estimate the level of perceived reverberation suppression achieved. As the rake designs primarily lead to improvements in DRR, it showed that on average, only modest improvements over the PWD baseline performance could be observed, with high variability across the test cases. On the other hand, the rake designs demonstrated more substantial improvements in the PESQ and speech recognition scores they were able to achieve over existing methods, with the Rake-DS giving the largest improvements. These results highlight the potential utility of the rake designs as general speech enhancement techniques.

It is worth noting that although the pre-echoes introduced by the Rake-DS were not deemed perceptually damaging, as indicated by informal listening and the PESQ scores obtained, the incorporation of null steering into the rake designs is a potentially useful feature for other applications. For example, in situations where there are competing sources (as in [11]), acoustic rake receivers can lead to improved SNIR in the presence of an

interferer. Whilst it may be unrealistic to expect to have knowledge of up to six reflection DOAs, the incorporation of even a single additional reflection was shown to usually result in significant improvements in the evaluation metrics considered.

Chapter 5

Conclusions

We conclude this work by summarising the contributions made in this thesis, as well as outlining possible areas of future research.

5.1 Thesis contributions

The aims of this thesis were to develop an experimentally validated reverberation measure that correlated with human perception, and to investigate the constructive use of early reflections to perform dereverberation.

In the author's opinion the main contributions of the thesis regarding the former aims are the proposed extensions to, and experimental validation of, the Reverberation Decay Tail (R_{DT}) measure. Having initially reviewed the signal-based measure, ways in which it could be extended and improved were identified. The resulting modifications included extension to wideband operation, the incorporation of equal loudness curves in the calculation of the perceptually motivated Bark spectral representation of the signals, and an improved decay curve detection scheme. The developed extensions were initially validated by assessing the relationship of the computed R_{DTx} measure with objective metrics known to be related to perception. Across a large test set of reverberant speech the R_{DTx} exhibited positive correlation with both the T_{60} and C_{50} metrics, and lower variance for each reverberant condition under test, when compared to the R_{DT} . The

experiments conducted thus showed the signal-based R_{DTX} to be a superior predictor of two acoustic metrics that are closely linked to reverberation, thereby validating to some extent the proposed extensions. Since speech signals captured by a microphone will often suffer from the effects of reverberation as well as additive noise, the impact to, and the robustness of, the extended measure to common types of noise was also investigated.

In order to further validate and develop the R_{DTX} , experimental listening tests into reverberation perception were conducted. The tests provided insight into a human assessor's ability to discern different levels of reverberation in speech produced and captured by a microphone in common acoustic environments, such as offices and meeting rooms. Using the data gathered, the influence of different acoustic metrics on perception was determined. This information was used to propose further improvements to the R_{DT} , which took the form of a weighting scheme that emphasised the relative importance of certain estimated features in the calculation of the measure. Data collected from the listening test data was used to assess the reliability of the different R_{DT} variants as predictors of perceived reverberation. The performance of existing reverberation metrics, models and measures was also evaluated. The results demonstrated that the application of the weighting scheme improved correlations with subjective assessments, and that out of all the signal-based measures considered, the wR_{DT} displayed strongest correlations with subjective assessments. To improve its practical utility, a parametric non-linear mapping function was devised to convert the output of the measure into a perceptual score in the range of 0-100, where a score of zero corresponds to anechoic speech and the highest score to the level of reverberation associated with a room with T_{60} greater than 1 second and a DRR of approximately 0 dB.

The main contributions of this thesis regarding the second stated objective (the constructive use of early reflections to perform dereverberation), was the development of three spherical microphone array acoustic rake receivers. Formulated in the spherical harmonic domain, the proposed designs showed that speech dereverberation could be achieved through the constructive use of strong early reflections. Using experimental simulations, the dereverberation capabilities of the different rake designs were compared

against existing beamforming approaches, using a range of evaluation metrics in a highly reverberant environment. The relative merits of each design were identified, with the practical utility of the spherical delay-and-sum rake design clearly demonstrated across the majority of the evaluation metrics considered. Through the incorporation of even a few additional reflections, the design was able to achieve improvements in the direct-to-reverberant ratio, overall perceived quality, and improved speech recognition scores, where the latter was determined by the reduced word error rate achieved for a baseline automatic speech recognition system.

5.2 Future work

At the completion of this thesis, several areas for potential future work were identified. Regarding the extensions and improvements proposed to the R_{DT} measure, further validation across a larger set of subjectively assessed speech data would be a useful development. The impact of individual extensions on performance could then be determined in a modular fashion and more general conclusions drawn, for example in judging which of the proposed extensions was actually improving correlations with perception. Moreover, a larger set of data could be used to determine optimum weighting values to apply to the individual features that go into calculating the wR_{DT} , through a comprehensive training and testing phase.

Whilst the R_{DT} variants proposed in this work, most notably the $w\tilde{R}_{DT}$, are potentially useful research tools, their practical utility could be greatly improved by developing non-intrusive equivalent designs. As the measure currently uses a reference signal to identify regions where the effects of reverberation would be most perceptible, this could perhaps be accomplished using only the reverberant test signal, by directly identifying and analysing the decays and relative energies of the signal in a similar manner to how blind T_{60} and DRR estimators operate [132–134].

As the rake designs proposed were tested under the assumption that the DOAs of the reflections being raked were exactly known, the effect of imperfect knowledge on

performance should be the subject of further investigation. Additionally, DOA estimation algorithms such as those presented in [124, 135], could be integrated into the designs. Closely related to the robustness of the rakes to DOA errors, this would allow the practicality of the proposed designs to be determined experimentally. Another interesting avenue for further research would be the application of the rake designs to more general scenarios, for example in reverberant environments that are noisy and contain competing sound sources. It is believed in such cases that there is potential utility in exploiting null steering. Incorporated into some of the rake designs, nulls could be used to suppress competing sources rather than, or even in addition to, pre-echoes. Upon exploring the value of the rake in the situations outlined above, simulating and evaluating its performance in more dynamic cases, such as scenarios with moving sources, would be a natural next step. Such work, which would require a source and reflection tracking stage, would provide valuable insight into the practicality of the rake as a means of improved signal acquisition in reverberant environments that mimic real world scenarios.

Bibliography

- [1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.
- [2] M. Wolfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons, 2009.
- [3] T. Gustafsson, B.D. Rao, and M. Trivedi, “Source localization in reverberant environments: modeling and statistical analysis,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, 2003.
- [4] D. A. Berkley, *Acoustical factors affecting hearing aid performance*, chapter Normal Listeners in Typical Rooms - Reverberation Perception, Simulation, and Reduction, pp. 3–24, University Park Press, Baltimore, 1980.
- [5] P. N. Reinhart, P. E. Souza, N. K. Srinivasan, and F. J. Gallun, “Effects of reverberation and compression on consonant identification in individuals with hearing impairment,” *Ear and hearing*, vol. 37, no. 2, pp. 144–152, 2016.
- [6] A. K. Nábelek and D. Mason, “Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms,” *J. of Speech and Hearing Research*, vol. 24, pp. 375–383, 1981.
- [7] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, “The Reverb Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

- [8] 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.
- [9] N. D. Gaubitch and P. A. Naylor, “Analysis of the dereverberation performance of microphone arrays,” in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2005.
- [10] R. Price and P. E. Green, “A communication technique for multipath channels,” *Proceedings of the IRE*, vol. 46, no. 3, pp. 555–570, 1958.
- [11] I. Dokmanić, R. Scheibler, and M. Vetterli, “Raking the cocktail party,” *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 825–836, Aug 2015.
- [12] D. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer, 2016, (to appear).
- [13] M. H. Acoustics, “EM32 Eigenmike microphone array release notes (v17.0),” Oct. 2013.
- [14] B. Rafaely, “Analysis and design of spherical microphone arrays,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.
- [15] J. Y. C. Wen and P. A. Naylor, “An evaluation measure for reverberant speech using tail decay modelling,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, 2006.
- [16] G. A. Soulodre, N. Popplewell, and J. S. Bradley, “Combined effects of early reflections and background noise on speech intelligibility,” *J. of Sound and Vibration*, vol. 135, pp. 123–133, 1989.
- [17] J. P. A. Lochner and J. F. Burger, “The influence of reflections on auditorium acoustics,” *J. of Sound and Vibration*, vol. 1, no. 4, pp. 426–454, Oct. 1964.
- [18] R. H. Bolt and A. D. MacDonald, “Theory of speech masking by reverberation,” *J. Acoust. Soc. Am.*, vol. 21, pp. 577–580, 1949.

- [19] P. Assmann and Q. Summerfield, *The perception of speech under adverse conditions*, Springer, 2004.
- [20] A. K. Nábělek and P. K. Robinson, “Monaural and binaural speech perception in reverberation for listeners of various ages,” *J. of Speech and Hearing Research*, vol. 71, pp. 1242–1248, 1982.
- [21] DeLiang Wang and Guy Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley, 2006.
- [22] Rob Drullman, Joost M Festen, and Reinier Plomp, “Effect of reducing slow temporal modulations on speech reception,” *J. Acoust. Soc. Am.*, vol. 95, pp. 2670, 1994.
- [23] Rob Drullman, Joost M Festen, and Reinier Plomp, “Effect of temporal envelope smearing on speech reception,” *J. Acoust. Soc. Am.*, vol. 95, pp. 1053, 1994.
- [24] T. Houtgast and H. J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.
- [25] S. Greenberg and B. E. D. Kingsbury, “The modulation spectrogram: In pursuit of an invariant representation of speech,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, vol. 3, pp. 1647–1650.
- [26] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.
- [27] J. F. Santos, M. Senoussaoui, and T. H. Falk, “An improved non-intrusive intelligibility metric for noisy and reverberant speech,” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2014, pp. 55–59.
- [28] W. Koenig, “Subjective effects in binaural hearing,” *J. Acoust. Soc. Am.*, vol. 22, no. 1, pp. 61–62, 1950.

- [29] J. P. Moncur and D. D. Dirks, “Binaural and monaural speech intelligibility in reverberation,” *J. of Speech and Hearing Research*, vol. 10, pp. 186–195, 1967.
- [30] B. Libbey and P. H. Rogers, “The effect of overlap-masking on binaural reverberant word intelligibility,” *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3141–3151, 2004.
- [31] J. Paulus, C. Uhle, and J. Herre, “Perceived level of late reverberation in speech and music,” in *Audio Engineering Society Convention 130*, London, UK, May 2011.
- [32] A. J. Watkins and N. J. Holt, “Effects of a complex reflection on vowel identification,” *Acustica*, vol. 86, pp. 532–542, 2000.
- [33] A. J. Watkins, “Perceptual compensation for effects of reverberation in speech identification,” *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 249–262, 2005.
- [34] G. J. Brown, A. V. Beeston, and K. J. Palomäki, “Perceptual compensation for the effects of reverberation on consonant identification: A comparison of human and machine performance,” in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTER-SPEECH)*, Portland, September 2012.
- [35] A. V. Beeston and G. J. Brown, “Consonant confusions provide further evidence that time-reversed rooms disturb compensation for reverberation,” in *Proc. EAA Forum Acusticum, European Congress on Acoustics*, Krakow, September 2014.
- [36] A. V. Beeston, G. J. Brown, and A. J. Watkins, “Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli a,” *J. Acoust. Soc. Am.*, vol. 136, no. 6, pp. 3072–3084, 2014.
- [37] 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, June 1992.
- [38] “Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” Recommendation P.862, International Telecommunications Union (ITU-T), Feb. 2001.

- [39] “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,” Standard P.863, International Telecommunications Union (ITU-T), Jan. 2011.
- [40] K. Kokkinakis and P. C. Loizou, “Evaluation of objective measures for quality assessment of reverberant speech,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 2420–2423.
- [41] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, fourth edition, 2000.
- [42] J. B. Allen, “Effects of small room reverberation on subjective preference,” *J. Acoust. Soc. Am.*, vol. 71, no. S1, pp. S5, 1982.
- [43] A. A. de Lima, T. de M. Prego, S. L. Netto, B. Lee, A. Said, R. W. Schafer, T. Kalker, and M. Fozunbal, “On the quality-assessment of reverberated speech,” *Speech Communication*, vol. 54, pp. 394–401, 2012.
- [44] J. M. F. del Vallado, A. A. de Lima, T. de M. Prego, and S. L. Netto, “Feature analysis for the reverberation perception in speech signals,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 8169–8173.
- [45] S. Goetze, E. Albertin, M. Kallinger, A. Mertins, and K.-D. Kammeyer, “Quality assessment for listening-room compensation algorithms,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, Mar. 2010, p. 24502453.
- [46] D. Lee, D. Cabrera, and W. L. Martens, “Equal reverberance matching of running musical stimuli having various reverberation times and spls,” in *Proc. Intl. Congress on Acoust.(ICA)*, Sydney, Australia, Aug. 2010.
- [47] S. Goetze, A. Warzybok, I. Kodrasi, J.O. Jungmann, B. Cauchi, J. Rennies, E.A.P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, “A study on speech

- quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms,” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014, pp. 233–237.
- [48] J. D. Polack, *La transmission de l’énergie sonore dans les salles*, Ph.D. thesis, Université du Maine, Le Mans, France, 1988.
- [49] D.W. Robinson, “The relation between the sone and phon scales of loudness,” *Acta Acustica united with Acustica*, vol. 3, no. 5, pp. 344–358, 1953.
- [50] 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,” Corpus LDC93S1, Linguistic Data Consortium (LDC), Philadelphia, 1993.
- [51] “Normal equal-loudness level contours,” Recommendation, Intl. Org. for Standardization (ISO), 2003.
- [52] 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.
- [53] E. A. P. Habets, “Room impulse response generator,” Tech. Rep., Technische Universiteit Eindhoven, 2006.
- [54] W. C. Sabine, *Collected Papers on acoustics (Originally 1921)*, Peninsula Publishing, 1993.
- [55] M. Karjalainen, P. Antsalo, A. Mäkilvirta, T. Peltonen, and V. Välimäki, “Estimation of modal decay parameters from noisy response measurements,” *J. Audio Eng. Soc. (AES)*, vol. 11, pp. 867–878, 2002.
- [56] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, “The ACE Challenge - corpus description and performance evaluation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

- [57] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 3, no. 3, pp. 247–251, July 1993.
- [58] M. Jeub, M. Schäfer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Sydney, Australia, July 2009, pp. 1–4.
- [59] D. Griesinger, “How loud is my reverberation?,” in *Proc. Audio Eng. Soc. (AES) Convention*, Feb. 1995.
- [60] T. Hidaka and L. L. Beranek, “Objective and subjective evaluations of twenty-three opera houses in europe, japan, and the americas,” *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 368–383, 2000.
- [61] A. Farina, “Acoustic quality of theatres: correlations between experimental measures and subjective evaluations,” *Applied Acoustics*, vol. 62, no. 8, pp. 889–916, 2001.
- [62] D. A. Berkley and J. B. Allen, *Normal listening in typical rooms: The physical and psychophysical correlates of reverberation*, pp. 3–14, Needham Heights, MA: Allyn and Bacon, second edition, 1993.
- [63] W. G. Gardner and D. Griesinger, “Reverberation level matching experiments,” in *Proc. of W. C. Sabine Centennial Symposium*, Cambridge, Massachusetts, USA, June 1994, pp. 263–266.
- [64] A. A. de Lima, T. de M. Prego, S. L. Netto, B. Lee, A. Said, R. W. Schafer, T. Kalker, and M. Fozunbal, “Feature analysis for quality assessment of reverberated speech,” in *Proc. IEEE Intl. Workshop on Multimedia Signal Processing (MMSP)*, Rio de Janeiro, Brazil, Oct 2009, pp. 1–5.
- [65] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*, John Wiley & Sons, 2007.

- [66] F. E. Toole, “Listening tests—turning opinion into fact,” *J. Audio Eng. Soc. (AES)*, vol. 30, no. 6, pp. 431–445, 1982.
- [67] “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms,” Recommendation P.835, International Telecommunications Union (ITU-T), Nov. 2003.
- [68] A. Warzybok, I. Kodrasi, J.O. Jungmann, E.A.P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, “Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms,” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2014, pp. 333–337.
- [69] “Method for the subjective assessment of intermediate quality levels of coding systems,” Nov. 2003.
- [70] 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.
- [71] B. Cauchi, H. A. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. A. Naylor, “Perceptual and instrumental evaluation of the perceived level of reverberation,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
- [72] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, “Potential biases in mushra listening tests,” in *Proc. Audio Eng. Soc. (AES) Convention*, 2007.
- [73] B.-D. Man and J. D. Reiss, “A pairwise and multiple stimuli approach to perceptual evaluation of microphone types,” in *Proc. Audio Eng. Soc. (AES) Convention*, Rome, Italy, 2013.
- [74] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, “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.

- [75] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, “The single- and multichannel audio recordings database (smard),” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014, pp. 40–44.
- [76] “Objective measurement of active speech level,” Recommendation P.56, International Telecommunications Union (ITU-T), Geneva, Switzerland, Mar. 1993.
- [77] E. Zwicker, H. Fastl, and C. Dallmayr, “Basic program for calculating the loudness of sounds from their 1/3-oct. band spectra according to iso 532 b,” *Acustica*, , no. 55, pp. 6367, 1984.
- [78] “Objective quality measurement of telephone band (300-3400 hz) speech codecs,” Recommendation P.861, International Telecommunications Union (ITU-T), 1998.
- [79] B. C. J. Moore, B. R. Glasberg, and T. Baer, “A model for the prediction of thresholds, loudness, and partial loudness,” *J. Audio Eng. Soc. (AES)*, vol. 45, no. 4, pp. 224–240, 1997.
- [80] J. Timoney, T. Lysaght, and M. Shoenwiesner, “Implementing loudness models in matlab,” Naples, Italy, October 2004.
- [81] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*, Springer, 2011.
- [82] M. Senoussaoui, J. F. Santos, and T. H. Falk, “SRMR variants for improved blind room acoustics characterization,” in *Proc. ACE Challenge Workshop, a satellite of IEEE-WASPAA*, New Paltz, NY, USA, 2015.
- [83] “Single-ended method for objective speech quality assessment in narrow-band telephony applications,” Recommendation P.563, International Telecommunications Union (ITU-T), 2004.
- [84] 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.

- [85] E. A. P. Habets and P. C. W. Sommen, “Speech dereverberation using spectral subtraction and a generalized statistical reverberation model,” *Submitted to Elsevier’s Speech Communications journal*, June 2006.
- [86] S. T. Neely and J. B. Allen, “Invertibility of a room impulse response,” *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, July 1979.
- [87] 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.
- [88] 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.
- [89] 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, Sept. 2013.
- [90] 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, Sept. 2014.
- [91] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, “Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge,” in *Proc. REVERB Challenge Workshop*, Florence, Italy, 2014, vol. 1, pp. 1–8.
- [92] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, “Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2008.

- [93] T. Yoshioka and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
- [94] A. Jukić, N. Mohammadiha, T. V. Waterschoot, T. Gerkmann, and S. Doclo, “Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 96–100.
- [95] A. Jukić, Z. Wang, T. V. Waterschoot, T. Gerkmann, and S. Doclo, “Constrained multi-channel linear prediction for adaptive speech dereverberation,” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, p. 15.
- [96] T. Bietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. V. Waterschoot, “Partitioned block frequency domain kalman filter for multi-channel linear prediction based blind speech dereverberation,” in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.
- [97] P. Annibale, F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, E.A.P. Habets, J. Filos, W. Kellermann, K. Kowalczyk, A. Lombard, E. Mabande, D. Markovic, P.A. Naylor, M. Omologo, R. Rabenstein, A. Sarti, P. Svaizer, and M.R.P. Thomas, “The SCENIC project: Space-time audio processing for environment-aware acoustic sensing and rendering,” in *Proc. Audio Eng. Soc. (AES) Convention*, Oct. 2011, number 131.
- [98] 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.
- [99] E. Jan, P. Svaizer, and J. L. Flanagan, “Matched-filter processing of microphone array for spatial volume selectivity,” in *Circuits and Systems, 1995. ISCAS’95, 1995 IEEE International Symposium on*. IEEE, 1995, vol. 2, pp. 1460–1463.

- [100] T. Nishiura, S. Nakanura, and K. Shikano, “Speech enhancement by multiple beam-forming with reflection signal equalization,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2001, vol. 1, pp. 189–192.
- [101] A. E. O Donovan, R. Duraiswami, and D. N. Zotkin, “Automatic matched filter recovery via the audio camera,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 2826–2829.
- [102] R. Scheibler, I. Dokmanić, and M. Vetterli, “Raking echoes in the time domain,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 554–558.
- [103] J. Meyer and G. Elko, “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2002, vol. 2, pp. 1781–1784.
- [104] T. D. Abhayapala and D. B. Ward, “Theory and design of high order sound field microphones using spherical microphone array,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1949–1952.
- [105] Z. Li, R. Duraiswami, E. Grassi, and L. S. Davis, “Flexible layout and optimal cancellation of the orthonormality error for spherical microphone arrays,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, vol. 4.
- [106] B. Rafaely, B. Weiss, and E. Bachmat, “Spatial aliasing in spherical microphone arrays,” *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, Mar. 2007.
- [107] B. Rafaely, “Sampling-efficient numerically-robust spherical microphone arrays,” in *Proc. 19th Intl. Congr. Acoust. (ICA)*, Madrid, Spain, Sept. 2007.
- [108] B. N. Gover, J. G. Ryan, and M. R. Stinson, “Microphone array measurement system for analysis of directional and spatial variations of sound fields,” *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 1980–1991, 2002.

- [109] Y. Peled and B. Rafaely, “Study of speech intelligibility in noisy enclosures using spherical microphones arrays,” in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2008, pp. 160–163.
- [110] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, “Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 2–16, 2010.
- [111] Z. Li and R. Duraiswami, “Flexible and optimal design of spherical microphone arrays for beamforming,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 702–714, 2007.
- [112] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, “Optimal modal beamforming for spherical microphone arrays,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.
- [113] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, “Spherical microphone array beamforming,” in *Speech Processing in Modern Communication: Challenges and Perspectives*, Israel Cohen, Jacob Benesty, and Sharon Gannot, Eds., chapter 11. Springer, Jan. 2010.
- [114] H. Sun, S. Yan, and U. P. Svensson, “Robust minimum sidelobe beamforming for spherical microphone arrays,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 1045–1051, May 2011.
- [115] Y. Peled and B. Rafaely, “Method for dereverberation and noise reduction using spherical microphone arrays,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 113–116.
- [116] P. K. T. Wu, N. Epain, and C. Jin, “A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4053–4056.

- [117] Z. Lin and W. Qingyu, "Localization of multiple acoustic sources using optimal spherical microphone arrays," in *Proc. 9th Intl. Conf. on Signal Processing (ICSP)*, Oct. 2008, pp. 349–352.
- [118] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.
- [119] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, Sept. 2015.
- [120] V. Tourbabin and B. Rafaely, "Utilizing motion in humanoid robots to enhance spatial information recorded by microphone arrays," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 147–151.
- [121] B. Rafaely, *Fundamentals of Spherical Array Processing*, Springer Topics in Signal Processing. Springer, Berlin Heidelberg, 2015.
- [122] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, first edition, 1999.
- [123] B. Rafaely, "Spherical microphone array with multiple nulls for analysis of directional room impulse responses," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 281–284.
- [124] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *J. Acoust. Soc. Am.*, vol. 131, pp. 2828–2840, 2012.
- [125] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

- [126] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
- [127] D. P. Jarrett, "Spherical Microphone array Impulse Response (SMIR) generator," <http://www.ee.ic.ac.uk/sap/smirgen/>.
- [128] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sept. 2012.
- [129] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, , N. Goel, M. Hannemann, P. Motlicek, Y. Qian, Jan P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, US, Dec. 2011.
- [130] P. P. Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4718–4722.
- [131] M.R.J. Gerrits, R.C. Hendriks, N.D. Gaubitch, J. Jensen, and M.S. Pedersen, *Evaluation of instrumental measures for the prediction of musical noise in enhanced noisy speech*, 2014.
- [132] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2811–2816, Apr. 2012.
- [133] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

-
- [134] T. de M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, “Blind estimators for reverberation time and direct-to-reverberant energy ratio using sub-band speech decomposition,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.
- [135] W. Xue, M. Brookes, and P. A. Naylor, “Cross-correlation based under-modelled multichannel blind acoustic system identification with sparsity regularization,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2016.

Appendix A

Listening test instructions

The instruction sheet provided to participants prior to starting the listening test, is shown below. It is worth noting that participants were also shown the evaluation GUI they would be using (Figure 3.2), how they would be expected to enter their scores (through the touch screen monitor or the mouse) as well as the listening booth the test would be taking place in (a photo of which is illustrated in Figure 3.3).

Instructions for Listening Test Participants

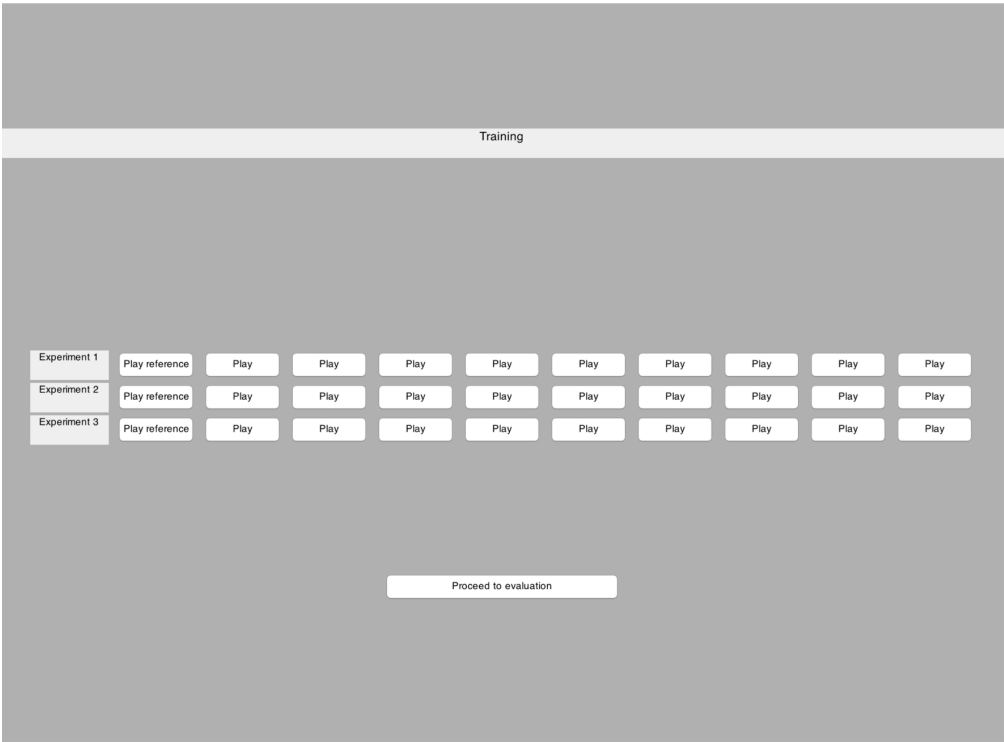
Thank you for agreeing to participate in this listening test. This document details what is expected of you as an assessor. Please read it carefully. If anything is unclear, ask the test conductor for clarification.

Aim:

In the tests that follow, you will listen to several speech recordings and be asked to score the level of **perceived reverberation**. Reverberation is defined in the temporal sense- how much a sound runs on (the level of echo you perceive).

User Information & Training Phase:

At the first stage of the experiment you will be asked to enter your age and gender. Note that the tests are conducted anonymously, and this information is required solely to describe the test subjects. The next stage is a training phase. During this stage you will have the opportunity to familiarise yourself with the test material; all speech recordings you will be asked to assess are presented at this stage. Please listen to them all. You may listen to them as many times as you like before you proceed to the next stage.



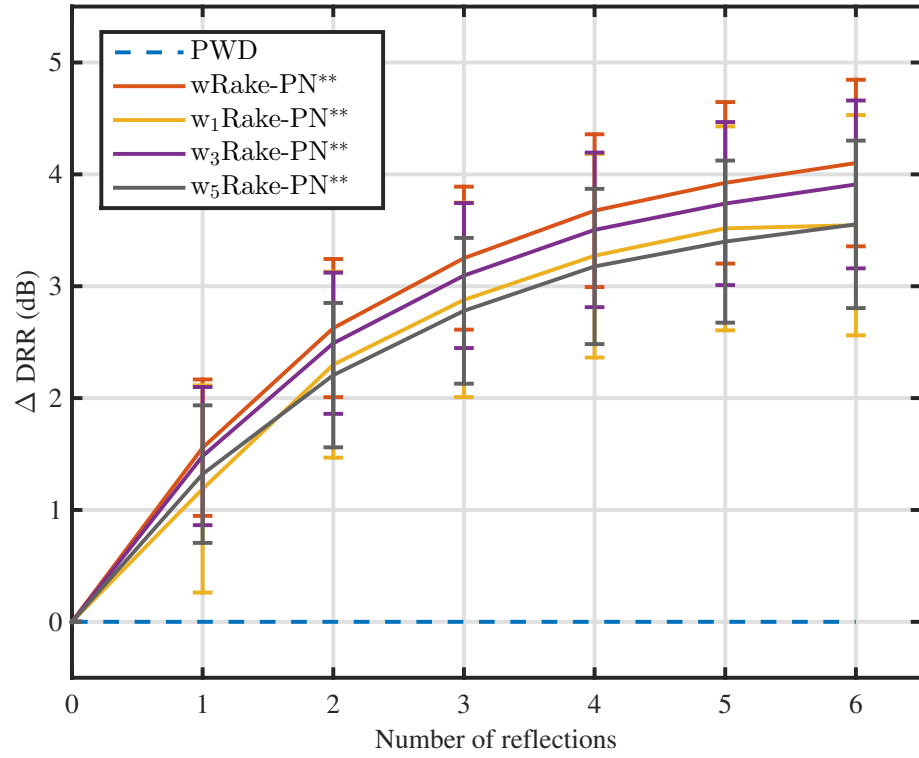
Test Phase:

After the training phase, you may take a break before starting the test. The test is composed of three trials. In each trial you will be presented with 10 test stimuli. After listening to the samples, you are required to score each file based on the level of reverberation you perceive. Use the sliders to input your score. Scores can take a value between 0-100, where a score of zero represents no perceived reverberation and a score of 100 corresponds to a signal you perceive to be extremely reverberant. In addition to the test signals, you will also have the option to listen to a reference signal, which does not suffer from reverberation. You are required to identify a hidden reference signal amongst the test files, which you must give a score of zero. Besides this requirement, there are no other constraints on how you score the test files. You may listen to files as many times as you like, and once you are ready, you can save your results and proceed to the next trial by clicking on the 'Save and proceed' button.

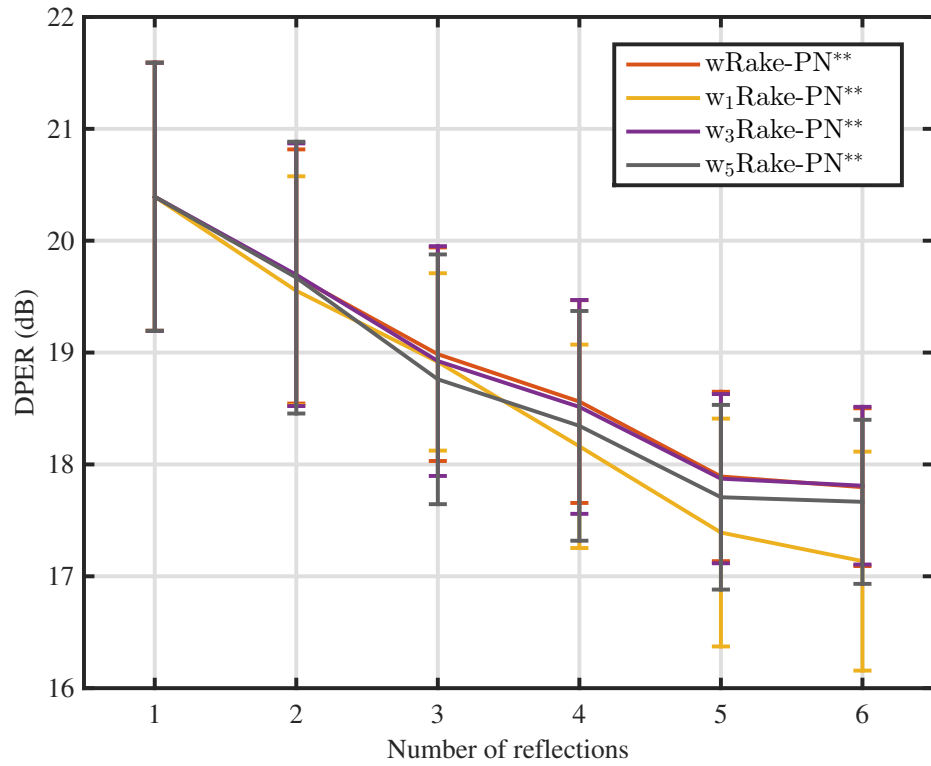
Appendix B

Impact of the pre-echo null Rake weighting scheme

The impact of a de-emphasising exponent other than 2 for the wRake-PN design of 4.16 is investigated in this section. We evaluate the DRR and DPER performances for exponents of 1, 3 and 5, comparing it to the performance of the wRake-PN assessed in the main thesis body, within room B. Overleaf the results show that an exponent of 2 leads to the largest DRR gains, whilst giving comparable DPER performance.



(a)



(b)

Figure B.1: Performance of the wRake-PN rake design with respect to (a) relative improvements in DRR compared to a PWD beamformer baseline and (b) absolute DPER, for different weighting schemes employed by wRake-PN, operating in room B, as the number of reflections raked varies. Error-bars denote mean and ± 0.5 standard deviation.

Appendix C

Modelling the Eigenmike

The ‘Eigenmike em32’ is a commercially available spherical microphone array from mh acoustics [13]. The array consists of 32 microphones placed on a rigid baffle of radius 4.2 cm. Table C.1 details the location of every microphone on the array, by providing the inclination and azimuth angles of each individual microphone, relative to the centre of the array. Using this information and the spherical microphone array impulse response generator (SMIRgen) [127, 128], the room impulse responses between a source and the Eigenmike array can be simulated for a given room setup, as was done in this thesis. Moreover, the quadrature weights required to approximate the spherical Fourier transform, as per equation 4.4, are also detailed.

Table C.1: Eigenmike individual microphone sensor placements, and corresponding quadrature weights required for spherical harmonic domain processing.

<i>Mic. #</i>	Inclination (degrees)	Azimuth (degrees)	Quadrature weight
1	69	0	$9\pi/70$
2	90	32	$5\pi/42$
3	111	0	$9\pi/70$
4	90	328	$5\pi/42$
5	32	0	$5\pi/42$
6	55	45	$9\pi/70$
7	90	69	$9\pi/70$
8	125	45	$9\pi/70$
9	148	0	$5\pi/42$
10	125	315	$9\pi/70$
11	90	291	$9\pi/70$
12	55	315	$9\pi/70$
13	21	91	$9\pi/70$
14	58	90	$5\pi/42$
15	121	90	$5\pi/42$
16	159	89	$9\pi/70$
17	69	180	$9\pi/70$
18	90	212	$5\pi/42$
19	111	180	$9\pi/70$
20	90	148	$5\pi/42$
21	32	180	$5\pi/42$
22	55	225	$9\pi/70$
23	90	249	$9\pi/70$
24	125	225	$9\pi/70$
25	148	180	$5\pi/42$
26	125	135	$9\pi/70$
27	90	111	$9\pi/70$
28	55	135	$9\pi/70$
29	21	269	$9\pi/70$
30	58	270	$5\pi/42$
31	122	270	$5\pi/42$
32	159	271	$9\pi/70$