BLIND ESTIMATION OF ROOM ACOUSTIC PARAMETERS FROM SPEECH AND MUSIC SIGNALS

Paul KENDRICK

Built and Human Environment (BuHu)
School of Computing, Science and Engineering
University of Salford, UK

Submitted in Partial Fulfilment of the Requirements of
the Degree of Doctor of Philosophy, March 2009

Table of Contents

Table of figures

List of tables

Acknowledgments

I am deeply indebted to my supervisor Prof. Trevor Cox for his guidance and excellent supervision throughout the duration of this project. Throughout our regular meetings he provided with me with the support and most importantly the encouragement that has enabled me to complete this research. I am also grateful to my co-supervisor Dr Francis Li for his invaluable guidance and support. I would welcome any opportunities that may arise in future to once again work with both Prof. Cox and Dr. Li.

This project was performed in conjunction with the University of Cardiff Centre for Digital Signal Processing, where Dr Yonggang Zhang and his supervisor Prof. Jonathon Chambers carried out a concurrent research project, utilising blind source separation techniques to perform blind estimation of acoustic parameters. The team at Cardiff were responsible for the development of the envelope segmentation algorithm used within this thesis. I am grateful to the team at Cardiff for offering their expertise, advice and support during a number meetings and correspondences.

A number of in-situ orchestral measurements were carried out and thanks go to Margaret Barry, Head of Orchestras and Ensembles at the Royal Northern College of Music, and to Mark Robinson, Artistic Director & Fellow in Music the University of Bradford for allowing me to record and perform measurements during rehearsals. Thanks also go to Henrik Möller for supplying the real room impulse responses.

I am extremely thankful for the help and support I received from my parents and my friends, in particular my dad, Dr. John Kendrick, for his help with proof reading and the occasional seemingly intractable maths problem. I must also thank my sister, Dr. Emma Kendrick and my mum, Helen Winthorpe Kendrick for offering their proof reading skills on the final draft. I must also thank my partner Laura for putting up with the many late nights and frustrations caused by particular problems I encountered.

I would also like to acknowledge the support of the Engineering and Physical Sciences Research Council, UK for funding this project (GR/S77530/01).

Nomenclature

Abbreviations

ANN	Artificial Neural Network
BSS	Blind Source Seperation
C_{80}	Clarity Index
D	Definition/Deutlichkeit
EDT	Early Decay Time
ELEF	Early Lateral Energy Fraction
FIR	Finite Impulse Response (filter)
HMM	Hidden Markov Model
Hz	Hertz
IACC	Inter-Aural Cross-Correlation
LG	Late Lateral Strength
MLE	Maximum Likelihood Estimation
MTF	Modulation Transfer function
PCA	Principal Components Analysis
RIR	Room Impulse Response
Rt	Reverberation Time
RTF	Room Transfer Response
RMS	Root Mean Square
SPL	Sound Pressure Level
STI	Speech Transmission index
t_s	Centre Time
LTI	Linear Time Invariant
dB	Decibel
ASW	Apparent Source Width
LEV	Listener EnVelopment
G	Strength
MLS	Maximum Length Sequence
IRS	Inverse-Repeat Sequences
JND	Just Noticeable Differences
JNND	Just Not Noticeable Differences
DL	Difference Limens
CMTF	Complex Modulation Transfer Function
CD	Compact Disk
p(x)	Probability density function
t_{mixing}	Mixing time
w.r.t	with respect to
ML	Maximum Likelihood
SNR	Signal to Noise Ratio

Symbols

$f(u)$	Activation function
$x(t)$	Anechoic source signal
ω	Angular frequency
In_i	ANN input
Out_i	ANN output
$S_{xx}(f)$	Auto power spectrum density
log	Base 10 logarithm
$C_{1,2}$	Covariance matrix
CR	Cramér–Rao
$S_{xy}(f)$	Cross Power Spectrum Density
$d(t)$	Decay curve
$\delta()$	Delta function
$env(t)$	Envelope of signal
$E[]$	Expected value
τ	Exponential time constant
f	Frequency
$H[]$	Hilbert transform
$y_h(t)$	Hilbert transform of signal
η	Learning rate
$L()$	Likelihood function
$lnL()$	Log likelihood function
D_m	Mahalanobis distance
μ	Mean
θ	Model parameter vector
ln	Natural logarithm
b_i	Neuron bias
w_i	Neuron weight
$n(t)$	Noise signal
$y(t)$	Received signal
$Y(f)$	Received signal Fourier transform
$h(t)$	Room Impulse response
$H(f)$	Room Transfer Function
Fs	Sampling frequency
$sign(x)$	Sign of x
$X(f)$	Transmitted signal Fourier transform
σ^2	Variance
V	Volume
$F(x)$	μ-law transfer function
$\otimes$	Convolution operator

"I see the parts but not the whole,
I study saints and scholars both,
No perfect plan unfurls,
Do I trust my heart or just my mind,
Why is truth so hard to find in this world."

Dustin Kensrue - 'Stare at the Sun' – 2004

Abstract

The acoustic character of a space is often quantified using objective room acoustic parameters. The measurement of these parameters is difficult in occupied conditions and thus measurements are usually performed when the space is un-occupied. This is despite the knowledge that occupancy can impact significantly on the measured parameter value. Within this thesis new methods are developed by which naturalistic signals such as speech and music can be used to perform acoustic parameter measurement. Adoption of naturalistic signals enables passive measurement during orchestral performances and spoken announcements, thus facilitating easy in-situ measurement.

Two methods are described within this work; (1) a method utilising artificial neural networks where a network is taught to recognise acoustic parameters from received, reverberated signals and (2) a method based on the maximum likelihood estimation of the decay curve of the room from which parameters are then calculated.

(1) The development of the neural network method focuses on a new pre-processor for use with music signals. The pre-processor utilises a narrow band filter bank with centre frequencies chosen based on the equal temperament scale. The success of a machine learning method is linked to the quality of the training data and therefore realistic acoustic simulation algorithms were used to generate a large database of room impulse responses. Room models were defined with realistic randomly generated geometries and surface properties; these models were then used to predict the room impulse responses.

(2) In the second approach, a statistical model of the decay of sound in a room was further developed. This model uses a maximum likelihood (ML) framework to yield a number of decay curve estimates from a received reverberant signal. The success of the method depends on a number of stages developed for the algorithm; (a) a pre-processor to select appropriate decay phases for estimation purposes, (b) a rigorous optimisation algorithm to ensure the correct maximum likelihood estimate is found and (c) a method to yield a single optimum decay curve estimate from which the parameters are calculated.

The ANN and ML methods were tested using orchestral music and speech signals. The ANN method tended to perform well when estimating the early decay time (EDT), for speech and music signals the error was within the subjective difference limens. However, accuracy was reduced for the reverberation time (Rt) and other parameters. By contrast the ML method performed well for Rt with results for both speech and music within the difference limens for reasonable (<4s) reverberation time. In addition reasonable accuracy was found for EDT, Clarity (C_{80}), Centre time (Ts) and Deutichkeit (D). The ML method is also capable of producing accurate estimates of the binaural parameters Early Lateral Energy Fraction (LEF) and the late lateral strength (LG).

A number of real world measurements were carried out in concert halls where the ML accuracy was shown to be sufficient for most parameters. The ML method has the advantage over the ANN method due to its truly blind nature (the ANN method requires a period of learning and is therefore semi-blind). The ML method uses gaps of silence between notes or utterances, when these silence regions are not present the method does not produce an estimate. Accurate estimation requires a long recording (hours of music or many minutes of speech) to ensure that at least some silent regions are present. This thesis shows that, given a sufficiently long recording, accurate estimates of many acoustic parameters can be obtained directly from speech and music.

Further extensions to the ML method detailed in this thesis combine the ML estimated decay curve with cepstral methods which detect the locations of early reflections. This improves the accuracy of many of the parameter estimates.

1 INTRODUCTION

The acoustical character of a room is very important in many situations. Acoustical properties which can enhance speech intelligibility are desirable in lecture theatres and classrooms. In concert halls and opera houses, acoustics which enhance the listening experience are highly prized. In factories or train stations, the goal is to have acoustics that help control the overall noise level and aid speech intelligibility of public announcements. The lay listener regularly makes judgments regarding the acoustic quality of a space. In the case of speech the judgments are based on the ease of understanding the speaker. For music the judgments are much more subjective and often limited to descriptions of 'good' or 'bad' acoustics. Although, when pressed, the lay listener often cannot define what is meant by good or bad acoustics, many studies have identified particular features in the acoustics of the room that are preferred or disliked by the general public. By being mindful of particular features of a room that cause bad acoustics and focusing on features that are preferred, a room can be designed to be good acoustically. It requires extensive knowledge and experience to predict the effect that the geometry and materials will have on the acoustics. More recently, reasonably accurate computer models have been employed to predict the room response, so that the acoustical properties can be predicted before building commences. These computer models enable auralisations of the room to be made so the acoustics of the space may be experienced prior to building. One of the biggest challenges for the designer of an acoustic space is how to marry the understanding of physical acoustics, i.e. the behaviour of sound in an enclosed space, with the understanding of the subjective perception of the acoustics by a listener. In many ways this means the acoustic designer must be part scientist and part craftsman.

1.1 Importance of room acoustic parameters

The subjective impression of room acoustics is largely determined by the reflections that are created when a sound source is activated in a room. Particular aspects of the subjective impression are related to the different properties these reflections may have, such as the rate of the decay of reflections, the direction of the incoming reflections with respect to the listener and the variation of the response with respect to frequency. To

design an acoustic space, particular subjective aspects of room acoustics must be identified and related to the physical acoustics properties. To achieve this, many listening tests have been carried out, both in real rooms and in simulations. By carefully correlating the listener's response with various quantifiable properties of the room acoustics, a vast myriad of objective parameters have been developed that correlate with subjective descriptions of the sound. For example, perceived reverberance has been shown to correlate with parameters such as the reverberation time. Reverberation time is defined as the time it takes for the reverberation to decay to 60dB below the initial level.

These acoustic parameters are essentially the 'cement' by which the science and craft of acoustic design are joined. This makes acoustic parameters tremendously important, both in designing a room for a required purpose, or when measuring an existing room to determine any potential problems.

1.2 Blind estimation of room acoustic parameters

Blind estimation of acoustic parameters refers to when estimations are made only using a recording of the reverberated signal i.e. without using the original transmitted signal. Acoustic parameters are generally determined by analysing the response of the room to artificial test stimuli, the room response is usually calculated by comparing both the reverberated and un-reverberated test signals. Signals are either played back over loudspeakers, or alternatively gunshots can be used. Over loudspeakers, pseudo-random noise or sine sweeps are generally used because this yields accurate and reliable results. One of the major drawbacks of such methodology is a logistical one. In order to yield realistic results the space should be occupied to its usual capacity. This is because human bodies absorb sound and occupancy level can drastically alter the room acoustics.

The unoccupied and occupied condition are compared and contrasted by Hidaka [1]. Hidaka shows that the Rt differs depending on the occupancy. The difference between occupied and unoccupied Rts varies from about 0.25s with short Rts (1s @1kHz) to 2s at long Rts (3s @1kHz). The effect on Clarity (C_{80}) and Early Decay Time (EDT) are also significant, especially for EDT which correlates well with the Rt. C_{80} shows that

introducing an audience to the room can reduce the clarity by 1.4dB (@1 kHz). Hidaka developed empirical compensation equations to enable extrapolations from occupied to unoccupied measures to be made, but these are only very approximate and based on a small number of rooms. This evidence shows that acoustic parameters can significantly alter when the room is occupied, and while measurements are generally performed when the room is unoccupied, in reality it is the parameters measured under occupied conditions that are required.

The artificial stimuli are often unpleasant for an audience to experience. Additionally, in order to obtain sufficient signal-to-noise (SNR), measurements are required to be carried at high sound pressure levels. The procedure can be very time consuming as measurements at multiple locations are required. These factors make it difficult to carry out measurements without disturbing the occupants. Audience members nearest to the source may also be required to wear hearing protection. In addition, finding a few thousand willing souls to fill a concert hall to spend 20 minutes listening to test signals is quite problematic.

Therefore, the interest in this thesis lies in trying to measure acoustic parameters using sounds that are already present in the room; of particular interest is the use of speech or music as a test signal. Measuring using naturalistic sounds should make occupied measurement easier as the signals do not disturb the room occupants. Consequently, it should also facilitate the monitoring of in-use conditions, enabling background measurements to be made while the audience is unawares. These are the motivations of the research, to make measurements using speech and music the methods developed must be blind, this is because the system will not have access to a recording of the source signal.

1.3 Scope, aims and objectives

This thesis documents work carried out under the EPSRC grant GR/S77530/01, entitled "Room Acoustics Parameters from Music". The aim of the project was to facilitate the estimation of acoustic parameters from music signals. Researchers have previously utilised speech signals to perform blind estimations. By also enabling estimations to be

performed using music signals, the range of rooms in which a blind approach can be adopted is increased (e.g. concert halls or opera houses in addition to lecture theatres).

The objectives of the work set out a number of specific goals,

1) Achieve accurate estimation (i.e. within the just noticeable difference range for each parameter) of a number of monaural room acoustic parameters such as; Reverberation Time, Early Decay Time, Clarity and Centre Time (t_s) for both speech and music signals.

2) Achieve accurate estimation (i.e. within the just noticeable difference range for each parameter) of a number of binaural room acoustic parameters such as; Early Lateral Energy Fraction (ELEF), Late Lateral Strength (LG), Inter-aural Cross-Correlation Coefficient (IACC), for both speech and music signals.

3) Source independence, for a blind measurement method to be useful, the performance of the system needs to be independent of the source signal. Achieving source independence with music is expected to be more challenging due to the low similarity between different signals as opposed to different speech signals which are much more comparable.

4) Realistic validation of the method. Performing a large number of measurements in different rooms to validate the methods is very time consuming and difficult, therefore a large database of room responses is to be generated using modern geometric modelling and simulation techniques. The room responses are to be as realistic as possible and also to be representative of a wide range of possible rooms. This is intended to make the validation of the methods rigorous and to ensure the measurement system is applicable to a wide range of possible rooms.

As often happens in research projects, the methodologies and approaches outlined in the case for support presented with the grant application changed to take advantage of methods developed by others. Two approaches have been focused on in this thesis. As much of this work has been inspired by the work of Li [2], the starting point was the continuation of the envelope spectrum machine learning method developed for speech

signals to also include music signals. The second method was inspired by the work of Ratnam *et al.* [3] [4] who developed a model-based, maximum likelihood (ML) estimation method. In this thesis the ML method was further developed to include a more realistic model of sound decay. The following two short sections describe the work that was carried out in these two areas.

1.3.1 Artificial neural network method

Li and Cox [5] developed a machine learning method to determine the Speech Transmission Index (STI) from received running speech. This method is quasi-blind as source signals do not need to be monitored during measurements, but they are required during the training phases of the machine learning algorithm. A key limitation of this method is that it is an empirical approach which requires extensive training before use. Even so, it can be shown that approximately one minute of speech can provide high accuracy for Rt estimation. With slightly compromised accuracy the method can be made completely blind, as the low frequency statistical properties of speech are not very different from speaker-to-speaker [6]. This method is termed the 'envelope spectrum method' because of the pre-processor used.

The envelope spectrum method was originally developed to be used with narrated speech. For parameters used in the evaluation of concert halls, however, it is natural to examine the use of music as an excitation signal, and this has not been considered in the context of the envelope spectrum method. In comparison to speech, music offers a larger bandwidth of excitation and so acoustic parameters may be able to be measured over a wider range of frequencies. In particular, in comparison to speech, music is often played at greater overall sound pressure levels, this is advantageous for measurement purposes as it provides a higher signal-to-noise ratio. Music, however, is a rather imperfect test signal, as shall be shown later. To work with music, the envelope spectrum method needed to be adapted to deal with the inherent statistical differences between speech and music, and these adaptations are outlined in this thesis (Section 6.3). Since the development of the envelope spectrum method, geometric room simulation techniques have improved. The use of these techniques to generate the dataset used to train the machine learning algorithm, improves the applicability of the method towards real rooms. The new dataset also affects the overall accuracy of the

method; this is because the neural network must perform a more complex mapping task than with the previous simpler dataset.

The limitations of the envelope spectrum method due to the inherent properties of music signals steered the research towards finding an alternative method. The method sought was one that is wholly blind and has much better Rt estimation accuracy. This method is described in the next section.

1.3.2 Maximum likelihood method

An alternative to the envelope spectrum method is one using Maximum Likelihood Estimation (MLE). This approach was originally developed by Ratman *et al.* [3]. The concept is to use decay phases following speech utterances or music notes and to estimate the decay curve using a model of sound decay. The method is inherently blind as it searches the signal for regions of free decay. The accuracy demonstrated by Ratman *et al.*, however, was insufficient for parameter measurement; this is because the model of sound decay used could only model very uniform reverberant sound decay. This thesis describes the development of a more realistic model of reverberant sound decay; one that allows for non-diffuse spaces. This model, when utilised in a maximum likelihood framework allows good prediction of both reverberation time and early decay time of a range of parameter values. The parameter range is limited at the lower end by the natural decay rate of the musical notes and utterances and at the higher end by the length of time in-between these notes and utterances.

One of the most important research challenges was how to yield robust parameter estimates from the hundreds or perhaps thousands of maximum likelihood decay curve estimates. A framework was developed to yield estimates of the room decay curve, from this, accurate estimates of the monaural parameters; Rt, EDT, C_{80} and t_s were gained. The estimates showed for both speech and music, over a wide range of room responses, a high percentage of the estimates (especially for Rt) were within the subjective difference limens. In addition, experiments with binaural parameters were carried out demonstrating excellent accuracy for both the Early Lateral Energy Fraction and the Late Lateral strength

Finally, cepstral analysis, which is a signal processing technique often used to detect echoes, perform dereverberation and analyse speech signals [88, 89, 90], was employed to improve the accuracy of the early sound field estimation. This enabled the pattern of the early reflections to be applied to the model of sound decay yielding a realistic sounding, blindly-estimated, impulse response (the maximum likelihood estimates, due to their stochastic nature, do not provide estimates of the fine-structure of the early reflections). While no significant improvement in parameter accuracy was found, informal subjective tests found that these impulse responses sounded much more subjectively realistic. This would enable audio engineers to mimic the acoustics of a room by analysing pre-recorded audio; new sounds could then be mixed in to the pre-existing audio and be perceived as if they were produced under the same acoustic conditions.

1.4 Thesis structure

The thesis is structured as follows; Chapter 2 is an introduction to the acoustic parameters that will be investigated and how they are traditionally measured. Chapter 3 is a brief review of existing methods for blind estimation of acoustic parameters. Chapter 4 describes the generation of a database of simulated room responses intended to represent the broad range of possible room impulse responses. Chapter 5 is an introduction into some of the techniques that are used throughout the thesis such as Artificial Neural Networks (ANN) and optimization. Chapter 6 describes work carried out by extending the envelope spectrum method, originally developed to estimate parameters from speech, to also work with music. Chapters 7 and 8 describe the development and validation of the maximum likelihood method for blind estimation of parameters from speech and music. It is based on a maximum likelihood fit of a stochastic room model to portions of the reverberant signal and analysing the results; this method extends previous work by using a more realistic room model and considering additional parameters. Chapter 9 describes a process by which the maximum likelihood method can be extended using cepstral analysis techniques to provide a better estimate of the early sound field. Chapter 10 compares the machine learning technique and the maximum likelihood methods. Chapter 11 draws conclusions from the work and some appendices are contained at the end of the thesis.

2 ROOM ACOUSTIC PARAMETERS – DESCRIPTION AND MEASUREMENT METHODS

This chapter introduces several acoustic parameters which are used to define the acoustic character of a space. These include monaural parameters such as reverberation time (Rt), early decay time (EDT), clarity index (C_{80}), centre time (t_s), the Speech Transmission Index (STI) and Definition (D). In addition a number of binaural parameters is introduced; these parameters use directional information to quantify subjective effects such as spaciousness and envelopment, and include the parameters Early Lateral Energy Fraction (ELEF), Late Lateral Strength (LG) and the Inter-Aural Cross-correlation Coefficient (IACC). Also introduced are several methods that are used to measure the room impulse response (RIR) and how, from that response, the various parameters are calculated. Finally, the concept of subjective difference limens is introduced, which is defined as the smallest change in a particular parameter that can be detected by human hearing, and how these are used to quantify the accuracy of a measurement system.

2.1 Transfer characteristics of rooms

Acoustic spaces are approximately linear-time-invariant (LTI) and generally passive systems. A linear system maps an input to an output using only linear operations. For example, if an input, $x_1(t)$ produces the output $y_1(t)$ and another input, $x_2(t)$ produces the output $y_2(t)$, the input $a_1x_1(t)+a_2x_2(t)$ will always produce $a_1y_1(t)+a_2y_2(t)$. Time invariance means that the system response does not alter with time. As a result, if the input signal is delayed by a certain amount the output signal will be delayed by that same amount. The room response to a sound source is pictured in Figure 2-1.

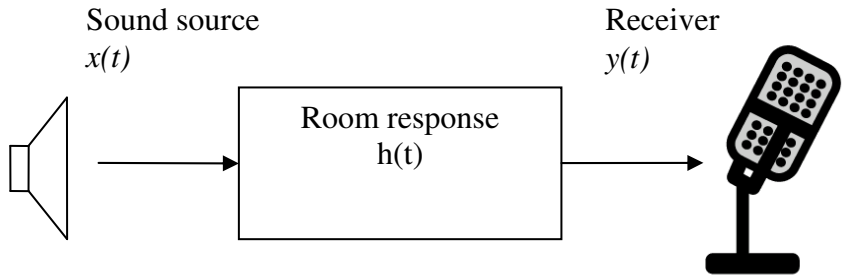


Figure 2-1. Effect of the room response on a souna source

The room response can be quantified theoretically by its response to a Dirac delta function, this is known as the system's impulse response, *h(t)*. A Dirac delta function is a short pulse of energy at time zero which is infinitesimally narrow and its integral of all time is unity. The response of the room to a sound source is the convolution of the impulse response, *h(t),* with the source signal *x(t)*;

$$y(t) = h(t) * x(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau \qquad (2\text{-}1)$$

This process has the useful property that when the system is described in the frequency domain, the convolution operation becomes a multiplicative one:

$$Y(f) = H(f)X(f) \qquad (2\text{-}2)$$

where *X(f), Y(f)* and *H(f)* are the Fourier transforms of *x(t), y(t)* and *h(t)*. The Fourier transform of *h(t), H(f)* is known as the system transfer function.

The transfer characteristics of a room are completely described, for a single source and listener orientation, by the room impulse response (RIR) or the room transfer function (RTF). This assumes linearity and time invariance, however the RIR is known to change with conditions within the space such as temperature (but this generally happens slowly over long periods of time) and often time-invariance is assumed. Public address systems can introduce significant levels of distortion but the RIR is a measure only of the passive system response and does not include any active systems that may be in use.

2.2 Quantifying the acoustic quality of a room

There are two ways to quantify the acoustical properties of a room; either perform some subjective test to gauge opinions of listeners, or record the response of the room to a sound and use that to quantitatively classify the acoustics. Subjective tests are time consuming and difficult to carry out whereas recording the response of a room to a sound is fairly straight forward.

Acoustic parameters are useful measures of the acoustic quality of a room which are calculable from the recorded response of a room to a sound. These parameters have been designed to correlate with opinions of listeners in subjective tests. Parameters are either calculated from the RIR or from the decay curve (the decay curve is the recorded reverberant decay when a steady noise source is turned off). It is much preferred to quantify the quality of acoustics using parameters, as it is less time consuming and much more consistent than the alternative which involves quantifying opinion using subjective tests. For example, an Rt of 1s will be the same no matter the location, however opinions of the perceived amount of reverberance may well differ based on listener experience: A person who lives outdoors and has never experienced the long reverberation times in cathedrals or churches may perceive relatively non-reverberant spaces to be quite reverberant.

2.2.1 The Decay Curve

The decay curve can be measured by switching off a white noise source that has been exciting the room for a period of time sufficient to yield a steady sound pressure level. The sound level is recorded after the source is turned off and the resulting change in sound level is known as the decay curve. It can be used to show particular features of the reverberation, such as the difference between the early and late responses. Figure 2-2 shows an example of a decay curve. A white noise burst is used to excite the room, the sound quickly builds up to a stationary level and then the sound source is switched off at time 0. The decay curve is measured from $t=0$. The EDT and Rt can be calculated directly from the decay curve. Due to the stochastic nature of the source signal, the measurement must be repeated multiple times to gain stable estimates for the

Rt and EDT. This method of recording the decay curve is known as the 'interrupted noise method' and is described in ISO 3382 [8].

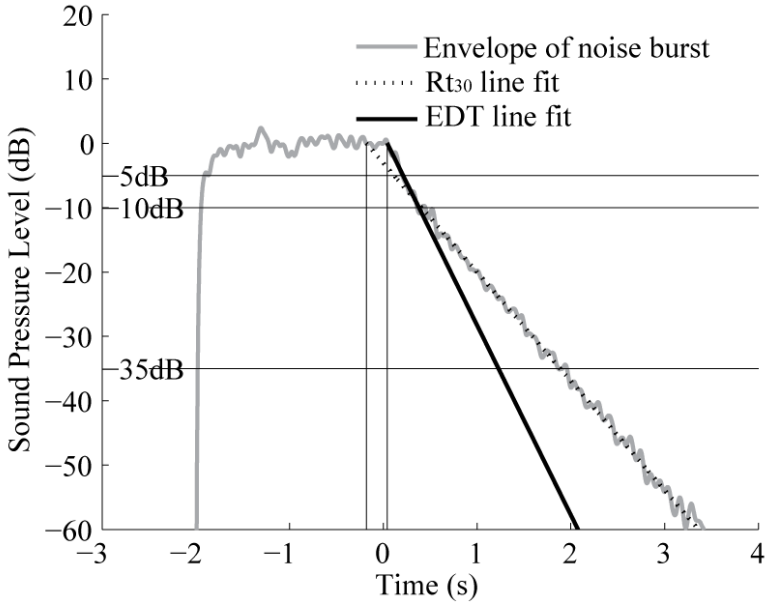


Figure 2-2. Decay curve of an enclosure, the straight lines demonstrate how the Rt_{30} and EDT are calculated. After Kendrick et al. [9]

The decay curve can also be calculated from the RIR via a procedure called Schroeder backwards integration [10]. This is equivalent to calculating the ensemble average of all possible decay curves that may be generated using the interrupted noise method and is calculated as follows.

$$decay(t) = \int_{t}^{\infty} h^2(t)dt \qquad (2\text{-}3)$$

The parameters EDT and Rt can now be calculated from this smoothed decay curve without having to average.

2.3 Acoustic parameters

Once the RIR has been measured, the acoustic parameters can be calculated directly from it. The next section introduces the acoustic parameters that will be referred to within this thesis and describes their purpose and method of calculating. Two

parameters that relate to the rate of reverberant decay are the reverberation time and the early decay time, other parameters can often give an indication to the clarity of the sound, such as clarity index deutlichkeit/definition and centre time. In addition there exist a number of parameters that have been designed to correlate with the sense of envelopment. These parameters are known as binaural parameters; this is due to the binaural nature of spatial hearing. Examples of binaural parameters include the inter-aural cross-correlation coefficient, late lateral strength and early lateral energy fraction. All of these parameters are defined in the international standard ISO3382 [8].

2.3.1 Reverberation time (Rt)

Reverberation time is a very important acoustic parameter both historically and practically. It was the earliest acoustic parameter developed by Sabine in the late 19th century [11]. Using organ pipes as a sound source Sabine discovered that by estimating the time it takes for the sound to become inaudible once turned off, the rate of sound decay was related to the size of the space and the amount of absorption within it. This work helped to define the parameter known as reverberation time. The level at which the sound became inaudible was assumed to be roughly 60dB below the initial level and Rt_{60} was more formally defined as the time taken for the sound pressure level to decay by 60dB when a stationary sound source is turned off. Later definitions (due to the difficulty in gaining a 60dB signal to noise ratio in recordings) utilised the range from -5dB to -35dB and performed a least-squares fit to this range. The time it takes for this line to decay by 60dB is extrapolated and is referred to as Rt_{30}. Often, when the signal-to-noise ratio is insufficient Rt is calculated on smaller dynamic ranges, for example from -5 to -25dB would be Rt_{20}. Differences between Rt values calculated over different level ranges arise due to the decay curve being not quite exponential. These are often referred to as non-uniform decay curves and the reason for the deviation from a true exponential decay may be due to the room being insufficiently diffuse, or often the room may be coupled with a secondary space having a differing Rt. Unless otherwise stated, in this thesis Rt_{30} is used.

2.3.2 Early decay time (EDT)

In more recent times, it has become increasingly common to characterise room reverberation using the EDT. This parameter is calculated in the same manner as the

Rt, only the decay rate is calculated over the first 10dB and multiplied by 6 to gain the time it would take to decay to -60dB (if the response was uniform). This parameter was introduced by Jordan [12] and later subjective tests [13] confirmed that EDT is more closely correlated with perceived 'reverberance' when listening to music signals.

2.3.3 Measures of clarity

The Clarity Index was defined by Reichardt *et al.* [14] and is intended to describe the transparency of music in a concert hall. It is calculated from the RIR and is described in the following equation representing the ratio between the energy that arrives in the first 80ms and the energy that arrives from 80ms to infinity and is measured in dB.

$$C_{80} = 10\log_{10}\left[\frac{\int_{0}^{80ms} h(t)^2\,dt}{\int_{80ms}^{\infty} h(t)^2\,dt}\right] dB \qquad (2\text{-}4)$$

Gade [13] describes the typical range of clarity values for concert halls as between -5 and +3dB. A similar parameter known as Definition or Deutlichkeit was defined by Thiele [95]. This is similar to Clarity only the integration period is over the first 50ms and the value is now calculated as the percentage of energy that is received in the first 50ms over the whole impulse response.

$$D = 100\left[\frac{\int_{0}^{50ms} h(t)^2\,dt}{\int_{0}^{\infty} h(t)^2\,dt}\right] \% \qquad (2\text{-}5)$$

Centre time [15] is another parameter that correlates with the definition, clarity and speech intelligibility, the difference between this index and C_{80} and D is that there is no fixed integration limit of 50ms or 80ms. This fixed integration limit is often regarded as a crude approximation of how the human hearing system works. Centre time is usually quoted in *ms* and is calculated as follows;

$$t_s = \left[\frac{\int_0^\infty h(t)^2 t dt}{\int_0^\infty h(t)^2 dt} \right] \quad (2\text{-}6)$$

This essentially measures the centre-of-gravity of the impulse response. Centre time has a high inverse correlation with speech intelligibility word score [16].

2.3.4 Speech transmission index (STI)

An important parameter that can be used to quantify the level of speech intelligibility of a space is the speech transmission index [17]. This parameter can also be used to quantify speech intelligibility over communication channels. The function, by which the STI is calculated is known as the modulation transfer function (MTF). The MTF describes the system response to the envelope of signals and therefore can be used to describe how well a signal envelope is preserved when the signal is reverberated. It is of particular importance as the effects of reverberation on the envelope are used to estimate acoustic parameters in the machine learning methods adopted in this work. In that respect the STI and MTF are explained in detail in the following sections.

The Envelope Spectrum

During the development of the STI, Houtgast [18] introduced the envelope spectrum. The envelope spectrum is a physical description of running speech that can be used to indicate the level of interfering noise or reverberation present in the signal. The envelope spectrum is the power spectrum of the signal envelope. The spectrum is normalised so that 0dB represents a modulation frequency where 100% modulation is taking place. Both reverberation and interfering noise reduce the level of modulation, but they have subtly different effects. Noise reduces the level of modulation equally at all modulation frequencies. Reverberation smoothes the envelope and any sudden sharp peaks in the envelope tend to be masked, while the slower varying features of the signal tend to be preserved, this has the effect of attenuating the higher frequencies while preserving the level of the lower frequencies in the envelope spectrum. This is similar to the effect a linear low pass filtering operation has on a signal. Houtgast [19] demonstrated that the area underneath the envelope spectra is an indication as to the

level of background noise and the amount of reverberation, both of which have significant impact on speech intelligibility. Quantifying the level of modulation transferred through a channel can give a good indication of the speech intelligibility.

Modulation Transfer function (MTF)

Both reverberation and interfering noise were found to significantly degrade the intelligibility of speech. These are generally linear processes and can be simulated as addition and convolution operations. Other, non-linear, operations also have significant impacts on speech intelligibility. Examples of nonlinear operations that degenerate speech intelligibility are: Distortion due to loudspeaker or amplifier overloading, signal artefacts due to data-compression operations and signal processing artefacts such as compression or limiting the signal. All of these operations impact on speech intelligibility and can be quantified by analysing the distortions these operations introduce to the signal envelope.

Steeneken *et al.* [17] developed a method to quantify the ability of a channel to transfer modulation by modulating noise signals with single frequency oscillations and measuring the level of modulation transmitted for a number of modulation frequencies. This was more robust than using speech signals to quantify modulation transmission levels.

Measuring each modulation frequency separately enables non-linear distortions to be accounted for. The function that quantifies the level of modulation transferred over a system is called the Modulation Transfer Function. To measure the MTF the level of modulation that is transmitted, at a number of different modulation frequencies and in a number of different audio bands, is measured by transmitting noise signals modulated by sine waves through the channel. The STI is calculated via a weighted sum of the modulation transmission level in all the frequency bands and at different modulation frequencies, these weights have been optimised against word score tests [17].

The MTF is time consuming to measure experimentally, so Houtgast [20] developed the Rapid Speech Transmission Index (RASTI) method. This method enabled a number of modulation frequencies and audio bands to be measured simultaneously and the

measurement duration was greatly reduced however, it is not as robust as STI when significant non-linearities are present. Another quick measurement method known as STI-PA [21] is also available, which was specifically developed to account for distortions in public announcement systems.

The Complex Modulation Transfer Function (CMTF)

The MTF takes no account of phase information but the Complex Modulation Transfer Function does. The CMTF can be estimated in the same manner as the MTF, by measuring the modulation level transferred when sine wave modulated random noise signals are applied to the system. The phase is determined with reference to the modulation source; the sine wave. This more rigorous definition of the transfer of modulation was defined by Schroeder [22].

To measure any non-linear processes in the transmission channel forces the MTF to be measured separately at individual frequencies. This causes the measurement method to be very time consuming. In many cases nonlinear effects are very small and can be ignored. For example, when measuring acoustic parameters it can often be assumed that nonlinear effects from loudspeaker, digital or analogue distortions are negligible. When only linear operators affect a channel, the MTF and CMTF can be determined from the normalised Fourier transform of the squared magnitude of the room impulse response [22] as shown in equation (2-7). Therefore the large number of repeat measurements that the STI method requires, to account for nonlinear distortions, can be avoided.

$$CMTF(\omega) = \frac{\int_0^\infty h^2 e^{-j\omega t} dt}{\int_0^\infty h^2 dt} \qquad (2\text{-}7)$$

Schroeder [22] demonstrated that this relationship was equivalent to the CMTF measured using the modulated noise method, given that the signal which passed through the system was a stationary random process modulated by cosine waves. Polack [23] extended the definition to other signals. In particular, he showed how music signals

could be used to estimate the CMTF. Polack accounted for the statistical non-stationary nature of music signals by using the anechoic signal envelope as a matched filter to estimate the transfer characteristics from the reverberant signal envelope.

2.3.5 Binaural parameters

Thus far, the types of parameters described have been restricted to those that can be measured using a single omni-directional microphone. This means that all information regarding the direction of arrival of the reflections is not accounted for. It is known that this directional information contributes strongly to another subjective effect known as the 'sensation of space' [16]. This 'sensation of space' or 'spatial impression', caused by reflections arriving at the listener from a variety of angles, is dependent on the angle of arrival of these reflections, the level of the reflections, and the time of arrival.

There is now a widespread agreement that spatial impression is not a one-dimensional factor but can be split into two more or less independent components. These components are (a) apparent source width (ASW), i.e. how wide the sound source is perceived to be, and (b) listener envelopment (LEV), i.e. how 'enclosed' or 'encased' the listener feels by the sound.

Both ASW and LEV are influenced by reflections arriving from lateral directions. If a reflection arrives before 80ms after the direct sound it contributes to the ASW and if it arrives after 80ms it contributes to the sensation of LEV.

A number of objective parameters have been developed that correlate with subjective opinion about ASW and LEV. These parameters must be measured using multiple microphones. The Early Lateral Energy Fraction [24] is related to the energy of the received reflections and their angle of arrival, and correlates with subjective views on envelopment. ELEF compares the energy of the reflections received from all directions with the energy of the reflections weighted according to a $cos\theta$ function (i.e. from lateral directions) in the delay range from 5ms to 80ms. The parameter is calculated as follows:

$$ELEF = \frac{\int_{5ms}^{80ms} h^2(t)\cos\theta dt}{\int_{0ms}^{80ms} h^2(t)dt} \qquad (2\text{-}8)$$

This parameter can be measured using an omni-directional microphone and a figure-of-8 microphone with the null facing the sound source. However, as the figure-of-8 response is squared, the actual pickup pattern is $cos^2\theta$. The subjective result correlates better when the pattern is $cos\theta$ therefore the following correction is often used [8]:

$$ELEF = \frac{\int_{5ms}^{80ms} h_{figure8}(t)h_{omni}(t)dt}{\int_{0ms}^{80ms} {h_{omni}}^2(t)dt} \qquad (2\text{-}9)$$

Another parameter that looks at the early reflections and their angle of arrival is the inter-aural cross correlation coefficient. This parameter uses the sound incident on each of the listener's ears and computes a value based on the similarity of the two signals to one-another. The function to compute this parameter is as follows:

$$\varphi_{rl} = \frac{\int_0^{t_0} h_r(t)h_l(t+\tau)dt}{\left(\int_0^{t_0}[h_r(t)]^2 dt \int_0^{t_0}[h_l(t)]^2 dt\right)^{\frac{1}{2}}} \qquad (2\text{-}10)$$

The value t_0 restricts the correlation to early reflections by limiting the impulse response to say 100ms. The maximum correlation (max φ) for the lag range $|\tau|<1ms$, is the IACC. The IACC can be measured using either a dummy head or by using in-ear microphones.

Both IACC and ELEF correlate well with ASW. Late lateral strength (LG) [25] correlates with listener envelopment. Bradley [25] indicated that listener envelopment is related to both overall sound pressure level and spatial or angular distribution of the reflections. He included these factors in the late lateral strength parameter LG. This

parameter is closely related to the strength factor G, restricted to late lateral reflections. It is the energy ratio of lateral reflections arriving after 80ms to the energy level of an impulse response transmitted via the same source in anechoic conditions to a receiver 10m from the source. This is often a difficult parameter to characterise in practice due to the requirement of careful calibration of the sound source

$$LG = 10\log_{10}\left(\frac{\int_{80ms}^{\infty}[h(t)\cos\theta]^2\,dt}{\int_{80ms}^{\infty}[h_A(t)]^2\,dt}\right)dB \qquad (2\text{-}11)$$

where $h_A(t)$ is the response measured in an anechoic chamber using the same sound source at a distance of ten meters.

2.1 RIR Measurement methods

There are a number of methods by which the RIR can be measured, this section introduces a number of them and discusses their particular advantages and disadvantages. ISO 3382 [8] is the international standard that sets out a number of recommendations relating to the measurement of room acoustics parameters. ISO 3382 does not however restrict RIR measurement methods to any one particular procedure. Each RIR measurement method has its own particular advantages and disadvantages. The next section will briefly introduce the various methods and their particular nuances.

Exciting the room directly with a delta function, thus measuring the impulse response directly, is not strictly possible. This is because the delta function is a theoretical signal which has infinite amplitude and infinitesimal duration. However, an impulse can be approximated by a sharp loud sound such as a pistol shot, a balloon pop or spark gaps (in scale models). It is difficult to simulate an impulse directly by playback over loudspeaker as these sounds must be extremely loud, at least 45dB above the background level for Rt_{30}. A pistol shot certainly achieves this but the audience and firer must be adequately protected from the high sound pressure levels (there is also the additional problem of carrying around what would appear to the lay observer as a deadly weapon!). Additionally, repeated measurements are required for every

source/receiver orientation to account for the stochastic variations in the pistol shot every time it is fired. This makes accurate measurements using a pistol as a sound source very time consuming. However, despite this, the pistol method remains a very easy and quick method to gain an approximation to the RIR.

In recent decades it has become common practice to excite the room using an artificially generated signal. From the resulting recording, and using the known input signal, the RIR is recovered (usually using a correlation based technique) while trying to minimise the effects of noise and nonlinear distortions on the recovered RIR. The room is generally excited using an omni-directional sound source and the RIR recorded at a number of receiver positions using an omni-directional microphone. It is usual to measure the broadband response. After the impulse response has been recorded the response may be filtered into octave or third octave bands to further analyse how the response alters with frequency. For a comprehensive review of a number of RIR measurement techniques please refer to reference [26]. The following subsections provide an outline to the range of options the acoustician has when measuring the RIR.

2.1.1 Maximum length sequence measurement

A linear system response, as assumed in the case of the RIR, can be modelled as the process shown in Figure 2-3. The input signal, *x(t),* is convolved with the RIR, *h(t),* to produce the reverberant signal *y(t).* However, noise *n(t)* is often present in the signal path particularly in the form of background electrical noise generated by, or picked up by the cables, microphones and preamps of the recording system.

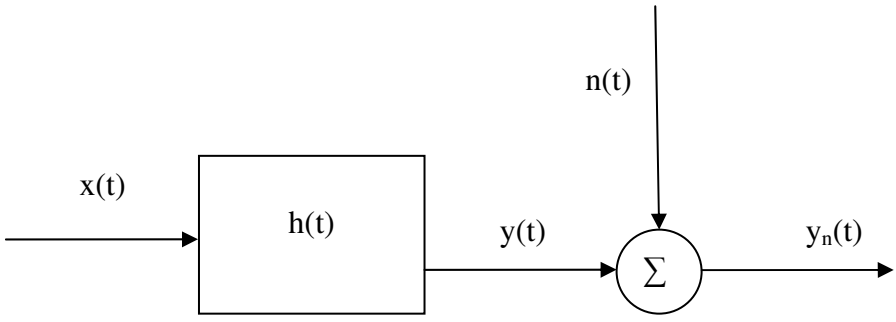


Figure 2-3. Room response to a signal including measurement noise

The spectra of $x(t)$ and $y_n(t)$, $X(f)$ and $Y_n(f)$, are used to compute the impulse response. Due to the limited available length (T) of the two signals, the spectra are calculated from

a finite time record producing $X_T(f)$ and $Y_{nT}(f)$. The cross-power spectrum density (CPSD) between the two signals is estimated by averaging;

$$S_{xy}(f) = E\left[\frac{X_T(f)Y_{nT}^*(f)}{T}\right] \tag{2-12}$$

where $E[]$ denotes expectation, by further expanding this relationship and replacing Y_{nT}^* with $(H.X_T+N)^*$ gives the following relationship;

$$S_{xy}(f) = E\left[\frac{X_T(f)\left(X_T(f)H\right)(f)^* + X_T(f)N(f)^*}{T}\right] = E\left[\frac{X_T\left(X_T(f)H(f)\right)^*}{T}\right] \tag{2-13}$$

As long as N is uncorrelated with the source signal, the factor $E[X_T(f)N^*]$ will tend to zero. Examining the above equation shows that the CSPD is equal to the multiplication of the RTF and the autospectrum of the input signal, this means;

$$H(f) = \frac{S_{xy}(f)}{S_{xx}(f)} \tag{2-14}$$

Good estimations of $S_{xy}(f)$ and $S_{xx}(f)$ are crucial to obtaining a good estimation of the RTF. Signals with broadband power spectra are required to produce reliable estimates over a range of frequencies, random signals such as white noise can be used but in order to yield accurate estimates many averages are required. The use of deterministic pseudo-random noise signals enable noise free measurements to be made without the need for averaging. However in the presence of background noise a number of averages will be required.

One such deterministic signal that has almost the same stochastic properties as white noise is a Maximum Length Sequence (MLS). This methodology was originally developed in the 1960s [27] but was first adopted by Schroeder in 1979 [28] for the purpose of RIR measurement. MLS is a periodic signal, which repeats every L sample where $L=2^N-1$ (N being an integer). The signal consists of a binary sequence where each

sample can be either +1 or -1. The MLS signal has a white spectrum (except at 0 Hz) and its autocorrelation is almost a delta function. A white spectrum is one that demonstrates equal amplitude at all frequencies often also referred to as a flat spectrum. The reverberated MLS signal is then cross-correlated with the input signal and as the auto-correlation of the periodic MLS signal is a series of impulse-like functions, period *L,* the resulting cross-correlation yields the RIR (or a series of repeating RIRs referred to as PIR, periodic impulse response). One must be careful of ensuring that *L* is long enough to encompass the whole RIR within one period and thus L must be greater than Rt. If *L* is too short then time-aliasing can occur where circular convolution causes the end of the RIR to be mapped over the top of the start of the RIR.

The noise tolerance of the MLS method is very good as noise is effectively spread out over the length of the estimated impulses response, and can be removed by averaging [26]. Additionally, the MLS sequence has a crest factor of unity giving it the best possible signal-to-noise ratio. The measurement time is very short and, in a low noise case, no averages are required, as the RIR can be accurately estimated using only two periods of the MLS signal, one period to ensure the system is excited into steady-state and the other to perform the measurement. The MLS signal is deterministic and therefore the input signal does not have to be recorded to perform the correlations with the output. The input signal can be regenerated as and when required. This means that only a single recording channel is required. The computation of the cross-correlation between input and output can be efficiently calculated using the Hadamard transform [29].

The biggest drawback with the MLS method is the poor behaviour when non-linear distortions are present, generally due to over-extrusion of the driver in the loudspeaker. Any such distortion artefacts cause erroneous peaks to be distributed along the impulse response. Some, but not all of these artefacts, can be suppressed using inverse-repeat sequences (IRS) [30], this maintains many of the advantages of MLS. Alternatively, long sequences can be used to spread out the distortion in time.

2.1.2 Swept-sine waves

In order to overcome the limitations due to loudspeaker non-linearities, another method was developed [31] which has a very efficient method for dealing with non-linearities. This method, known as the swept-sine method uses a sine wave whose frequency continuously increases with time, often logarithmic frequency sweeps are used to emphasise the low frequencies due to the generally lower SNR at lower frequencies. These swept sine waves, or chirps, can account for non-linearities because the loudspeaker distortions cause harmonic distortion at higher multiples of the excitation frequencies. The swept sine method utilises this to completely remove all harmonic distortions.

The RIR is calculated via a cross-correlation with the input signal $x(t)$. Because the swept sine wave effectively only excites a single frequency at any one time, any additional frequencies generated via harmonic distortions from cone over extrusion are immediately identified. After correlating $x(t)$ and $y(t)$ any harmonic distortions will cause frequencies to occur before the swept sine wave has excited that particular frequency. This simple fact causes the resulting impulse responses estimated when distortions are present to be non-causal; the system is outputting frequencies before they have occurred in the source signal. This causes all non-linear distortions to show up as peaks at negative time locations in the resulting $h(t)$ estimate. As all RIRs are causal, by simply removing all negative time components the swept sine method can account for these harmonic distortions. While this method shows excellent distortion immunity, the lower crest factor of the sine wave, when compared with the MLS signal, means that the noise immunity is lower; this can be improved by averaging.

2.2 Subjective difference limens

In order to judge the performance of a measurement method it must be compared against the ability of the human auditory system to detect changes in acoustic conditions. The subjective difference limens are the smallest change in a parameter value that the human auditory system can detect. Difference limens (DLs) are determined using just noticeable differences (JND) and just not noticeable differences (JNND) in parameters. To determine the DL, subjective tests must be carried out under

controlled conditions where sounds are presented to listeners with differing acoustic conditions. A statistical analysis of the results yields a value for the DL. Table 2-1 shows the DLs for a number of parameters, experimentally determined by various authors.

Parameter	Difference limens	
Rt (T_{30}) s	*5 % (>2s) 0.1s (≤2s)*	*[32]*
EDT (s)	*5 %(>2s) 0.1s (≤2s)*	*[32]*
C_{80}	*1 dB*	*[32]*
D_{50}	*5%*	*[32]*
Ts	*0.01ms*	*[32]*
ELEF	*0.048*	*[33]*
LG	*7 dB*	*Mean of results from [34] & [25]*

Table 2-1. Difference limens for acoustic parameters used in this thesis

Kuttruff [16] comments that with respect to Rt, there is no point in stating accuracy greater than 0.1s, this is also true for EDT, this is due to limitations on measurement accuracy. Plots of results will be presented in the manner depicted in Figure 2-4. The axes are the actual parameters value (x-axis) versus the estimated parameter value (y-axis), each result is presented as a dot on the graph. Difference limens are presented for reference as dashed lines.

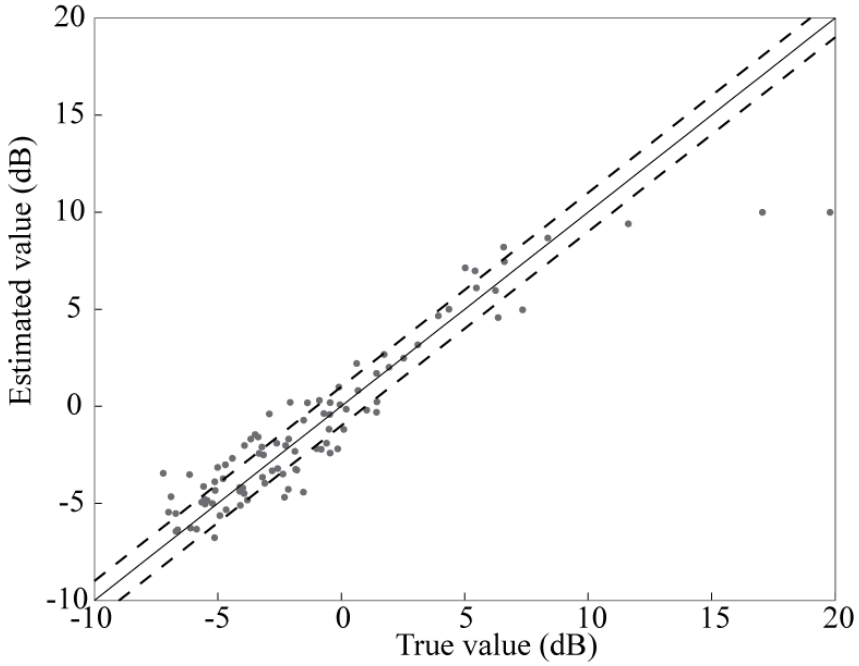


Figure 2-4. Example of plot showing accuracy of C_{80} estimations. Each dot represents an estimation and the difference limens are represented by dashed lines. This gives an indication of the accuracy of the method

2.3 Discussion

This chapter has introduced a number of acoustic parameters and explained their measurement methods. It has highlighted a number of issues that prevent occupied measurement. High sound pressure levels are required for good signal-to-noise ratios which requires safeguards for the audience. Multiple measures, with different locations for source and receiver, are required and generally for each position a number of averages are also required. This makes the task very time consuming and highly inconvenient for the audience. All these factors, coupled with the desire to measure the occupied state, motivate researchers to achieve accurate parameter estimates using only naturalistic sounds such as speech or music. This would facilitate in-situ measurement with the audience unaware of the procedure.

In addition to strengthening the motivation for the work, the parameters have been defined and the subjective difference limens for each parameter listed. The difference limens for each parameter define the required accuracy for any measurement system, and are therefore used throughout this thesis to validate each method.

3 REVIEW OF EXISTING OCCUPIED ACOUSTIC PARAMETER ESTIMATION METHODS

Successful occupied measurement of room acoustic parameters can be achieved either by modifying specialist test signals, so they are imperceptible, or by utilising naturalistic signals such as speech or music. When naturalistic signals or imperceptible signals are used, this provides the opportunity to measure in-situ without disturbing the audience. This approach could facilitate the estimation of changes to acoustic conditions in real time. Methods that estimate parameters without explicit knowledge of the input signal are known as blind methods. They are blind, as the non-reverberated signal is not available and therefore all parameters must be calculated without 'seeing' the originally transmitted signal. This chapter briefly reviews previous attempts at performing occupied measurements. The existing methodologies can be broadly classified as follows: 1) standard measurement methods at imperceptible levels. 2) 'Stop-chord' based methods, 3) model based methods, 4) machine learning methods, 5) pitch estimation based methods, 6) blind source separation based methods, 7) cepstrum based methods.

3.1 Standard methods at imperceptible levels

The maximum length sequence method for measuring the RIR, described in 2.1.1, appears to provide an opportunity for measuring the RIR without the audience being aware of the signal. Schroeder [28] used low level MLS signals during a lecture, and by averaging many times, an estimate for the early part of the impulse response was made. Other authors [35, 36 and 37] have continued this work to facilitate measurements of the whole decay curve using low level MLS signals. Unfortunately, due to the requirement of the MLS signal to be significantly lower than the music or speech, so that it is imperceptible, any acceptable parameter estimations made while an orchestra or speaker was performing would take too long. For example, even with a relatively high MLS signal (for example 30dB lower than a music signal at 75dB) a single measurement for Rt would take at least 11 hours [2].

3.1 'Stop-chord' methods

First attempts at measuring the decay curve using naturalistic signals used stop-chords. These are chords sounded by an orchestra simultaneously which terminate sharply. A particularly popular music passage that contains such feature is Beethoven's fifth symphony which encompasses the archetypal four note sequence. Cremer and Müller suggested that the reverberation time could be estimated from the decay curve following a loud stop-chord from this symphony [38]. Other authors have attempted similar procedures using stop-chords to gain approximate decay curves [1, 39]. Measurements using this technique are generally inaccurate for a number of reasons. Firstly, the stop-chord does not fully excite the room and the level does not reach a steady state before decaying. This means that the resulting decay curve is somewhere between the true decay curve and the level decay from impulsive excitation. In addition, the termination of the musical note is not precise as the musical instruments will not cease simultaneously and this persisting envelope could skew the results, and also music is not ideal for measuring the response as not all frequencies are sounded therefore the response measured is biased towards those frequencies that are.

Despite its limitations, a number of researchers have used this methodology to develop new algorithms for estimating the acoustic parameters from 'stop-chords' or regions in music or speech where the source signal is silent and the reverberation can be recorded. Cox *et al.* [40] used statistical machine learning to determine reverberation parameters from the decay phases after separated speech utterances. Baskind [41] estimated the decay curve by simply performing backwards integration on selected decay phases from speech signals. This method required the decay curves to be very uniform (which is not true for many rooms) and that the speech contains significant portions of silence. Baskind comments that the success of such a method is highly dependent on the pre-processing stage where the decay phases are selected. However, Baskind manually selected decay phases and thus the method is not totally blind.

Vesa *et al*. [42] developed a similar method where decay phases were manually selected and then backward integration was performed on the segments. The Rt was estimated

via a statistical analysis of the distribution of results. The unique aspect of the work lay with the use of a short term coherence function between binaurally recorded signals. This enabled Vesa to determine the 'diffuseness' of the sound field and thus specify upper and lower limits on the Schroeder backwards integration avoiding the direct sound, early reflections and the noise from the backwards integration.

Vieira [43] proposes another algorithm, similar to the one used by Baskind, which performs estimation on decay phases between speech utterances. The method differs in the manner in which the decay curve is calculated from the decay phases, and an alternative algorithm proposed by Xiang [44] is used. This method, rather than carrying out a straight line fit to the backwards integrated RIR as suggested by Schroeder, performs a non-linear regression which tries to compensate for the noise floor often seen at the end of RIRs.

Hansen [45] developed a method to calculate the Rt from the envelope of the auto-correlation function of reverberated music signals. The autocorrelation function of music signals is non-diminishing and therefore not ideal, so Hansen searched the music for more suitable regions. Hansen located musical notes that decayed rapidly in comparison to the rest of the signal. By calculating the reverberation time from the autocorrelation function of two of the fast decaying sections, Hansen yielded an estimate for the Rt. This methodology is very prone to errors due to the small number of decay phases used in the estimation and the reliance on suitable decay phases existing in the signal. The success of any 'stop-chord' method is dependent on a number of factors. 1) The success of the automatic determination of suitable regions of decay to gain estimates. 2) Choosing suitable statistical models and tools to correctly quantify the decay characteristics from numerous decay phases accounting for the presence of noises and source fluctuations. These two points are addressed in later chapters within this thesis, overcoming these issues forms a large part of the research carried out.

3.2 Model based methods

Ratnam *et al*. [3, 46] recently proposed a blind estimation method by performing Maximum Likelihood Estimation of decays following speech utterances. Although only one real speech example was presented, the results and the theoretical framework show

the potential of the MLE approach to blind estimation. However, the single exponential decay used is an idealised model based on the assumption of a diffuse field. In reality, non-exponential or weakly non-exponential decays are commonplace. This is why other reverberation parameters, most notably the Early Decay Time, are needed for room assessment. As shall be shown later, the single exponential model limits the accuracy of estimation, as a result, even with white noise burst excitations, errors found exceed perception difference limens.

Couvreur *et al.*[47] developed a method which uses a Hidden Markov Model (HMM) to model equivalent continuous sound level (Leq) sequences of anechoic speech signals. Using the HMM model and a very simple model of sound decay, the parameters of both were estimated via a maximum likelihood framework. The parameters of the room model can then be used to compute the Rt. The method assumes a very simplistic model of sound decay (purely exponential decay), and therefore is limited to only estimating Rt and to rooms with uniform decay curves.

3.3 Machine learning based methods

As discussed previously, Cox *et al.* [40] developed a machine learning approach to parameter measurements in which reverberation times were accurately extracted from speech signals received in spaces under investigation. This approach was limited by the quality of the data set, which in this case was generated stochastically. Therefore while accurate measures were gained, the method has not been validated for real room which exhibit more complex reflection patterns. Li and Cox [5] developed a method to determine the Speech Transmission Index (STI) from received running speech. Both of these methods are quasi-blind. Source signals do not need to be monitored during measurements, but are required during the training phases of machine learning algorithms and are high accuracy estimation methods intended for room acoustics measurement applications. With slightly compromised accuracy they can be made completely blind for other applications where only coarse estimates are needed [6]. The method is limited to speech signals as, when applying music signals to the methodology, the accuracy is compromised further. Overcoming the problems with regards to music signals, using this method, would make the method much more versatile in its application and as such is one area focused on in this thesis.

3.4 Pitch estimation based methods

Reverberation corrupts the harmonic nature of speech signals of speech. Wu *et al.* [48] utilised this to develop a measure of reverberation derived from the 'pitch strength' of speech signals. The pitch period is detected using a Hidden Markov Model based pitch tracker. The tracker is robust under reverberant conditions. The HMM pitch estimate is compared with the pitch in a normalised correlogram of the signal. The distribution of deviations from the HMM estimated pitch period is used to compute the reverberation. This distribution spreads out when the reverberation time is longer. The relationship between a measure of distribution spread and reverberation time is empirically calculated. This method is limited to short reverberation times due to the degradation in the performance of the HMM pitch tracker with increased reverberation.

Baskind [49] also developed a method based on pitch tracking of music signals. The pitch tracker was used to estimate the fundamental frequency of a note which was in turn used to design a comb filter which only passes that particular harmonic series. The novelty of this method lies with the use of the comb filter. When a note ceases, the reverberant decays are generally masked by subsequent notes. The comb filter removes these masking notes as long as a single instrument is playing and the subsequent note is different. This facilitates Rt estimations to be made on the masked decay phases. A short term coherence function was used, similar to Vesa [42], to detect the onset and end of each reverberant decay phase, and then backwards integration was carried out. This method has a number of limitations. Firstly, it requires the music to be monophonic i.e. a single instrument, multi-pitch estimation and comb filter design may be possible but any overlapping overtones in subsequent notes will limit the accuracy. Additionally the methodology means that decay curve estimates are made from single note excitation, this requires the method to search for a range of notes to ensure, when the average decay curve is calculated, that it is not biased by a particular range of notes. Finally, the pitch tracker needs to be reliable under all acoustic conditions, at longer reverberation times the accuracy of the fundamental frequency estimate will decrease. This will cause errors in the comb filter response and results will be compromised.

3.5 Blind source separation based methods (BSS)

Zhang *et al.* [4, 50] developed a method of estimating Rt that utilised blind source separation and automatic noise cancelling algorithms. Zhang used blind source separation to estimate a noise signal from passively received reverberated noise speech signals then an adaptive noise canceller removed the noise from the reverberated speech. A maximum likelihood estimation frame-work was utilised, similar to Ratnam *et al.* [3] which yielded a distribution of Rt estimates from which the final estimate was chosen. This methodology was limited by the performance of the BSS algorithm for which the current state of the art is only successful for short reverberation times.

3.6 Cepstrum based methods

Baskind [41] demonstrated how the early sound field could be estimated using homomorphic deconvolution [51]. Baskind then outlined the principles behind processing the early sound field estimates via a correlogram to yield estimates of the IACC and interaural time differences (ITD) although validation of the method was limited.

3.7 Discussion

This chapter shows that a range of algorithms have been utilised for the purpose of acoustic parameter estimation. However, it is apparent from the literature that no accurate solution to the problem of estimating acoustic parameters blindly from speech and music signals has been published. A common feature of these works is the lack of validation on real or even a range of simulated responses. With the exception of Li *et al.* [6], most only demonstrate estimations from a single room example. Any methodology intended for measurement purposes must be thoroughly validated using a large number of RIR examples.

4 ACOUSTIC IMPULSE RESPONSE DATABASE

This chapter describes the method by which a large database of reverberated music and speech signals was created using both real and artificially simulated impulse responses. The machine learning method developed later in this thesis requires a large number of realistic room impulse responses for training purposes. By using a more realistic training database the resulting measurement system will be more applicable to real rooms. Additionally, validation of any measurement method requires the method to be evaluated using wide range of examples. As it is impractical to measure such a large number of responses, a geometric room model was developed which can generate a set of feasible geometries at random. The resulting geometries were used in a commercial acoustic prediction software package [52] to generate a large number of impulse responses.

4.1 Building a database of impulse responses

The most rigorous validation of a blind estimation method would compare the results from any blind parameter estimation methodology from naturalistic sounds recorded in real rooms, with parameters extracted from impulse responses measured in the same conditions. However, this is problematic as it requires a large amount of time and is difficult to carry out. Recording real room impulse responses and later convolving them with anechoic signals is much less problematic. This poses some additional problems such as how to measure the huge number of examples required by a machine learning method, especially in occupied conditions.

Good validation of any blind method requires a wide range of examples with which to analyse performance. For the machine learning method developed in this research, the dataset has to include many thousands of examples far too many to be obtained from real room measurements. Previously [5], stochastically generated impulse responses were used, but in recent years there have been significant advances in the modelling of rooms using geometric algorithms. Consequently, a commercial package [32] with a proven track record that utilises randomized, tail-corrected cone-tracing was used to generate a training set of examples for teaching the machine learning algorithms and

examining the performance of the methods. To create the large number of required room geometries, a parametric algorithm is defined which randomly generates a set of realistic rooms. These geometries are then used in a commercial acoustic prediction software CATT acoustic [52], to generate a set of impulse responses.

4.2 Distribution of sound sources

An additional problem is that impulse responses are generally measured from a single sound source to a single receiver. The sound source and receiver are moved around and a number of RIRs captured. The resultant parameters are averaged to yield an empirical estimate of the acoustic parameters. The reality is more complex, for an orchestra each listener will hear the result of a large number of sources each with different paths and RIRs. It is not feasible to measure all combinations of individual source to receiver impulse responses, which is why the empirical averaging of parameters over a number of source and receiver positions is often carried out. When carrying out real blind measurements the distributed nature of the orchestra, and therefore the large number of source to receiver paths that create the true listener experience, is inherent in the measurement. The individual RIRs that are used to measure the room's acoustic parameters are all superimposed on one another. This has an impact on the subjective evaluation of the sound that is heard by the listener. For example, in a large orchestra the different instrumental sounds will not present themselves at the listener's ears simultaneously. Depending on the orientation of the listener some will be received after others. This could be perceived as a reduction in clarity which, importantly, is independent of the room's acoustic properties and only dependant on the relative orientation of the listener and the orchestra.

This problem is avoided in these simulations by ensuring the sound is from a single source. This is a reasonable assumption for non-amplified speech, but for distributed sources such as orchestras and speakers utilising sound reinforcement it is considered to be a source of inaccuracy.

4.2.1 Simple geometric room model

A geometric room model with variable dimensions is defined, surface properties and source and receiver locations are specified and from this, a large number of realistic room models generated. Impulse responses were then predicted using CATT acoustic for each model. Two room model types were used, a box shaped room and a fan shaped room. These two room models were thought to represent a large number of possible spaces that may be encountered. The basic room models are shown in Figure 4-1 and Figure 4-2.

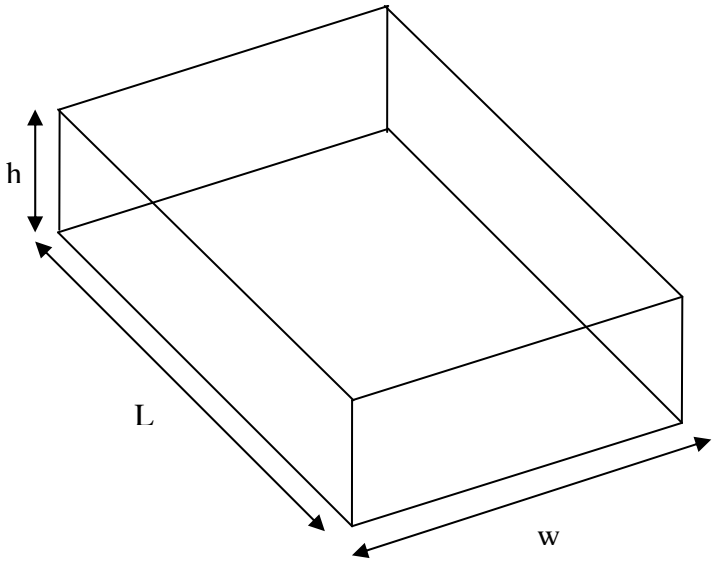


Figure 4-1. Model 1, Box shaped room model.

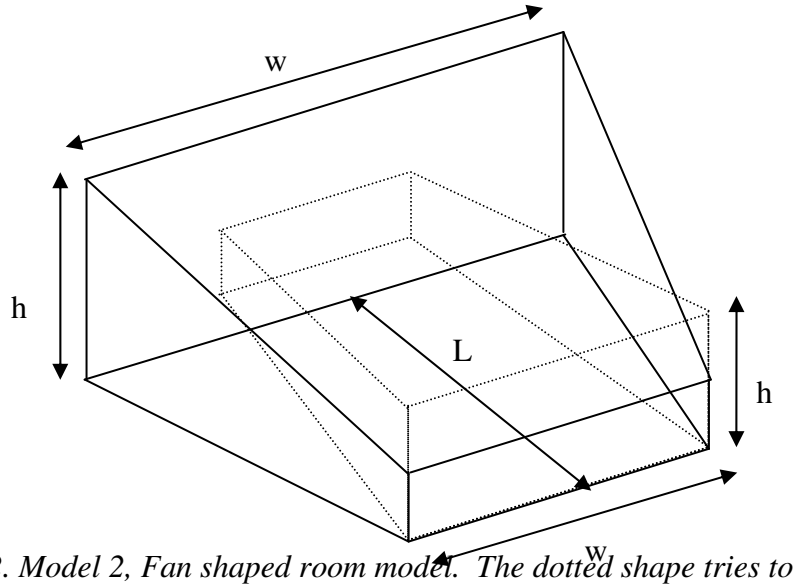


Figure 4-2. Model 2, Fan shaped room model. The dotted shape tries to demonstrate how the geometry may vary, and demonstrates how a sloping floor could occur.

A number of constraints (listed in Table 4-1.) are placed on the geometries and aspect ratios to ensure only realistic shaped rooms are possible. To create the rooms, each dimension is randomly chosen using a random number constrained within appropriate (realistic lengths) bounds. Once the geometries have been generated, further checks are carried out to ensure that the room aspect ratios are within the reasonable bounds. Should the result fall outside of the stipulated bounds, all of the geometries are regenerated and then the check is re-run. This 'brute force' approach is repeated until the room passes all the geometrical criteria. The constraints are all listed in Table 4-1. The reasonable bounds of room geometry were defined pragmatically by considering typical room sizes in the built environment constrain. Room volumes ranged from 75 to 30,000 m^3.

Constraints: Box shaped room	Constraints: Fan shaped room
$3 < h < 10$ m	$3 < h_1 < 10$ m; $3 < h_2 < 10$ m
$5 < w < 55$ m	$5 < w_1 < 55$ m; $5 < w_2 < 55$ m
$5 < L < 55$ m	$5 < L < 55$ m
$h > \max(l,w)/5$	$\max(h_1,h_2) > \max(L,w_1, w_2)/5$
$h < \min(l,w)/2$	$\max(h_1,h_2) < \min(L,w_1, w_2)/25$
$w/L > 0.25$	$w/L > 0.25$
	$0.3 < w_1/ w_2 < 3$

Table 4-1. Geometrical constraints imposed on room shapes, geometry must fall within all of these bounds to be classed as valid. Constraints developed pragmatically by considering typical room sizes in the built environment.

Within each model an audience area was created that was 1m tall, $^2/_3$ the length of the room (with a space at the front representing the stage area) and ¾ the width of the room (all the way along the length of the audience area). Each model had an omni-directional sound source positioned on the stage 2m above the floor and within a rectangular area in the centre of the stage ensuring the source is at least 1m from any surface [8]. A grid of omni-directional receivers were spread over the audience area, one in four of these receivers was chosen at random. Receiver placement close to any surface was avoided (at least 1m from any surface) as recommended in ISO 3382. Receivers were not placed exactly in the centre of the room to avoid any anomalies where reflections exactly constructively interfere.

Appropriate absorption properties (appropriate as wall material properties were only applied to walls etc.) were randomly selected from a database of materials provided with the software. Scattering properties were selected using a function designed to randomly assign scattering levels in each frequency band, once again limits were placed on the function to ensure scattering properties were realistic. The distribution of possible scattering coefficients is demonstrated in Figure 4-3. The distribution of scattering coefficients was chosen pragmatically to represent realisable surfaces. Scattering is often frequency dependant. In order to scatter low frequencies the surface must, in general, have features of comparable size to the wavelength. Therefore it is common so see material scattering coefficients increase with frequency. The randomised scattering function generates a series of coefficients for different octave bands. The function is constrained so that the coefficient always increases with frequency. The possible range of scatter coefficients for each octave band is increased

for each adjacent octave band. The audience surface properties had their own pre-defined scattering properties which were used in the prediction.

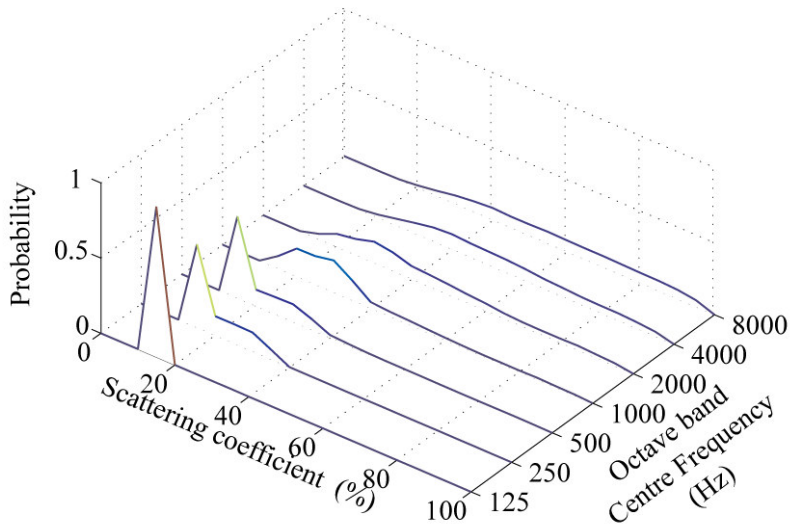


Figure 4-3. Distribution of possible scattering coefficients used in defining the material properties of the room, each line represents the distribution of possible scattering coefficients used in the room model for each octave band.

The audience area was also assigned a surface property. The level of occupancy was varied by selecting audience surface properties representing differing degrees of occupancy (e.g. 0%, 50%, 75% and 100%). Once the material and geometrical properties had been defined, a quick estimation of the reverberation time was calculated (using Sabine's classical formulation formulation). Rooms with an estimated Rt less than 4s were deemed acceptable. If this condition was met, then a successful room design had been achieved, if not then the whole process was repeated.

4.2.2 Prediction of acoustic impulse responses

A large number of valid geometries was generated and from these, a room impulse response predicted using CATT acoustic. The room geometries were generated very quickly, however each room prediction took around ½ hr. Using the sequence programmer built into the software, a number of responses can be generated in sequence. 8400 different room geometries were generated and from each an impulse

response was predicted; 50% of the examples were box shaped rooms and 50% fan shaped rooms. The elapsed time for performing these calculations was significantly decreased by running the prediction software simultaneously on a large number of machines and updating a remote file store via the FTP protocol.

Often longer reverberation times are seen in the results, this is because of the non-diffuse nature of some of the spaces and the inability of the Sabine equation to accurately predict the Rt of these spaces. These cases are particularly challenging for a measurement system and while these cases may be rare in reality, they are kept in the data set so the machine learning algorithm learns from all possible cases which may occur. Results from such a wide range of different acoustic conditions as generated from this database provide a level of robustness and confidence to the blind measurement algorithms developed in this thesis. A second set of impulse responses was also used to evaluate performance, this set consisted of 20 measured room impulse responses from Finnish concert halls [53].

4.2.3 Simulated impulse response example

A fan shaped model is generated and illustrated in Figure 4-4. The randomly selected surface properties are listed in Table 4-2

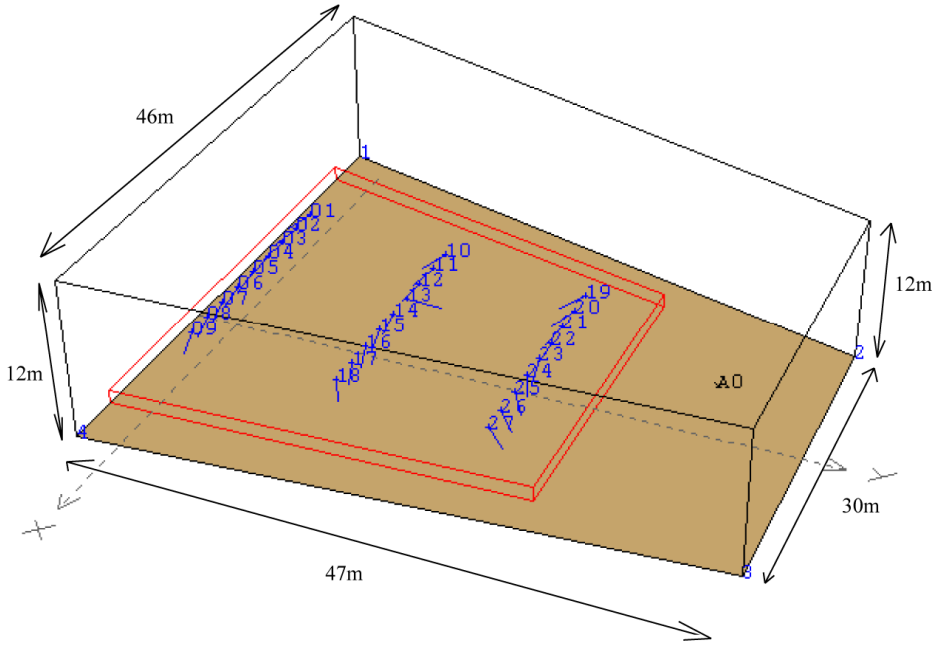


Figure 4-4. View from CATT acoustic showing one of the randomly generated room geometries. The sound source is labelled as A0, the grid of receivers is numbered from 1 to 27.

Surface	Material	Absorption coefficients (%) Octave bands (Hz) 125:250:500:1000:2000:4000	Scattering coefficients (%) Octave bands (Hz) 125:250:500:1000:2000:4000
Floor	Carpet, heavy, cemented to concrete	2:6:14:37:60:65	15:15:15:33:49:64
Walls	lime cement plaster	2:2:3:4:5:5	15:31:31:31:31:65
Ceiling	plaster smooth on lath	14:40:6:4:4:3	15:15:15:23:36:49
Audience area	Medium upholstered concert hall chairs, average. unoccupied	56:64:70:72:68:62	30:40:50:60:70:70

Table 4-2. Surface properties for the room model shown in Figure 3-3, please note 8k and 16k bands are extrapolated from the 2 and 4 kHz octave bands in the CATT software

Figure 4-5 a) shows the predicted 1 kHz octave band impulse response at receiver 14. Backwards integration of the response yields the decay curve presented in Figure 4-5b. The acoustic parameters of this room model are presented in Table 4-3. The decay curve in Figure 4-5b shows a more or less exponential decay and the time delay between the direct sound and the early reflections can be seen in Figure 4-5a. The room acoustic parameters show the room to be very reverberant due to the very non absorbent walls.

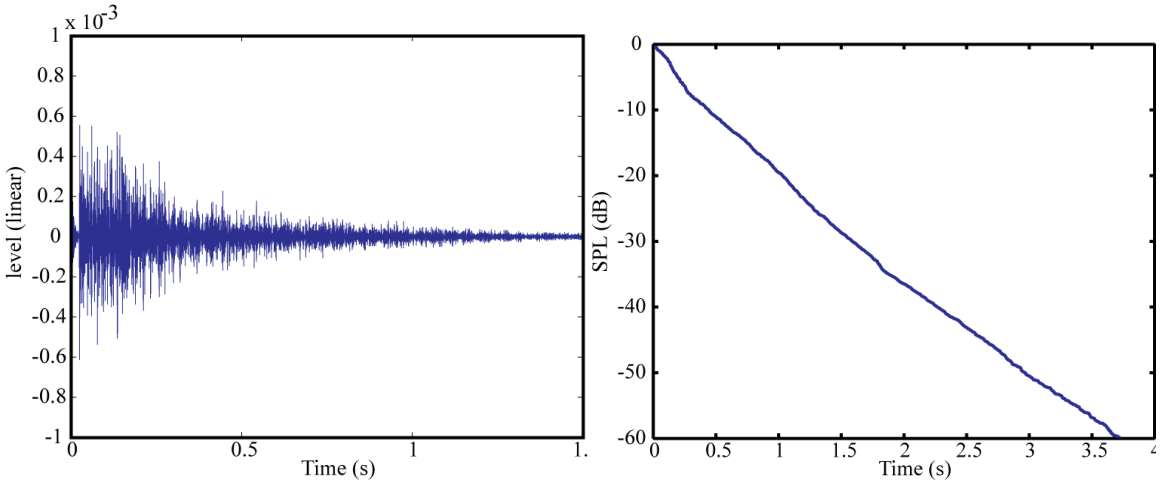


Figure 4-5. a) Simulated room impulse response and b) decay curve produced from simulated impulse response for the example room model.

Octave bands (Hz)	Rt (s)	EDT (s)	C_{80} (dB)	t_s (s)	D (%)
125	4.4	3.7	-2.4	0.252	17
500	4.9	4.7	-2.8	0.242	23
1000	4.9	4.8	-4.1	0.274	21
2000	3.4	2.5	-3.5	0.191	21
4000	2.4	1.8	-0.7	0.134	31
8000	1.8	1.6	-0.3	0.113	35

Table 4-3. Acoustic parameters for the simulated room

4.3 Distribution of acoustic parameters in the impulse response database

It is interesting to look at the distribution of parameters over the whole database of simulated impulse responses and to compare the parameters that occur in real rooms. The following table is presented from work in references [54 and 13]. These parameters refer to concert halls only and the acoustic parameters Rt, EDT and C_{80} have been averaged over a number of octave bands.

Parameter	Range
Rt	1.4s to 2.8s [54]
EDT	1.8s to 2.6s [54]
C_{80}	-5dB to +3dB [13]

Table 4-4. Range of parameters seen in concert halls, C_{80} from 14 European concert halls, Rt and EDT from data collected, worldwide, during the career of Jordan [54.]

Auditoria with Rt and EDT values greater than or less than these values do occur and the simulated RIR database should also include a number of examples with non-ideal responses. In general it is recommended that for the purposes of music the Rt and EDT should not be less than 0.5s [16]. A lower level of reverberation is required for rooms in which speech is the primary focus. The acceptable range of Rt for small to medium rooms for speech purposes is from about 0.4s to 1s according to DIN 18041 [55]. There is small number of spaces which may be much more reverberant. Spaces such as churches and railway stations can have very long Rts. The distribution of acoustic parameters represented by the database generated in this chapter covers the range of parameters seen in reality, Figure 4-6 shows that the acoustic parameter distribution for artificial responses covers the range of values expect in real spaces but the distribution also extends and covers a significant number of 'extreme' cases.

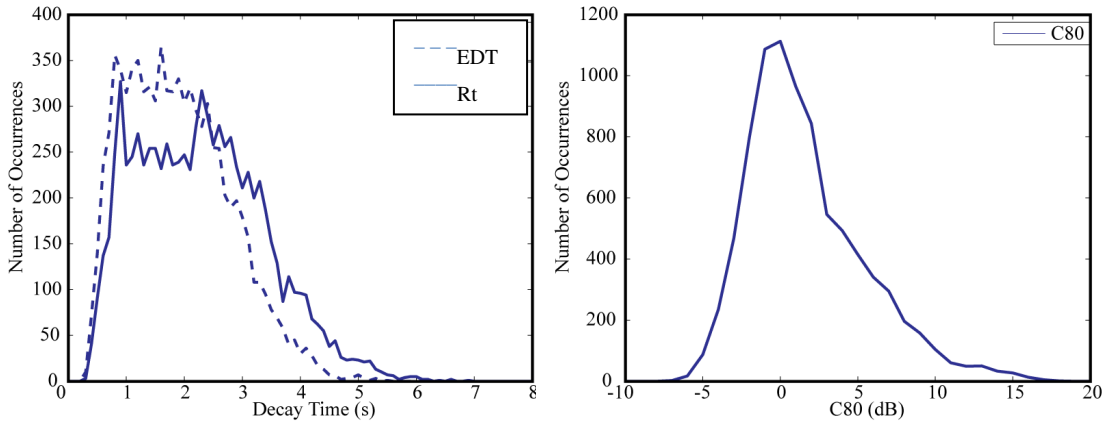


Figure 4-6. Distribution of Rt and EDT and C_{80} values in the database of 8400 simulated room impulse responses averaged over the octave bands from 63Hz to 8000Hz.

Figure 4-7 to Figure 4-10 provide more detail on how the parameters vary with frequency. Parameters such as Rt and EDT are lower at higher frequencies, this is consistent with real rooms, as at higher frequencies there is generally less reverberation, due to the higher level of absorption at these high frequencies. The distribution of C_{80} shows that the mean parameter value increases with frequency. This is because at higher frequencies the late reverberation level is very low in comparison to the direct sound which accounts for the increase in clarity. The centre time shows that at higher

frequencies, the 'centre of gravity' of the RIR is nearer to the start of the response, this is consistent with the quicker decay of sound in higher octave bands.

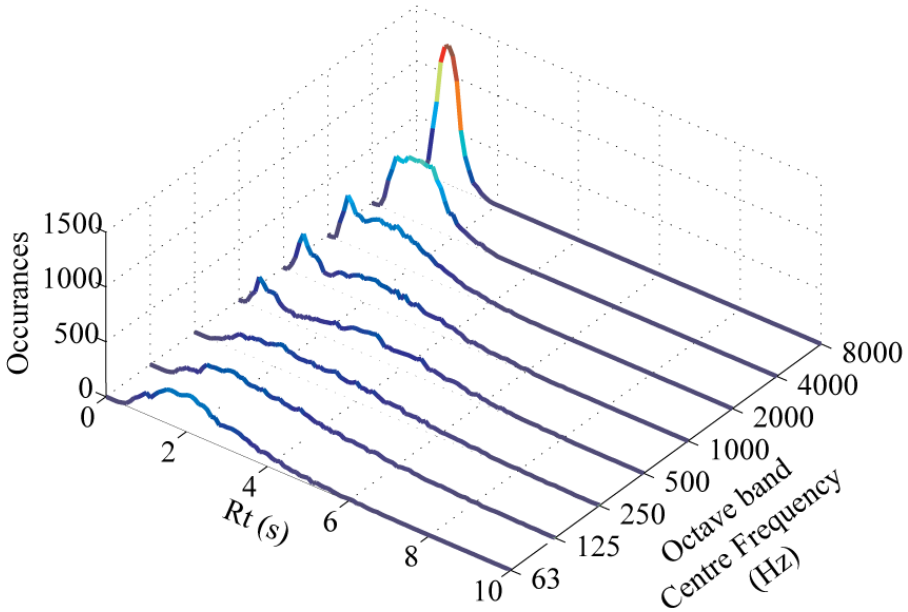


Figure 4-7. Distribution of Rt values in the database of 8400 simulated room impulse responses for the octave bands from 63 Hz to 8000 Hz.

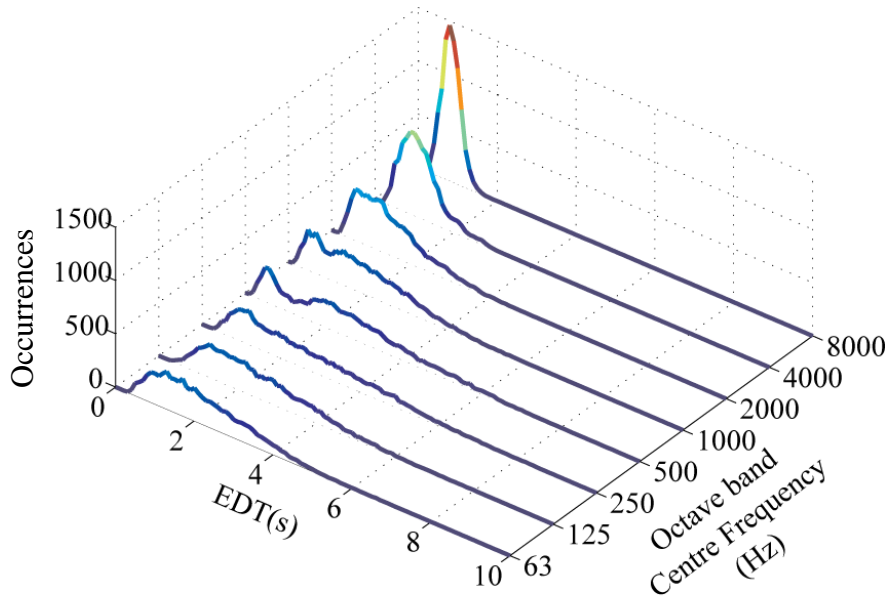


Figure 4-8. Distribution of EDT values in the database of 8400 simulated room impulse responses for the octave bands from 63 Hz to 8000 Hz.

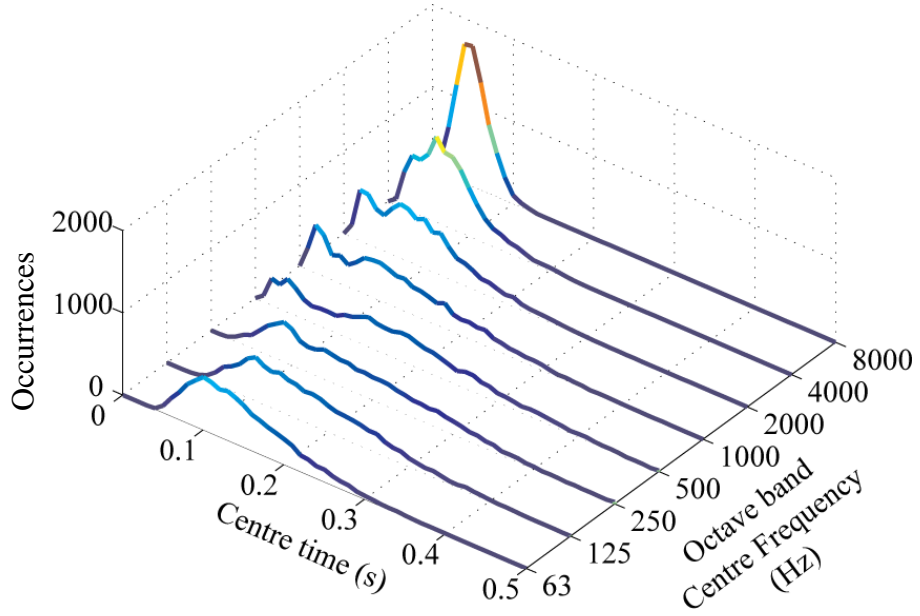


Figure 4-9. Distribution of centre time t_s values in the database of 8400 simulated room impulse responses for the octave bands from 63 Hz to 8000 Hz.

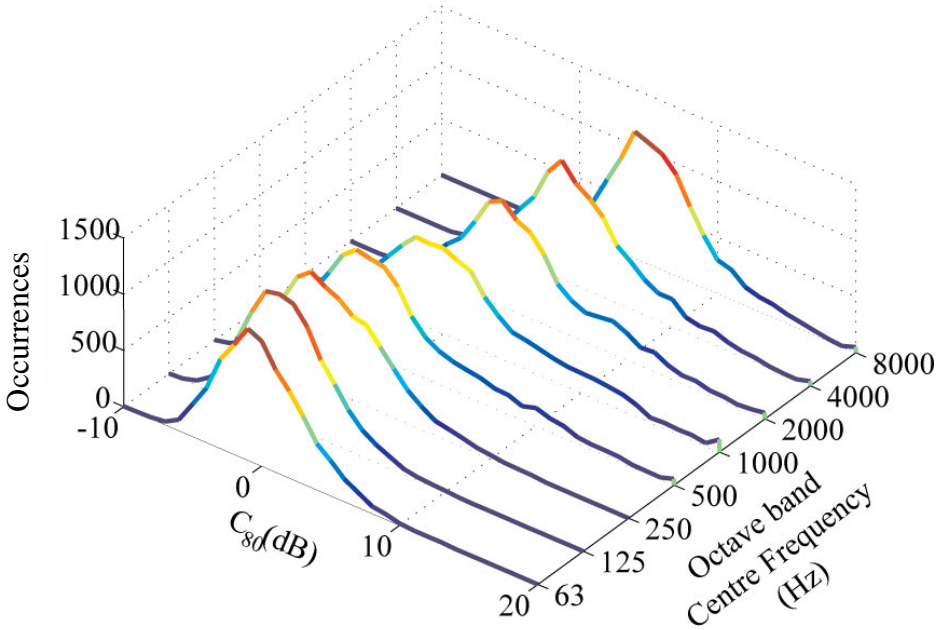


Figure 4-10. Distribution of clarity C_{80} values in the database of 8400 simulated room impulse responses for the octave bands from 63 Hz to 8000 Hz.

4.4 Limitations of the database

The RIR database is intended for two purposes, firstly to provide a large number of RIRs as training data for a machine learning method, and secondly to provide validation for the estimation methods over a broad range of room responses. Therefore the RIRs must be as realistic as possible and be representative of all possible responses that may occur. If the trained machine learning method encounters RIRs not represented within the database, the method will fail. The previous section discusses the success of the database at covering all the possible parameters that may occur in lecture theatres or concert halls. The database also represents RIRs with more extreme parameters, such as very reverberant spaces and very non-diffuse spaces. Despite this the RIR database has a number of limitations which are discussed in this section.

The models are very simplistic in their internal geometrical complexity and all scattering is approximated using scattering coefficients applied to surface materials rather than any physical objects. This has the effect of making scattering quite uniform over a large surface area which could make the room more or less diffuse than a real room with similar geometries. In real rooms, diffusing objects are placed in certain locations to increase the diffusion over the audience area rather than having large areas with uniform scattering properties. This is not expected to be a limiting factor of the database; the limitation only applies when comparing a real room with a simulated room model. As the geometries and surface properties are randomised the database is still expected to contain RIRs that can represent these room responses.

The responses are simulated using the commercial software CATT acoustic which employs cone tracing. There are several of limitations to this method of prediction that must be considered when utilising this database. Geometrical simulation methods such as ray tracing, the image source method and cone tracing do not take into account the wave nature of sound propagation. This means that the methods are limited to higher frequencies and wave effects that occur when sound interacts with features of the room having a similar size to its wavelength, such as diffraction and room modes, are not always modelled. This means that the database is only realistic in the higher octave bands (when the room size is much larger than the wavelength). In practical terms, the

results using these simulated responses are not considered in the lower octave bands (63 and 125 Hz).

The prediction software computes the ray behaviour in discrete octave bands. By calculating the ray behaviour in separate bands the response over each band remains quite uniform, this is not representative of real rooms whose decay rate alters continuously with frequency. One of the reasons for this is that the material properties are presented to the simulator in octave bands. This is generally not a problem for this research as the simulations are used for predicting parameters which are also measured over octave bands. One potential issue may arise is when the response is significantly different in adjacent octave bands. Octave-band filters have a finite roll-off and therefore acoustic parameters can be influenced by the out-of-band response. By ensuring the surface properties avoid sudden jumps in value between octave bands these discontinuities in acoustic parameters have been minimised. (The surface absorption properties were selected from a database of real measured values which exhibited relatively smooth change of absorption over the frequency range and the scattering coefficients were generated using an algorithm that tried to avoid sudden changes in scatter properties between octave bands.)

One feature of the database is the significant number of non-uniform decays where the EDT and Rt values differ significantly. With a machine learning method, these non-uniform decays may cause the learning architecture to learn complicated features that may not occur that frequently in reality. However, non-uniform decays certainly do occur and their over-representation in the simulated database enables the machine learning method to deal with the occasional ‘odd response’.

4.5 Discussions

This chapter has outlined a method for generating a large database of RIRs. Sets of room geometries and surface material properties are randomly generated within a framework, limiting the models to realistic geometries. The two room model types (fan shaped and rectangular) represent a large number of room configurations representing a high proportion of the shapes that are actually found for many concert and lecture halls.

The range of parameters that the simulated RIRs exhibit easily covers those that may be expected from most real rooms in addition to a number of potentially 'odd responses' from non-diffuse spaces. The simulated RIR database limitations are not thought to be significant and the RIRs generated are representative of a very broad range of possible responses.

5 OPTIMISATION AND ARTIFICIAL NEURAL NETWORKS

This chapter introduces the subjects of optimisation and Artificial Neural Networks (ANNs). Optimisation and Artificial Neural Networks are utilised in the acoustic parameter estimation methods presented in this thesis and therefore chapter introduces the underlying principals. Typically, a room response model contains parameters which change the behaviour of the model. Optimisation is used to determine the 'best' model parameters which reproduce available experimental room response information. The function can occupy a large, multi-dimensional parameter space and have numerous solutions. This makes the problem of finding the true 'globally optimal' solution, rather than a 'locally optimal' solution, difficult. This chapter discusses the various schemes that have been developed to find global optima, and their advantages and disadvantages. Also in this chapter is an introduction to machine learning techniques using artificial neural networks, how they work and their peculiarities. The purpose of the ANN method is somewhat different to optimisation, as there is initially no underlying model. Instead the model is automatically developed during the training of the neural network.

5.1 Optimisation

Later in this thesis a method for estimating acoustic parameters from received signals using a model of sound decay is defined, this is called the maximum likelihood method. This method uses the reverberant decays in-between speech utterances and formulates a likelihood function incorporating the sound decay model. The input to the likelihood function consists of the received reverberant sound and all of the model parameters. As the parameters are unknown, the likelihood function must be 'optimised' so that the most likely set of model parameters (e.g. decay time) responsible for generating the decay are found. In the case of a likelihood function this is a parameter set that produces the maximum function value for all possible parameters. Locating the optimum of a complicated, multi-dimensional function is a difficult problem. This is a separate area of research in which many researchers are searching for faster and more accurate methods of optimisation. There is no ultimate optimisation algorithm and each

method has its pros and cons. A good literary source that explains the many facets to optimisation is contained in [56].

It is important to note that maximising a function is equivalent to minimising the negative of the function. In the case of maximum likelihood optimisation, as presented later in this thesis, the optimisation minimises the negative likelihood value.

5.1.1 Optimisation overview

A multi-dimensional function, such as a likelihood function, can be thought of as the landscape of an undiscovered world. Upon arriving, how does one find the deepest valley? One could travel to all points in that world measuring the height at each. This, however, may take an extremely long time to complete (this is analogous to the *direct search method* mentioned in 5.1.3). Alternatively one could start walking downhill and continue until there is nowhere to go except down (this is analogous to the *gradient search method* mentioned in 5.1.4). However there is no guarantee that this peak is 'globally' the deepest ravine, it may merely be a small 'local' valley. More intelligent methods of searching could be employed, such as looking into the distance, using rough initial estimates of the landscape, or perhaps using satellite technology to gleam a rough overall impression of the landscape.

The problem described in the previous paragraph is analogous to the investigator's search for the global minimum of a function. Search methods can often get stuck on local peaks in the function; to prevent this, algorithms must use more intelligent search methods. One may wish to evaluate the function at every set of parameter values (grid co-ordinates) but this is untenable as the parameter space may be very large and as the number of variables (dimensionality of the function) increases, the grid size increases. For instance, if the function is N dimensional and the problem requires 100 steps to explore each dimension, the problem will scale by N^{100}.

5.1.2 Types of Optimisation algorithms

There are a number of types of optimisation methods two of which include *direct search methods* and *gradient search methods*

5.1.3 Direct search methods

Direct search methods use only function evaluations. Different direct search algorithms explore the search space in various ways to find the solution. A simple, commonly used search method is the *brute force* method where the function is evaluated over a large grid of parameter values. This can be a very computationally expensive method, but it has the advantage of knowing that the whole parameter space has been explored. The *random walk* method is another direct search method. First a starting point is chosen and the function evaluated at that point, then a new location is chosen by adding a randomly generated value to the parameter(s). The two locations are then compared and the most optimal location is chosen as the new position. Next a new random direction is generated and the process repeated. This method is not reliable in practice as successful global optimisation requires a careful choice of the statistics (mean/standard deviation) of the random walk direction. Another direct search method known as the pattern search algorithm [57] (or the Hooke-Jeeves algorithm) chooses a starting point, then evaluates the function at trial points in all N positive and negative (N number dimensions) axis directions. The algorithm then moves to the most optimal location and repeats the process. If a more optimal solution is not found, then the operation is repeated with a smaller step size. In general, direct methods are not as efficient at finding a local optimum solution as methods which use gradient information, but they are often used in hybrid global optimisation methods [82]. The *brute force* approach can be particularly useful when the parameter search space is limited and the required accuracy of the estimates is finite. A grid of parameters over the search space range with resolution according to the required accuracy can give a small grid of parameter estimates that is computationally efficient to evaluate and indicates the optimal solution to within a specified accuracy. In this thesis a *brute force* method is used to provide rough estimates for parameters prior to determining the local minimum with a more accurate method. Other properties of the function can be useful in reducing the search space, for example functions may be symmetrical, as is the case for the likelihood function developed later. This effectively halves the search space.

5.1.4 Gradient search methods

Where derivatives of the function with respect to the parameters are available, gradient search methods can be used. First and second order derivatives contain information about the local direction of the function and a search direction can be chosen accordingly. Examples of gradient search methods include the famous *steepest descent method* and the *Newton-Raphson method*. In general they are very efficient at determining a local extrema. These choose the locally optimal search direction at every iteration, but can only ever guarantee convergence to local extrema. Gradient search methods will be utilised in this thesis to provide quick convergence to a local optimum. As previously mentioned the starting point for the optimisation will be determined from a *brute force* evaluation of the function at a number of chosen grid points, this helps to avoid local peaks and find the optimal solution for a given parameter search space.

5.1.5 Simulated Annealing

Simulated Annealing (SA) [58] is a scheme to avoid getting stuck at local optima and can be applied to many optimisation methods. SA allows the search method to choose 'up-hill' directions i.e. less optimal solutions. The probability of this being allowed decreases in proportion to the difference between the current function value and the function at the 'up-hill' position. The probability of an up-hill direction being chosen also decreases with a temperature parameter which has a value that decreases with time. In other words after a reasonable length of time SA forces the search algorithm to look more locally. The algorithm is called Simulated Annealing as the method is analogous to the process of the controlled cooling of a molten material to increase crystal size i.e. reach a minimum energy configuration. Whilst simulated annealing methods are global optimisation methods, there is no way of determining if they have sampled sufficient of the parameter space to include the global minimum.

5.1.6 Evolutionary Algorithms

Evolutionary Algorithms are a set of optimisation algorithms that use a Darwinian based decision process to choose more optimal solutions. After a starting point is defined (initialisation – or birth in this case!) the algorithms use mutation and recombination to gain new estimates and a selection process selects the fittest result. These processes are

repeated until some stopping criteria are reached. Examples of evolutionary algorithms include Genetic Algorithms, Evolutionary Strategies, Differential Evolution and Evolutionary Programming. As with all global optimisation methods, evolutionary methods have no real convergence criteria. It is not possible to state when the global optimum solution has been found. It is also thought that this methodology may not be particularly computationally efficient as the number of function evaluations may be large. Due to the large number of optimisations required, speed of the optimisation is very important.

5.1.7 Constrained optimisation

A further distinction in type of optimisation is whether there are any bounds on the parameter values. The optimisation is either constrained or un-constrained. Constraints can be placed on parameters to ensure compliance of the solution with a physical reality. An example of a constraint to a physical reality, in the case of reflection decay estimation, would be ensuring a model cannot allow the build up of sound to occur when no sound source is present. Price [59] comments that functions involving constraints can often be intractable as they have many local solutions and interacting mixed-type variables. Despite this, constrained optimisation will be utilised to prevent the search algorithm yielding unrealistic solutions. The parameters will be constrained so that only realistic sound decays can be considered as solutions. It is hoped that constraining the algorithm to physical reality will yield fewer problematic parameter estimates.

5.1.1 Choice of algorithm

The choice of algorithm is dictated by the type of problem that needs to be solved. The search space for the globally optimal model parameters is finite because the parameter space is defined by the range of possible decay behaviours that may be encountered in real rooms. For example, reverberation times greater than approximately 15s are generally not realisable due to air and surface absorption. This places a lower limit on the possible rate of decay that may be encountered. This steers the choice of algorithm towards constrained optimisation, to ensure the model parameter estimates fall within the bounds of the physically realisable. Additionally, the desired parameter estimation accuracy is not infinitesimal but related to the subjective difference limens. Therefore a

brute force approach is adopted to provide a quick, coarse parameter estimate prior to performing gradient based constrained parameter optimisation. This combination of techniques can help ensure that the parameter estimates will be optimal to within a specified accuracy.

5.2 Introduction to Artificial Neural Networks

The human brain can solve very complex and difficult problems and store huge amounts of information. The mode of operation of the human brain is very different to that of a conventional computer. Neural networks are vast networks of interconnecting processing elements called neurons. In biological neural networks these neurons are living cells. The neural network is very powerful and versatile due to its massively parallel nature and its ability to learn and adapt.

The development of the artificial neural network was driven by the desire to harness this computational power and adaptability, and as such the architecture and functional principals of biological neural networks were studied and modelled. This section gives a brief overview of neural networks and details the operation of a few training algorithms. For a good review of the historical development of ANNs refer to Haykin [60].

5.2.1 The Neuron Model

The starting point in the development of the ANN was the development of a model of the most basic component of the neural network, the neuron. The development was initiated by McCulloch and Pitts [61] in 1943 who developed the first artificial neuron model. The basic structure of the McCulloch and Pitts neuron is still used today. The artificial neuron has two stages. The first stage takes the inputs to the neuron, weights each input differently and then combines them in some manner, this is called the basis function. The basis function can also apply a bias value to the weighted combination of inputs. In the second stage, the combined inputs are passed through another function known as the activation function. This structure is shown in Figure 5-1.

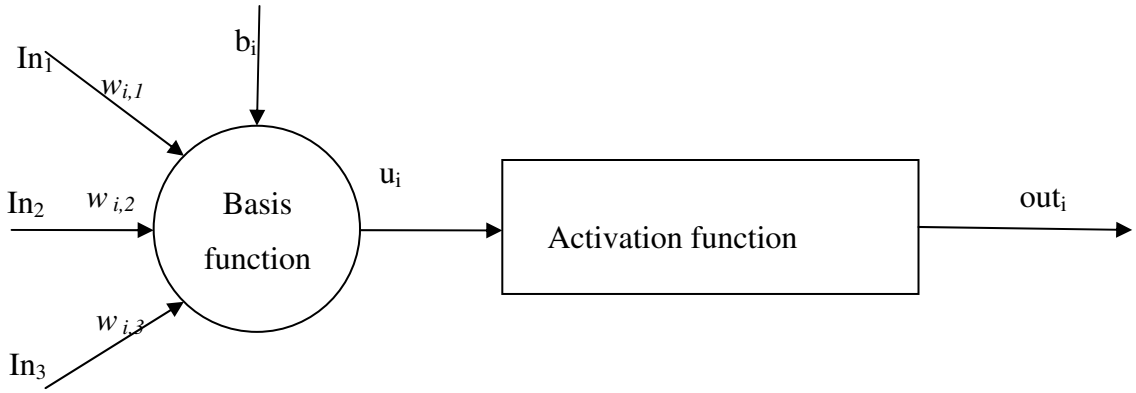


Figure 5-1. Artificial neuron model

McCulloch and Pitts' neuron used a simple weighted sum of the inputs as the basis function. The linear basis function is written as follows;

$$u_i = \sum_{j=1}^{N} w_{i,j} In_j + b_i$$

(5-1)

where i represents the i^{th} neuron, In_x represents input number x to the neuron, u_i is the signal combined by the basis function, $w_{i,x}$ are the weights for each input, b_i is the basis value and out_i is the output of the neuron.

McCulloch and Pitts' original neuron model used a simple threshold activation function (also known as a Heaviside step function) which outputs one of two values depending on the input. Other activation functions can also be used, such as sigmoid functions:

$$f(u_i) = \frac{1}{1 + e^{-au_i}}$$

(5-2)

Or linear activation functions:

$$f(u_i) = \begin{cases} 1 & \text{if } u_i \geq 1/2 \\ x & \text{if } 1/2 > u_i > -1/2 \\ 0 & \text{if } u_i \leq -1/2 \end{cases}$$

(5-3)

Figure 5-2 compares the threshold and sigmoid activation function. Continuous activation functions, such as the sigmoid function, allow nonlinear mapping of a continuous function from once space to another. Therefore the sigmoid activation function will be used in this thesis in order to perform the nonlinear mapping between the reverberant signal and the acoustic parameter being estimated.

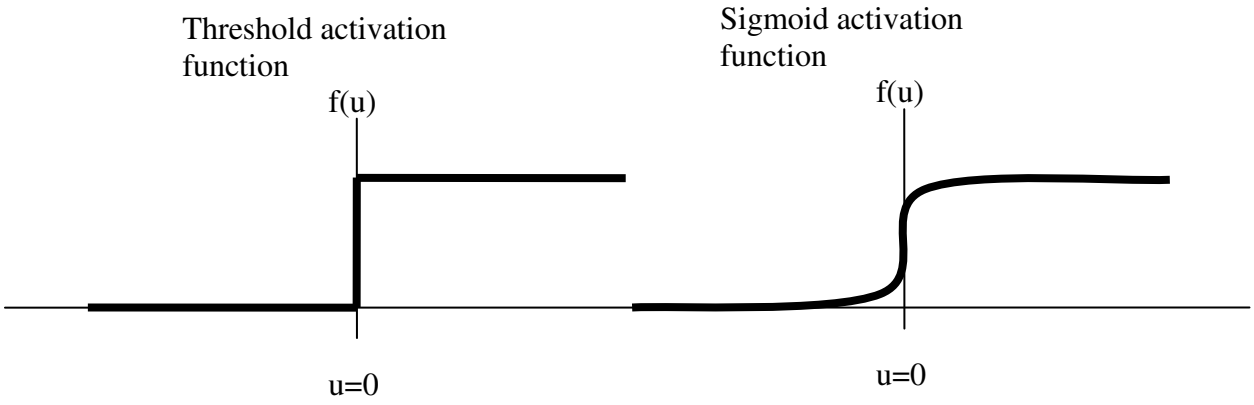


Figure 5-2. Examples of two types of activation function: 1) threshold activation functions can take one of two values, 2) sigmoid activation functions are continuous.

5.2.2 Network architecture

A single neuron is a very simple function with limited processing ability. It is not until many such neurons are combined into an interconnected network that the powerful computational ability is realised. The perceptron was one of the first neural network architectures. It was first developed in 1958 by Rosenblatt [62]. This network has a single layer of interconnecting neurons with linear basis functions and threshold activation functions. It is used as a simple classifier to differentiate between two groups of data. The perceptron can only classify between datasets when the data is separable in a linear fashion (equivalent to finding a straight line that can separate two clusters of data points). Multiple layers of neurons are often used to increase the processing ability

of neural networks. A multi-layer neural network architecture features a number of layers of interconnecting neurons in sequence, where each layer feeds the results from the previous layer forward to the next layer. Hence these networks are known as feed-forward neural networks. Figure 5-3 shows an example of a multi-layer feed-forward ANN.

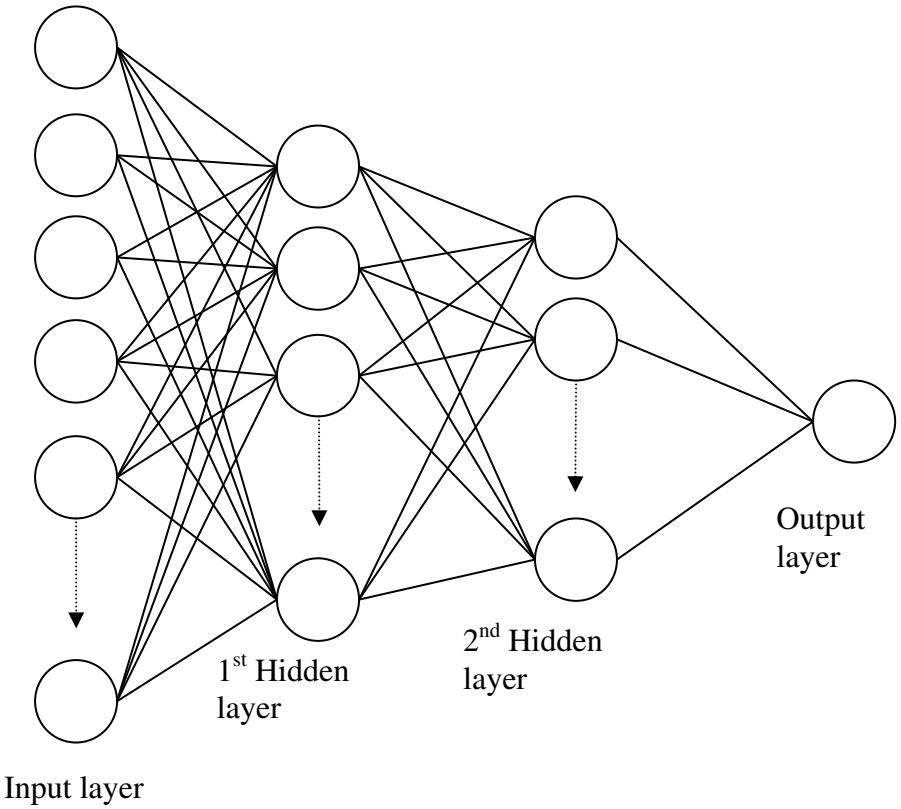


Figure 5-3. Multi-layer feed-forward neural network architecture

Within this thesis, 10-3-1 is used to describe, for example, an ANN with 10 inputs, a hidden layer with three neurons and a single output. The size of the ANN must be appropriately selected for the task. Table 5-1, replicated from [63] describes the mapping ability of ANNs with different numbers of hidden layers. Heaton also points out that ultimately the network architecture will be determined by trial and error. An ANN with two non-linear hidden layers is often referred to as a universal approximator. This powerful statement refers to the network's ability to map any function, provided it is sufficiently smooth and the training dataset is rich enough.

Number of Hidden Layers	Result
none	Only capable of representing linear separable functions or decisions.
1	Can approximate any function that contains a continuous mapping from one finite space to another.
2	Can represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy.

Table 5-1. Determining the Number of Hidden Layers, replicated from [63].

5.2.3 ANN learning algorithms

Once the neural network architecture has been defined the next stage is to train the network to perform the desired task. There are two broad types of training algorithms, supervised or unsupervised learning (reinforcement learning is another paradigm but this is not covered here). Unsupervised learning is a technique where features of the data set are learnt without a known output. An example of an unsupervised neural network is the self-organising map (Haykin p465 [60]) . The self-organising map can take high dimensional data and convert it into a lower dimensional representation, often in the form of a two dimensional map where similar data sets are clustered together. They are often used as a visualisation tool to identify trends within a dataset.

Supervised learning involves training a network to map a series of inputs to a series of outputs where the ANN is trained using known outputs. The desired outputs are referred to as teacher data. The output of the ANN is monitored and for every input data, the output is compared with the teacher data. The training procedure aims to minimise the total error in the output for all possible inputs. By minimising this error, the ANN's aim is to replicate the teacher data when presented with an input vector. To achieve this, neuron weights and biases must be updated within the ANN accordingly. An algorithm is required that uses the output error to update the weights in a manner that minimises the squared error. The Delta learning rule, developed by Widrow and Hoff [64], is one such algorithm that was developed to train a single layer perceptron. The delta learning rule uses the difference between the ANN output and the teacher, $e_i=(teacher_i - out_i)$, to update the neuron weights w_{ij}

$$\Delta w_{ij}(n) = \eta\, e_i(n)\, In_i(n)$$

(5-4)

where n refers to the iteration number in the network training process and η is a positive constant value that sets the rate of learning. This value is important as if it is too large the algorithm may miss the optimal solution and if it is to0 low the training process will be very long. The network weights are updated at each iteration.

$$w_{ij}(n+1) = w_{ij}(n+1) + \Delta\, w_{ij}(n)$$

(5-5)

The Delta rule is limited to training single layer perceptrons, to extend the method to multilayer neural networks the backpropagation algorithm was developed by Rumelhart *et al.* [65]. This is a generalisation of the delta learning rule and can be applied to multi-layer networks with non-linear, continuous activation functions. The total squared error for all the training data is minimised in a similar manner as the delta rule but here multiple layers must be updated. The backpropagation algorithm updates the output layer first using *(5-4)* and *(5-5)*. The error is recalculated and the next layer back is updated in a similar manner. This continues back through each layer of the network to the input layer, hence the name backpropagation.

These methods are gradient descent algorithms, as the weights are updated according to the negative gradient of the global error, this directs the error towards a local minimum. In this respect the problem of training an ANN is similar to that of optimisation, the desired solution being the global error minimum rather than local solutions. There are several training frameworks that have been developed to avoid local minima and find the global minima. A momentum term can be added to the backpropagation algorithm [65]. The momentum term uses previous values of the rate of change of neuron weights to influence the rate of change of the current weight calculation. This provides the algorithm with a memory of the previous speeds of convergence and the learning algorithm has 'momentum' to continue searching in promising directions while 'skipping' past sub-optimal solutions. Many more methods of network training have been developed in recent years and these broadly fall into two categories 1) Heuristic

methods which build on the standard backpropagation algorithms such as using an adaptive learning rate or momentum. 2) Methods that utilise optimisation algorithms such as the conjugate gradient method. In this thesis methods utilising gradient based optimisation algorithms were utilised as they offer a large order of magnitude increase in speed compared with heuristic methods. Neural network weights must be initialised prior to training; it is most common to initialise each weights using a uniformly distributed random number between 1 and -1; this approach is adopted later in the thesis.

5.2.4 ANN applications

There are two broad categories of applications for supervised neural networks

1. Classification.

2. Function approximation.

Classification involves separating the data into discrete groups or classes. An example of a classification task would be determining the colour (e.g. red, green or blue) of an object from a digital image. Function approximation the most appropriate methodology for learning continuous acoustic parameters from speech and music signals.

It is usual when working with ANNs to split data into multiple data-sets. The training of the network is carried out on one set called, appropriately, the training set. One feature of a neural network is the ability to adapt to new situations. When a correctly trained network is presented with new data from the same global population as the training set, the ANN can still perform the regression or classification task. This is known as generalisation. The ANN uses features of the data-set, empirically learned during training, to classify the data. To test the network's ability to generalise, another data set called the test set is fed through the network (without updating the weights). By analysing the error on the test set, and comparing with the error on the training set, the generalisation ability of the ANN can be quantified.

Divergence between the training and test set total squared error indicates overtraining. Overtraining occurs when the training set is not representative of the global population

of possible data inputs. It can also occur when the network is larger than the problem requires. The training process finds a way to minimise the error that is not representative of underlying relationships between features in the data set. For example, in the colour classification task, the network is taught to classify the colour of a series of pictures of fruits and the training set yields perfect results, i.e. the banana is classified as yellow, the apple as red and the pear as green. However, when presented with an orange, the ANN classifies it as red because it is the same shape as the apple.

To prevent overtraining, another unique data set is defined called the validation set. During training, after all training data has been passed through the network and the weights updated, the validation set is fed through the network and the error recorded, but the weights are not updated. By comparing the error in the training and validation sets overtraining can be detected, as the error in the validation set will begin to rise as the training error continues to fall.

5.3 Discussions

This chapter has introduced concepts relating to optimisation and neural networks. Optimisation enables a model of a physical system to be optimised according to some cost function, to yield the optimum parameters for that model given some input data. Supervised artificial neural networks are similar in that training involves the minimisation of some cost function, however the ANN method does not have a specific model of the system being measured. Features of the system are empirically learnt as the ANN is trained, this adaptability makes the ANN very good at performing complex non-linear mappings.

There are various ‘off the shelf’ algorithms for performing optimisation and training ANNs. Matlab [66] by Mathworks contains a number of toolboxes by which optimisation and neural network training can be carried out. These methods are optimised for speed and memory consumption and have been used throughout the thesis to carry out neural network training and optimisation tasks.

6 ROOM ACOUSTIC PARAMETER ESTIMATION USING ARTIFICIAL NEURAL NETWORKS

This chapter discusses the development of a machine learning method, based on previous work by Li [2] and Cox *et al.*[5], for estimating room acoustic parameters from speech and music using Artificial Neural Networks. ANNs with a large number of inputs are difficult to train. Generally, ANNs are limited to less than 200 inputs [2] therefore a preprocessing stage must be employed to reduce the quantity of data inherent in reverberated speech and music signals before being fed into the ANN. Reverberation is known to smooth signal envelopes and this is similar to low pass filtering. By estimating the transfer characteristics of this low pass filtering operation, the acoustic parameters can be estimated. An envelope spectrum detector [18] is used to compress the data prior to the machine learning stage. Speech envelope spectra are known to be relatively stable features of speech signals and Houtgast [19], Li [2] and Cox *et al.*[5] have shown that reverberation time and noise levels can be estimated using them.

Developments in this chapter focus on extensions to the method developed by Li *et al.* for speech, to allow music signals to be used in estimations. Work in this chapter details the development of a multi-band envelope spectrum detector to account for the uneven spectrum of music signals. The training of such a system requires a large database of realistic room impulse responses. Previously, Li *et al.* used stochastically generated impulse responses, in this thesis geometric room prediction software is used to generate the training examples, as described in Chapter 4. As the machine learning method uses reinforcement learning, a more realistic RIR database will yield a more accurate measurement system.

6.1 Background – The Envelope Spectrum Method

Li *et al.* [2] and Cox *et al.*[5] developed a machine learning approach to estimating room acoustic parameters. They trained an ANN to recognise acoustic parameters from the envelope spectrum of received reverberant speech signals (this method will henceforth be referred to as the envelope spectrum method). The method uses a

supervised learning approach and thus is semi-blind, as a period of training is required where the ANN is trained by envelope spectra with known acoustic parameters. The total squared error is minimised by comparing estimates with actual parameters and updating the network weights accordingly. During the training phase, the ANN is taught to associate envelope spectra characteristics with acoustic parameters. The envelope spectra are applied to the ANN inputs and the output of the ANN is compared with the true acoustic parameter value. The ANN is then updated according to the error between the acoustic parameter and the ANN output. This scheme is presented in Figure 6-1.

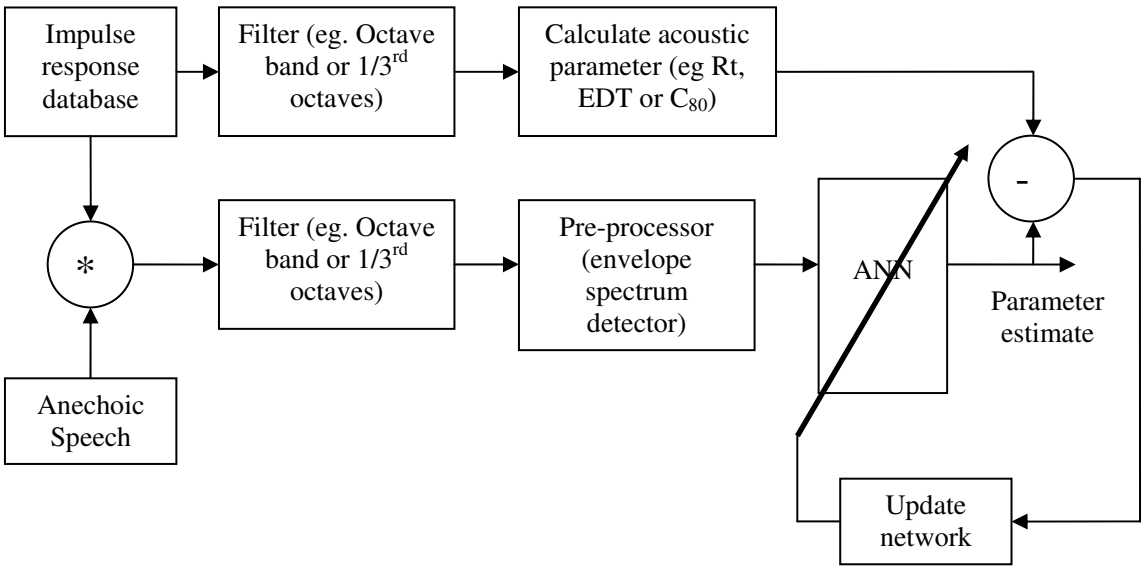


Figure 6-1. Envelope spectrum method - ANN training phase.

To determine if the network training is complete a second data set, called the validation data set, is used, this data is not used in the training phase. After each training cycle the validation data is fed though the network and the output error calculated to gauge its performance. The training data error is used to update the ANN weights while the validation data error is used to determine whether to stop training. When the validation error ceases to reduce, the training is halted. This is known as early stopping and is used to avoid over-fitting and ensure the network can generalise. Ensuring the network is capable of recognising parameters from data not used in the training phase is an important ability and is required for any acoustic parameter measurement system.

Once the training phase has been completed, the network is ready to estimate room acoustic parameters. This phase is known as the retrieval phase, named as such because the network is retrieving information stored within itself in the training stage. Figure 6-2 shows the system flow for the retrieval phase.

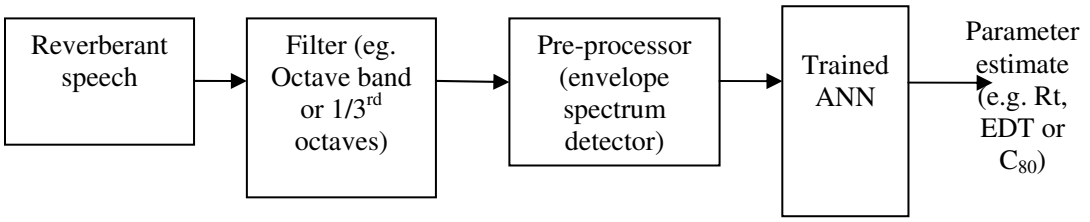


Figure 6-2. Envelope spectrum method - ANN retrieval phase

The test data set is then used to classify the ANN performance. Unless otherwise stated, all presented results are from the test data set. The following sections give more information on the separate parts of the systems outlined in Figures 6-1 and 6-2.

6.1.1 Envelope Detection

Sound pressure changes can be evaluated using the envelope of the signal. Envelope detection using a Hilbert Transform is a mathematically deterministic and robust definition of signal envelopes, which is thought to be superior to other techniques for most applications [16]. For simplicity, continuous signal expressions are used here to outline the procedure. For a time *t,* let the received sound signal be y*(t)* and the envelope *env(t)* is evaluated using:

$$env(t) = \sqrt{y^2(t) + y_h{}^2(t)} \qquad (6\text{-}1)$$

where $y_h(t)$ is the Hilbert Transform of speech signal y(t) and is defined by,

$$y_h(t) = H\left[y(t)\right] \equiv \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{y(t-t')}{t'}dt \qquad (6\text{-}2)$$

The Cauchy principal value is used to account for the singularity at *t=0.* Hilbert transforms shift each frequency by 90^o therefore can be implemented efficiently using an FFT; phase shifting by $\pi/2$ for positive frequencies and $-\pi/2$ for negative frequencies.

6.1.2 The 'Welch' envelope spectrum detector

As previously discussed, Houtgast and Steeneken [18] showed that the low frequency envelope spectrum can be used as an indicator of the level of reverberation added to speech. The original envelope spectrum was calculated using a hardware spectrum analyser which outputed a power spectrum of the envelope. Li [2] utilised modern signal processing techniques to give better estimations of the envelope spectrum with higher frequency resolution and lower variance. Li used Welch's periodogram method [67] to estimate high resolution envelope spectra. Welch's method is a standard non-parametric spectrum estimator that provides reliable power spectrum estimates.

The Welch envelope spectrum method calculates the power spectrum of a signal as follows:

1. The envelope is windowed into short windows - 50% overlap is used.

2. Each windowed section is multiplied by a Hanning window (to reduce the amount of spectral leakage).

3. An FFT is carried out on each window where zero padding is used to interpolate between frequency bins.

4. An average of all the spectra yields the final estimate.

5. The spectrum is normalised to the mean intensity of long term speech signals. This is achieved by setting the output of the analyser as 0dB when a sine wave with amplitude equal to the average sample value is applied to the spectrum analyser. The normalisation has important implications: envelope spectra are independent of excitation level, and noise and reverberation have different effects on the envelope (noise reduces the overall level and reverberation has a filtering effect).

There are several of different parameters that can be used to fine-tune the spectrum detector including: window width, FFT length and how the spectrum is sampled. All of these can have a significant impact on the ANN performance. For speech signals 1

minute is sufficient to yield a stable envelope spectrum estimation [19]. Window widths of between 2.5 and 4s gave the best results when training an ANN. Reverberation has predominantly a low frequency effect on speech envelopes and so taking more data from lower frequencies is known to aid accuracy. Therefore the spectrum is sampled from 0.15Hz to 25Hz in 1/6th octave steps. Three example envelope spectra are presented in Figure 6-3 and show the effect of reverberation on the envelope spectrum, where the low pass filtering effect of increasing the reverberation time can clearly be seen.

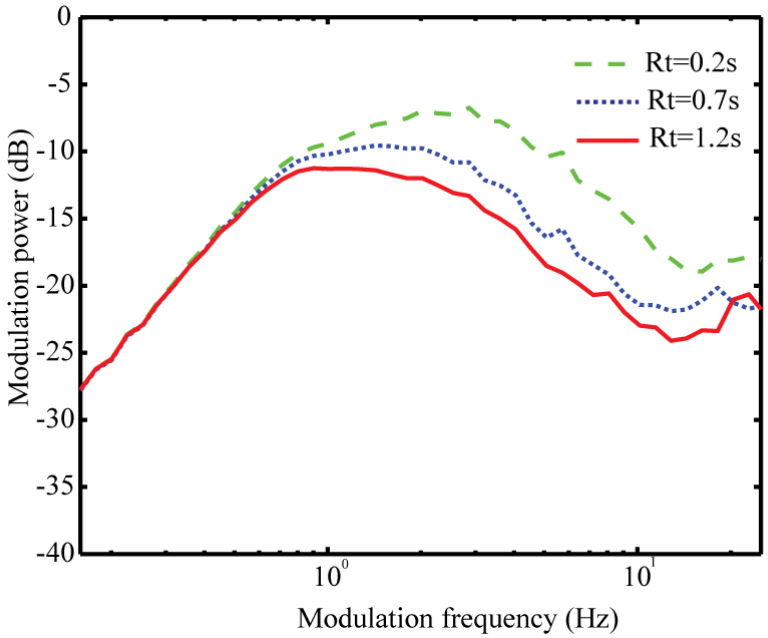


Figure 6-3. Example of speech envelope spectra received in different reverberant environments. Simple exponential impulse responses with reverberation times of 0.2s, 0.7s and 1.2s are generated and convolved with the speech signal. A 3.5s window is used in the spectrum estimation.

6.2 Data separability

Data separability refers to the ease with which the ANN can learn the features in the training set and distinguish between different data vectors, i.e. how similar or different envelope spectra with differing parameter values are. This data separability has tremendous influence on the measurement performance. The separability is related to the similarity of the data vectors over a range of parameter values and is dependent on

the level of noise present on the data vectors. Noisy data can cause the ANN training to fail or find sub-optimal solutions. The data set must fulfil a number of criteria to lead to successful ANN training:

1. The data must be representative of all likely inputs so the ANN can generalise in the retrieval phase.

2. The data vectors must have a high degree of data-separability for good ANN performance.

Training an ANN is computationally expensive, particularly if the network is large, and therefore it useful to gain knowledge of the quality of the data so that the pre-processor can be tuned prior to training.

6.2.1 Mahalanobis distance

The Mahalanobis distance [68] has been used as a measure of data separability. The distance between two data vectors can be quantified as the Euclidian distance between the two vectors. The Mahalanobis distance (D_m) is a similar measure which describes the distance between the mean vectors μ_1 and μ_2 of two groups of data, but also takes into the account the covariance matrix $C_{1,2}$ between the two groups

$$D_m = \sqrt{(\bar{\mu}_1 - \bar{\mu}_2)^T C_{1,2}{}^{-1} (\bar{\mu}_1 - \bar{\mu}_2)} \qquad (6\text{-}3)$$

where T denotes vector transpose. The ANN is performing a function approximation task as the parameters being estimated are assumed to be generated via some continuous function. Therefore the Mahalanobis distance is not ideal, as it is designed for measuring the distance between discrete groups of data. The Mahalanobis distance is ideal for predicting the classification performance of an ANN. However, as the parameter estimation methods do not require infinitesimal accuracy, the data can grouped. For example, with Rt the data can be grouped into sequential groups each with a width of 0.1s. The Mahalanobis distance between adjacent data vector groups can then be calculated. The data vectors (envelope spectra) are split into a series of adjacent groups, each a specified width in seconds (for EDT or Rt). D_μ in equation (6-4)

is then defined as the average Mahalanobis distance between a series of adjacent (in decay time) bins of the envelope spectra.

$$D_{\mu} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{(\vec{\mu}_n - \vec{\mu}_{n+1})^T C_{n,n+1}^{-1}(\vec{\mu}_{n1} - \vec{\mu}_{n+1})} \qquad (6\text{-}4)$$

where μ_n and μ_{n+1} are the mean vectors of the two adjacent groups and $C_{n,n+1}^{-1}$ is the inverse covariance matrix and N is the number of groups. D_{μ} shows excellent correlation with the ANN error performance and is quick to compute a repeatable measure of overall data separability.

6.3 Training an artificial neural network to estimate room acoustic parameters from reverberated speech

A number of trials were run using Li's method but utilising the new RIRs rather than the stochastically generated RIRs which were previously used. This was in order to examine how well the measurement system performs with this new impulse response database. An anechoic speech signal of length 90s was used [69]. All experiments were carried out using the 1 kHz octave band-pass filtered signal.

6.3.1 Training the neural network

To train the ANN, the data are first split into three groups: the training set, the validation set and a test set. The training set is made up of 7560 different room examples, and both the validation and test sets are made up of 420 different room examples. The impulse responses for each set are randomly chosen for each trial. The neural network training is carried out using the Matlab Neural Network Toolbox [66]. The training algorithm chosen is based on the second order gradient-decent algorithm; Levenberg-Marquardt [70] back propagation. This algorithm is quoted in the Matlab help file as being the fastest training method for moderate-sized neural networks and has a very efficient Matlab implementation.

The training is halted if a training cycle causes the squared validation error to increase or not improve. An ANN with the structure 45-20-10-5-1 is found emperically to work

well. It is known that ANNs only generally require two hidden layers for most problems [63] the third hidden layer is probably redundant and with hindsight a two hidden layer network may have been sufficient, but investigations indicated no overfitting problems occurred using a 3rd hidden layer. All activation functions are hyperbolic tangent sigmoids except for the output layer which is linear. Once the training stage is completed, the ANN performance is evaluated using the test data set. The error is presented as the percentage of parameter estimates that are within the difference limen.

6.3.2 Results using speech signals

The window width for the envelope spectrum detector is set at 3.5s, with 50% overlap, and the envelope spectrum is sampled from 0.15Hz to 25Hz in 1/6th octave steps to emphasise the low modulation frequency features (the windows are zero padded to increase the number of frequency bins in the spectrum). The ANN is trained using the Rt. Figure 6-4 shows the progress of the training, highlighting the training and validation errors up to the stopping point. Each epoch represents 7560 iterations through all training examples. Figure 6-4 shows how the total squared error in the validation set begins to diverge from the training error indicating no further training would be beneficial and training is therefore stopped when two consecutive epochs indicate an increase in the squared validation error.

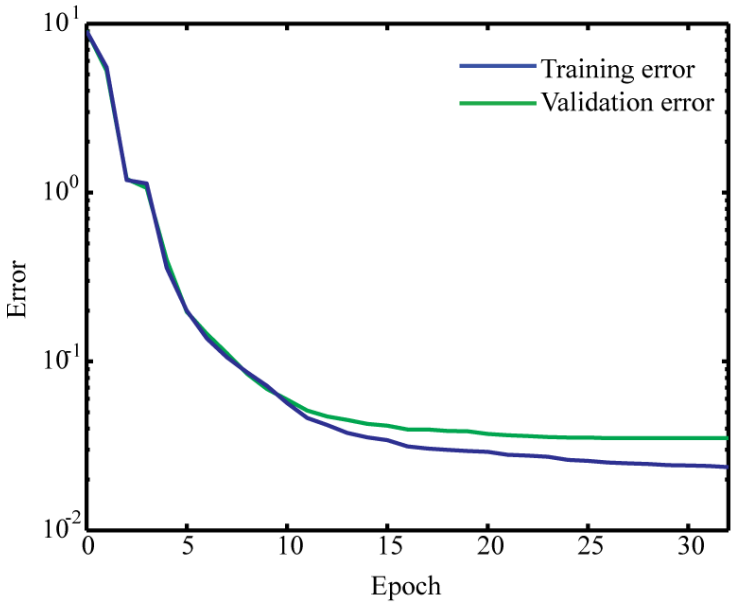


Figure 6-4. ANN training phase error showing the error in the training and validation sets. Each epoch represents a step through all of the data set.

After training, the ANN is presented with the test data set and the estimation performance is analysed. The resulting ANN estimates are then compared with the values calculated directly from the RIRs, as shown in Figure 6-5. This shows that 62% of the Rt estimates are within the difference limen using speech signals. This demonstrates a decreased performance when compared to previous researchers results using speech signals [2]. In order to investigate this, the system was checked by using the same stochastic impulse response model used in this earlier research. It was found that using this stochastic impulse response database, much higher accuracy for Rt was yielded, with 94% of results within the difference limen. This indicates that the method is working as originally intended and the difference in Rt accuracy between the stochastic and geometrically simulated RIR databases is due to additional complexities within the new RIR database. The most noticeable difference between the two databases is that the new geometrically generated database contains a significant number of non-uniform decays where the early decay rate differs significantly from the late decay rate. These non-uniform decays pose a much more complex mapping tasks for the ANN, for reasons that will be explained in the following paragraphs and in Section 6.4.

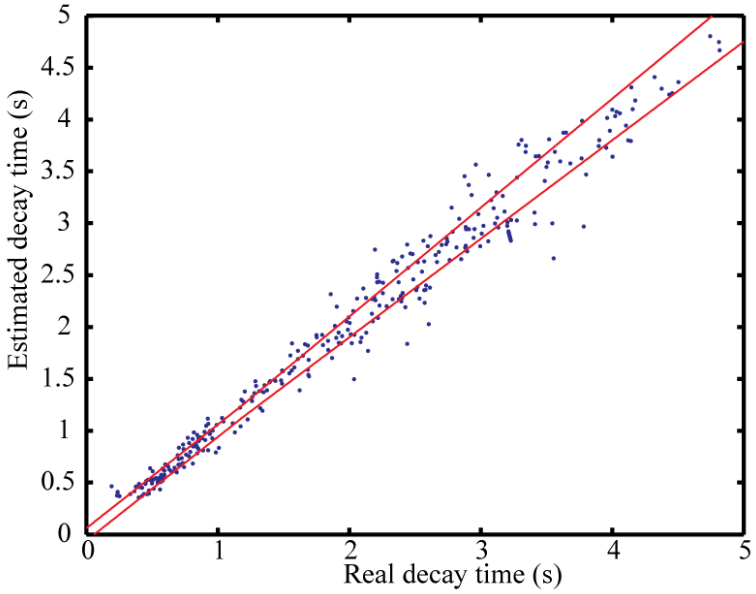


Figure 6-5. ANN estimation results on test data set for Rt using speech as test signal. Red lines indicate the subjective difference limens.

In addition to investigating Rt, different ANNs were trained, using the same data sets to estimate EDT, C_{80} and t_s. Figure 6-6 shows the estimation performance for EDT. It is apparent that the ANN method is more successful at estimating the EDT than the Rt, with 92% of the EDT estimates within the difference limen. This difference is due to the reverberant tails following speech utterances being masked by subsequent utterances. This ultimately means the received reverberation is more heavily influenced by the early rather than late reflections and therefore the effect of the early reflections on the envelope spectrum is emphasised.

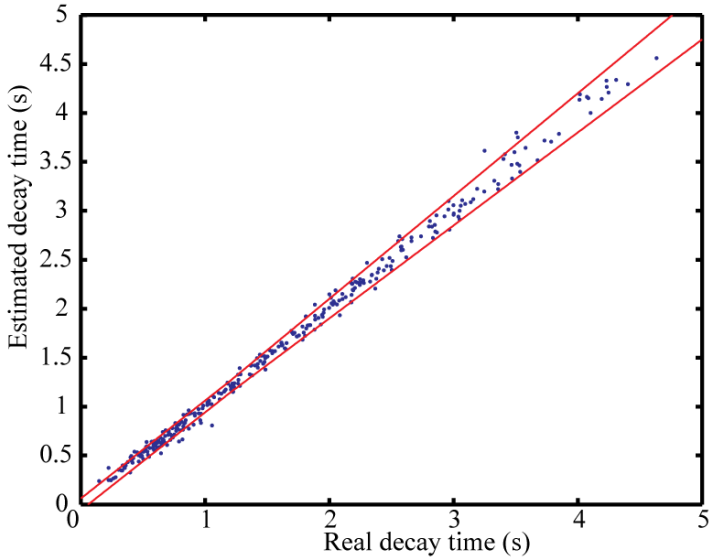


Figure 6-6. ANN estimation of EDT, from speech signal, results from test set.

The ANN learns, not from features in the time domain signal envelope, but from an averaged envelope spectrum. Section 6.4 investigates how this emphasis of the early reflections influences the envelope spectrum.

6.3.3 Influence of window width on estimation performance

The window width used in the envelope spectrum estimation influences the frequency resolution and variance of the resulting envelope spectrum estimate. A wide window provides a better spectral resolution, but short windows increase the number of averages used to compute the final envelope spectrum and therefore decrease the variance in the estimate. It is therefore empirically investigated if improvements in accuracy can be found by altering the window lengths to 2s, 3.5s 5s or 10s. The Mahalanobis distance is calculated as described in Section 6.2.

Figure 6-7 shows how the data separability varies with Rt and EDT. This shows that in general the longer windows appear to give greater data-seperability for both parameters. This is expected as the effect of reverberation on signal envelopes, especially for long decays, is a slowly varying process and therefore greatly influences envelope features at low modulation frequencies. The shorter windows results in lower resolution spectrum

estimations and this reduction in resolution, particularly at low modulation frequencies, causes the loss of features important in estimating the decay rate. For Rt and EDT times a window width of 5s appears to be optimal.

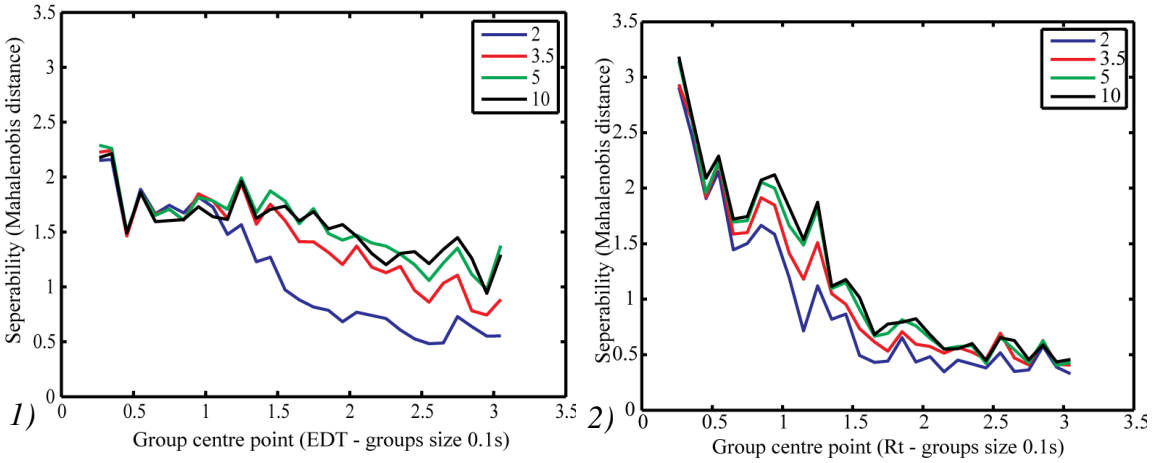


Figure 6-7. Mahalanobis distance as a function of Rt(1) and EDT(2). These plots indicate the separability of the data and how it changes with respect to decay rate. The legend indicates the window width (in seconds) used to compute the envelope spectrum.

An interesting feature of

Figure 6-7 (1) is that, for Rt, the separability decreases steadily and then flattens off at around 2s. This is because speech envelope spectra generally have their strongest modulation levels between 0.8Hz and 5Hz (see Figure 6-3). A modulation frequency of 0.8Hz corresponds to features about 1.25s in length. When taking into account that the Rt is calculated to -35dB (not -60dB), an RIR with an Rt of 2s calculates the parameter value from approximately the first 1.25s of decay. This is also apparent in Figure 6-5 which shows that the Rt estimation performance begins to deteriorate when Rt>2s. This explains the increase error in Rt estimation above 2s as more of the late part of the impulse response is masked by new utterances.

Figure 6-7 (2) shows that the performance is more consistent for EDT over the range of the whole possible parameter values. This is because the EDT is calculated from the first 10dB of decay therefore values can be calculated using a much smaller length. For

example, only the first 0.83s seconds of an RIR is used to compute an EDT of 5s. As demonstrated in Figure 6-3, speech contains many features with lengths between 0.2 and 1.25s (0.8 Hz and 5 Hz), sufficient to allow free decay for a wide range of EDT values.

Figure 6-8 and Figure 6-9 show the ANN estimation performance for C_{80} and t_s. Once again the ANN estimation is very good for these parameters with 97% lying within the difference limen for C_{80} and 98% within the difference limen for t_s. Again the reason for the excellent performance is that these parameters are most influenced by the early sound field, and the early sound field has the greatest effect on the speech envelope spectrum.

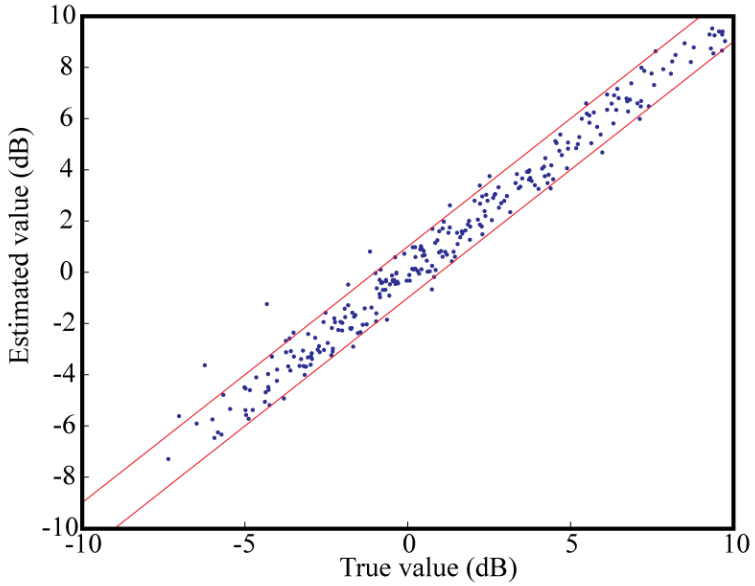


Figure 6-8. ANN estimation of Clarity from speech signal, results from test set.

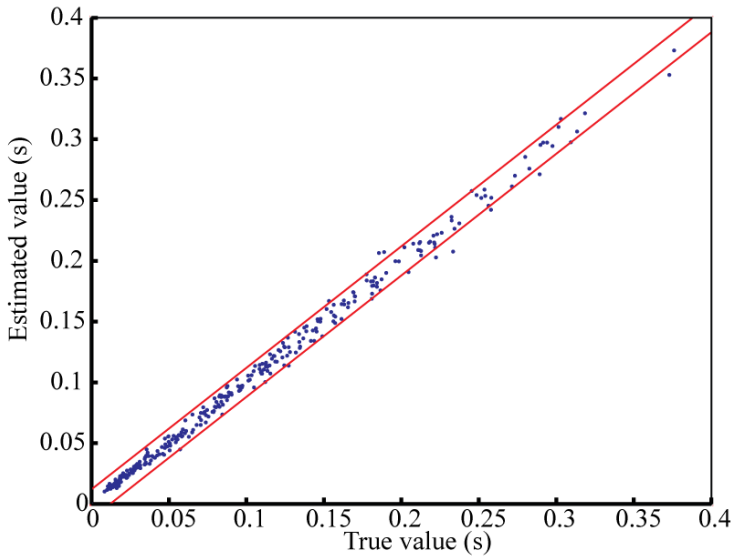


Figure 6-9. ANN estimation of Centre time from speech signals results from test set.

6.4 Effect of reverberation on the envelope spectrum

As mentioned previously, reverberation has the effect of smoothing features of signal envelopes and is similar to a low pass filter operation on the envelope. It is important to investigate this relationship so that the limitations of the method can be better understood. The MTF approximately describes the transfer function affecting the anechoic envelope spectrum in a reverberant space. This section investigates how different features of various RIRs can affect the MTF and thus also how particular RIRs affect the envelope spectrum. Of particular interest are the differing effects of the early and late portions of the RIR on the MTF and envelope spectra.

An idealised room impulse response $h(t)$ can be represented by an exponential with time constant τ modulating a Gaussian noise source $n(t)$;

$$h(t) = e^{-\frac{t}{\tau}} n(t) \qquad (6\text{-}5)$$

Calculating the Complex Modulation Transfer Function (CMTF) from equation (2-7) using (6-5) yields the following relationship as a function of ω.

$$E[CMTF(\omega)] = \frac{\int_0^\infty e^{-\frac{2t}{\tau}} E[n^2(t)] e^{-j\omega t} dt}{\int_0^\infty e^{-\frac{2t}{\tau}} E[n^2(t)] dt} = \frac{\int_0^\infty e^{-\frac{2t}{\tau}} e^{-j\omega t} dt}{\int_0^\infty e^{-\frac{2t}{\tau}} dt} = \frac{1}{\left(\frac{2}{\tau} + j\omega\right)} \frac{2}{\tau} \qquad (6\text{-}6)$$

$$E[CMTF(\omega)] = \frac{1}{1 + j\omega\tau/2} \qquad (6\text{-}7)$$

Interestingly, comparing (6-7) with the electrical impedance of a capacitor shows that they are quite similar. A capacitor in an electric circuit behaves like an acoustic space where energy enters the space/capacitor and is gradually released. A time varying signal applied to these systems smoothes (low-pass filters) the signal. In this idealised case, the Rt parameter is equivalent to capacitance.

As described by Schroeder [22], in this simple case the reverberation time can be calculated directly from the 3dB cut-off point, ω_c, of the filter's transfer function. However, in reality acoustic impulse responses differ considerably from the idealised case, especially in the early part of the response. The early sound field includes a time gap between the direct sound and the first reflections. Furthermore, the strong isolated reflections can cause ripples in the MTF where the periodicity of the ripples is related to the time between reflections. The early sound field is highly variable and depends on many factors. As a result, the MTF is often much more complex than a simple low-pass filter transfer function. In order to extract parameters from the envelope spectrum, estimates over a range of frequencies modulation are required in order to account for these ripples. In the case of non-uniform decays, which are often seen in real rooms, the modulation transform is even more complex. For example, when the impulse response is modelled by a sum of exponentials modulating a noise source as in equation (6-8);

$$h(t) = \left(e^{-\frac{t}{\tau_1}} + e^{-\frac{t}{\tau_2}} \right) w(t) \qquad (6\text{-}8)$$

By calculating the CMTF in a similar manner to (6-6), the Fourier transform of $h^2(t)$ is;

$$E[CMTF(\omega)] = \left(\left[\frac{\tau_1}{2 + j\omega\tau_1} \right]_1 + \left[\frac{\tau_2}{2 + j\omega\tau_2} \right]_2 + \left[\frac{2\tau_1\tau_2}{\tau_1 + \tau_2 + j\omega\tau_1\tau_2} \right]_3 \right) / \, Norm \qquad (6\text{-}9)$$

Norm is a normalisation factor arising from the denominator in equation (2-7) which calculates the CMTF from the RIR. Equation (6-9) shows that the CMTF of a non-uniform decay can be thought of as the superposition of a number of 'partial CMTFs'. These partial CMTFs are labelled as 1, 2 and 3 in equation (6-9). Partial CMTFs 1 and 2 are related to the modulation transfer functions of each region of decay calculated separately, while partial CMTF 3 is a cross term due to the squaring operation. Figure 6-10 shows the partial MTFs evaluated for the decay seen in equation (6-8) where $\tau_1=8$ and $\tau_2=0.5$. In Figure 6-10 we can see a complex relationship is formed, different regions of the decay curve affect modulation frequencies by differing amounts. Figure 6-10 shows that the slowly decaying late region (due to τ_1) is the predominant influence on the lowest modulation frequencies while the quickly decaying early region is more predominant at higher modulation frequencies. This result has a major impact on the accuracy of any estimation using the MTF of envelope spectrum. To gain adequate estimates of the decay for a complex, multi-rate decay function, accurate estimates of the modulation transferred over a range of modulation frequencies are required.

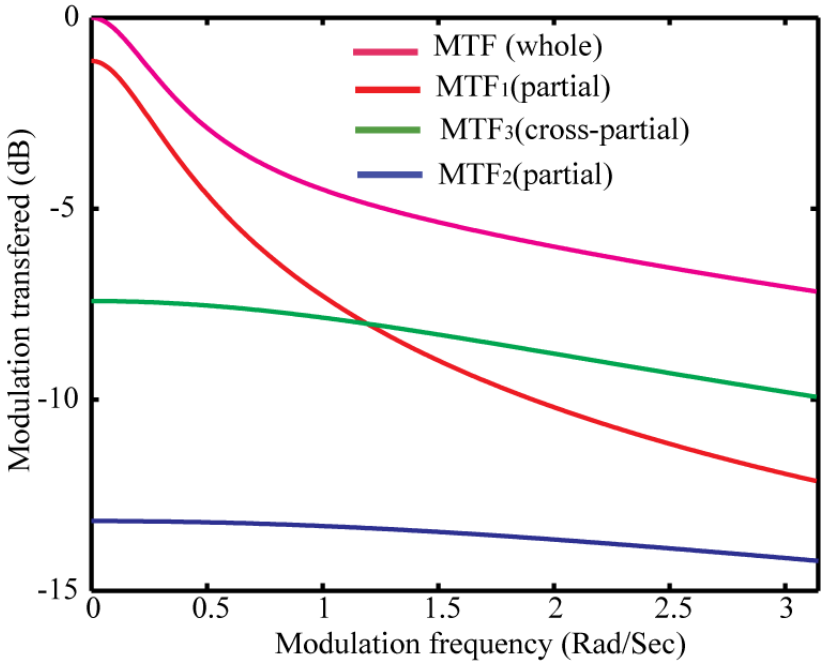


Figure 6-10. MTF evaluated for a model of non-uniform sound decay using a sum of exponentials. The partial MTFs are shown for each exponential as well as the cross term and the whole MTF

6.4.1 Effect of reverberation on the envelope spectrum - summary

The early and late portions of the RIR influence the envelope spectrum in different ways. While both early and late parts of the impulse response have an effect at all modulation frequencies, low modulation frequencies are most heavily influenced by the late response and the early response most heavily affects the higher modulation frequencies. This means that, particularly when the RIR demonstrates non-uniform decay, the envelope spectrum needs to contain modulations over a wide range of frequencies in order for the ANN to correctly differentiate between early and late components.

The temporal distribution of reflections in the early sound-field has a comb-filtering effect on the envelope spectra. The comb-filter's pattern of notches is dependent on the time delays between the early reflections. This phenomenon complicates the feature space and can cause problems in ANN training when a limited range of modulation frequencies are available.

The task of estimating the Rt is made more difficult as the calculation for Rt_{30} ignores the first 5dB of decay. This means there are components in the envelope spectra, due to the direct sound and early reflections, which are essentially noise. This can degrade the ANN performance when estimating the Rt. In the ANN method, this can cause very different envelope spectra to have the same Rt, and these conflicts can cause major problems when training an ANN.

In summary, to gain good estimation acoustic parameters requires the source signal envelope to have sufficient modulation over a range of frequencies.

6.5 Training the network to estimate the MTF directly

Training an ANN to estimate acoustic parameters such as Rt and EDT has been shown to yield reasonable results for EDT but yields errors in the late reverberation estimation. This is because speech does not provide modulation over a sufficiently wide range of frequencies and additional lower modulation frequencies are required (or alternatively, there are insufficient gaps in-between utterances). The other problem is that the parameters are calculated from a decay curve excited by broadband excitation. Speech only approximates a broadband signal and therefore the ANN needs to compensate for this.

These two issues motivated an investigation into an alternative approach. An ANN is trained using the envelope spectrum of speech but using the MTFs as targets instead of acoustic parameters. The MTF is a measure of the level of modulation transferred by a system for a range of modulation frequencies modulating a stationary broadband noise source. Therefore it is hoped that, by training an ANN to directly map between envelope spectrum and MTF, the ANN is being trained specifically to find a way to compensate for the non-broadband nature of the signal and the lack of some modulation frequencies.

A useful outcome of this methodology is that it yields a MTF estimates from which the decay curve can be directly calculated (via an inverse Fourier transform). From the MTF all the acoustic parameters could be estimated and, unlike previously, only a

single ANN needs to be trained to estimate all parameters. The ANN is essentially performing blind deconvolution of the MTF from the envelope spectrum.

The MTF only approximately describes the modulation transfer of speech and music signals (due to the deviation from broadband excitation) and the success of this methodology depends on the ANN finding ways to compensate for this approximate relationship. It was discovered that while the method was able to yield decay curve estimates, the resulting parameter accuracy was less than that achieved by training the ANN on the parameters directly. Increasing the number of dimensions and problem complexity causes a decrease in performance; therefore this methodology was not further developed. For completeness the investigation is detailed in Appendix E.

6.6 Companding to improve the Rt estimation

Companding is a signal processing technique that is often used to mitigate the detrimental effects of channels with limited dynamic range. Companding reduces the dynamic range of signals by applying a non linear processing technique to the time domain signal. There are two main types of companding; A-Law and μ-law [71]. Both are quite similar although μ-law is generally used in American telecommunications and the A-Law in European telecommunications. The nonlinear μ-law transfer function is as follows:

$$F(x) = sign(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)}$$

(6-10)

Figure 6-11 shows how companding compresses the upper dynamic range of the signal into a smaller dynamic range than the lower level signal components. This means that after companding, changes in lower-level signal components will be emphasised when compared with the higher-level components. Therefore, when companding reverberant signals, the effects of the late reflections are emphasised and the effects of the early reflections de-emphasised.

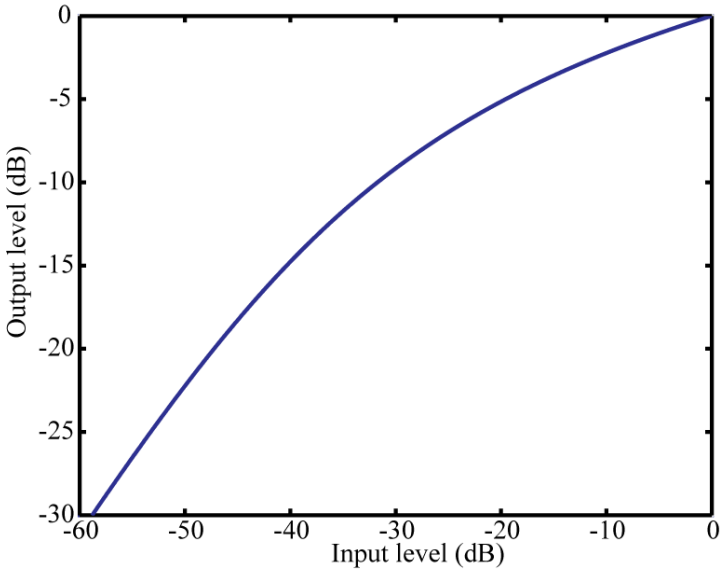


Figure 6-11. μ-law companding transfer function.

The companding stage is placed immediately prior to the envelope spectrum detector. Figure 6-12 shows the effect on Rt estimation accuracy when companding is used prior to envelope spectrum detection, a value of $\mu=150$ is found to improve the ANN estimation accuracy. The accuracy has been significantly improved from 61% within the DL without companding (Section 6.3.2), to 94% within the DL. The improvement is due to the reduction in importance of the early reflections on the envelope and the increase in importance of the later reflections. This is similar to the effect of increasing the signal-to-noise ratio, where the early reflections are the noise and the later reflections the signal. When estimating the Rt the first 5dB of decay is effectively noise as it is not included in the calculation. Companding a noisy signal however will simply decrease the signal-to-noise ratio and reduce the accuracy still further.

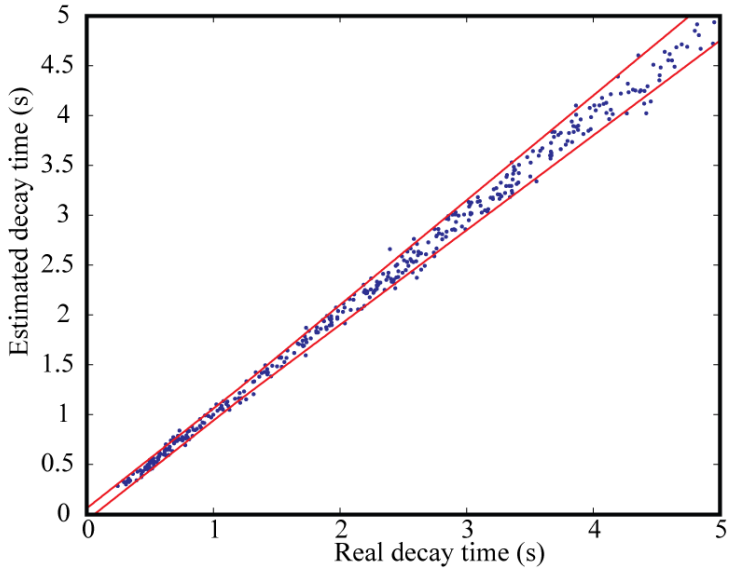


Figure 6-12. μ-law companding; effect on Rt accuracy.

6.7 Estimating room acoustic parameters from reverberated music

A specific aim of this work was to extend Li's methodology so that room acoustic parameter estimations could be made using music as well as speech. Applying the method developed by Li *et al.* to music signals yielded the results shown in Table 6-2. The music used was from the Denon anechoic orchestral music CD [72], the pieces of music used are described in Table 6-1.

Title	Duration (mm:ss)
1. Mozart: Le Nozze De Figaro: Overture	4:19
2. Mendelssohn: 4th Mov. -SymphonyNo.3 In A Minor, Op.56 'Scottish', Bars 396-490	2:20
3. Bizet: L'Arlesienne, Suite No.2: Menuet, Bars 396-490	4:13
4. J. Strauss: Pizzicate-Polka	2:35
5. Pushkin: Ruslan And Lyudmila	5:22
6. Verdi: La Traviata	3:27
7. Bruckner: SymphonyNo.4 In EFlat 'Romantic',Bars 517-573	1:41
8. Debussy: Prelude AL'Apres-Midi D'un Faune, Bars 1-20	1:55

Table 6-1. Denon anechoic music listing [72].

Table 6-2 shows that music generally produces very poor results for most parameters. In particular, Rt yields a very poor result where on average only 21% of the results are within the DL. The results are less bleak for the parameters that describe the early sound field, EDT, C_{80} and t_s. On average, for EDT, 37% of the estimates are within the subjective DL, for C_{80} 52% and t_s 55%. Additionally, it can be seen that one particular piece of music, track 4, performs much better for all parameters.

	ANN performance (% within DL)			
Track number	Rt	EDT	C_{80}	t_s
1	23	40	58	58
2	20	36	53	54
3	22	38	53	55
4	35	55	74	78
5	23	42	62	61
6	21	34	48	48
7	21	31	45	46
8	15	21	33	27

Table 6-2. ANN estimation accuracy for music signals using un-modified envelope spectrum method

6.8 Problems with music signals

In order to investigate the reduction in accuracy when using music signals, a number of empirical investigations were carried out. First the effect of the differences between envelope spectra of the music signals were investigated, followed by the differences between the audio spectra of the music signals.

6.8.1 Envelope spectra of music signals

Music and speech signals have very different low frequency envelope spectra. This is highlighted in Figure 6-13 which compares the envelope spectra for all of the pieces of music tested along with the speech envelope spectrum. The main difference between the speech and music envelope spectra is that the speech envelope spectrum is much smoother.

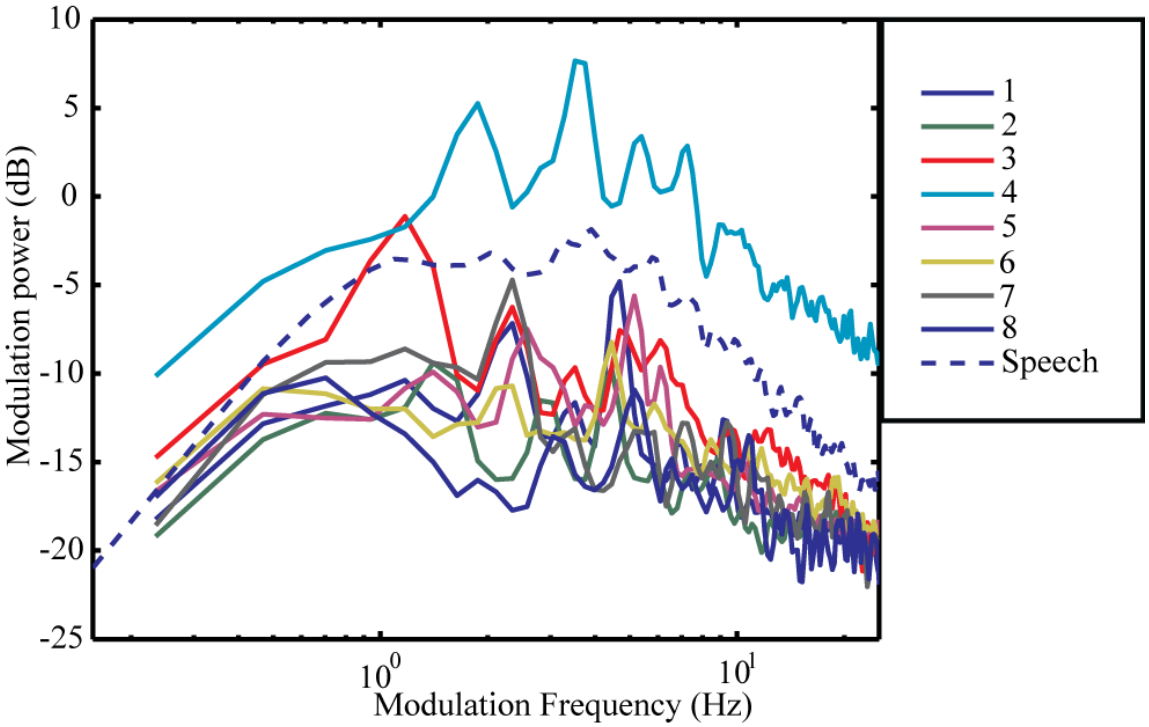


Figure 6-13. Envelope spectra of music signals compared with speech envelope spectra, figure legend indicates the piece of music

Speech diction varies more stochastically than music and unlike music, is not dictated by a musical score, or synchronised to a specific tempo. Therefore, the speech envelope spectrum is quite random and produces modulations over a range of low frequency modulations. This results in quite a smooth envelope spectrum over a range of modulation frequencies. In contrast, music's low frequency envelope spectra are much more uneven. A piece of music is played at a certain tempo which influences the speed and length of the individual notes. For example, a quaver in a piece of music with 120bpm in 4/4 time will be 0.25s long. If a sequence of quavers is present in a piece of music this will produce a component in the envelope spectrum at 4Hz. Due to the inherent patterns within music signals, the envelope spectrum will contain components at specific modulation frequencies related to the particular patterns of notes within the music. This means that the music envelope spectrum is characterised by a number of peaks due to the musical patterns and tempo, but is lacking modulation in-between these frequencies. This is problematic for the envelope spectrum method, as good estimates for a range of modulation frequencies are required.

Interestingly, in Figure 6-13 it is noted that one piece of music (track 4) has a higher envelope spectrum than the rest of the signals including speech. The reason for this is the particular style of music which is pizzicato strings, with many musical rests between notes. This means that the piece's envelope has a low average level, which when combined with the sharp pizzicato notes causes the envelope spectrum, when normalised, to demonstrate high levels of modulation.

To further investigate this problem a series of artificial signals were generated to show how the modulation frequency content of the low frequency envelope spectrum affects the acoustic parameter estimation. The following modulation frequencies were chosen; 0.5, 1, 2, 5, 10, 15 and 20Hz. From this, a set of signals were generated based on sums of a number of sine waves at those frequencies, the phase of each sine wave being randomly selected. A constant is added to each sine wave so that all vary between 0 and 2. Each modulator is then multiplied by random Gaussian noise to produce the test signal. The following modulators were used:

Modulator number	Modulator is the superposition of sine waves at the following frequencies (Hz)
1	0.5
2	0.5, 1
3	0.5, 1, 2
4	0.5, 1, 2, 5

Table 6-3. Test signals generated for modulators, each modulator is a sum of sine signals, the frequencies of which are listed.

Each of the test signals is convolved with the impulse response database and the reverberated signal analysed in terms of data separability using the Mahalanobis distance. Figure 6-14 shows the data-separability across each of the modulator series datasets for both EDT and Rt. The left figure shows that for Rt, increasing the number of modulation frequency points available in the envelope spectrum increases the data-separability. This is particularly pronounced where Rt<1s. A similar, more obvious, pattern is seen when analysing data separability for EDT, this shows a distinct correlation between the number of frequencies in the modulator and the data separability below about 1.2s.

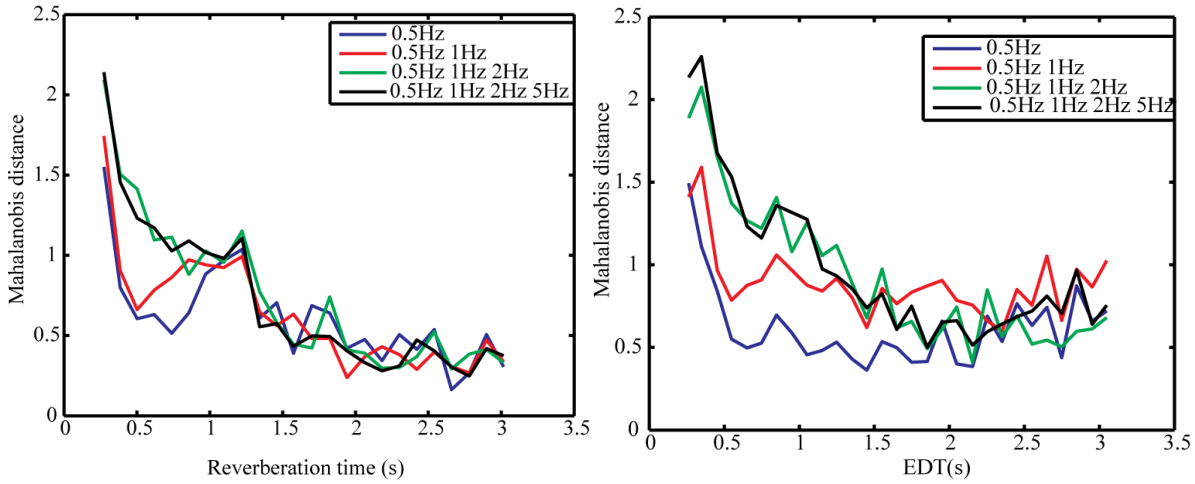


Figure 6-14. Data separability as described by the Mahalnobis distance for a sum of sine waves modulating white noise convolved with the RIR database for 1) Rt and 2) EDT. The Mahalnobis distance group step size was 0.1s.

There is a direct correlation between the number of modulation frequencies present in the envelope of a signal and the performance of the envelope spectrum method when estimating Rt and EDT. This is problematic, particularly for music signals, as the modulation frequency content of music envelopes is focused at specific modulation frequencies. This is a limiting factor in achieving good accuracy when using music signals in the envelope spectrum method.

The envelope spectra of the different music signals (Figure 6-13) are very different from one another. One of the interesting aspects of Li's work was to achieve source independence, in other words accurate estimation of the Rt using an arbitrary speech signal. This could be achieved as speech envelope spectra do not vary greatly between speakers. By training an ANN on envelope spectra of a number of speakers, parameters can be estimated with only slightly compromised accuracy. Figure 6-13 is evidence that achieving source independence with music signals using the same method cannot be achieved as the music envelope spectra vary too greatly between different pieces.

6.8.2 Audio spectra of music signals

Excitation within an octave band is uneven with respect to frequency for music signals. This has a negative effect on the accuracy of the method as acoustic parameters are calculated using a broadband signal. Figure 6-15 compares the audio spectra for two

pieces of music within the 1 kHz octave band. Also shown is the audio spectrum of the speech sample. Although the spectra of the music signals are very uneven, there is a clear pattern, in that most of the energy is focused within narrow regions relating to notes on the equal tempered scale. The amount of energy within each region is related to how many times and how loudly particular notes are sounded. It should be noted that track 4, the best performer in the ANN parameter estimations, does not vary as much over the octave band and has a spectrum closer to speech than the other signals. This is an indication that the ANN performance is closely related to the unevenness (in frequency) of the excitation signal.

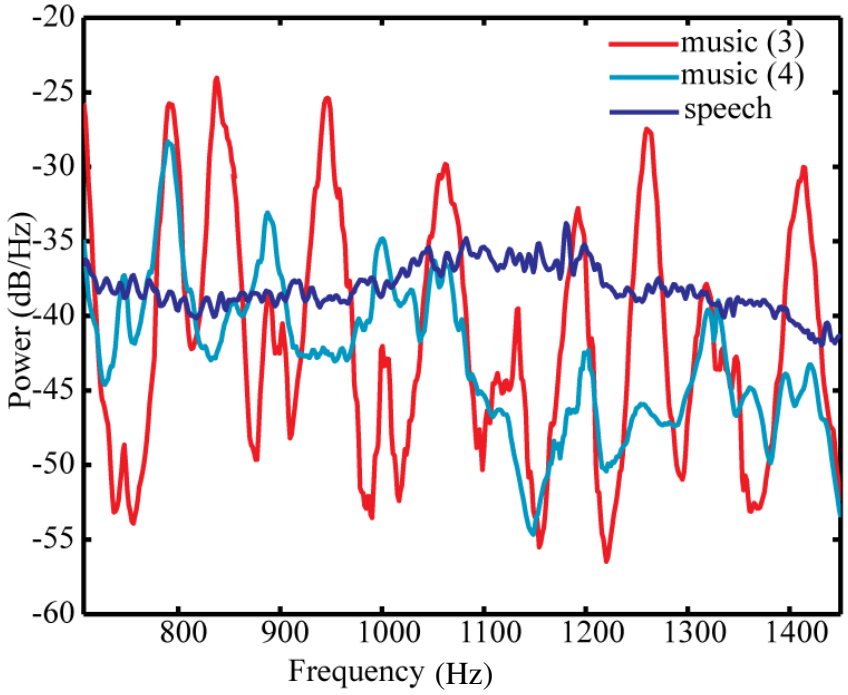


Figure 6-15. Audio spectra for a selection of the music signals (3&4) and speech in the 1kHz octave band.

Across the octave band, the speech spectrum varies by only roughly 5dB while the music spectra can vary by up to 30dB. This unevenness with respect to frequency in the excitation signal must be compensated for, as the parameters that are being estimated are defined assuming a uniform level of excitation. Consider comparing an acoustic parameter calculated from a standard impulse response with even excitation, and the same parameter calculated using an impulse response weighted (filtered) to have a frequency response corresponding to the average power spectrum of a piece of music – i.e. an uneven frequency response.

To produce the filtered RIR, the spectrum of the music extract was estimated using the Welch power spectrum method using 0.5s windows and 50% overlap Hanning windows. This frequency response is used to design a zero-phase finite impulse response filter (FIR) filter with the same frequency response as the average spectrum of the music signal. Zero-phase filters are used so the time response of the impulse response is least distorted as they minimise the start up transient response of the filter. A short tap length of 301 is used for the FIR filter. To achieve zero-phase filtering the matlab command 'filtfilt' is used which first filters the signal in the forwards direction then in the reverse direction (as the RIR was filtered twice, the filter was designed to have a magnitude response equal to the square-root of the music spectrum). The 'reverberation time' of the weighting filter was checked to ensure it would not significantly influence the room acoustic parameters because of its time response. Generally, the filter had a reverberation time of about 0.1s.

Figure 6-16 shows the reverberation time for an uneven excitation (excitation with similar spectrum to a music signal) plotted against the reverberation time with even excitation (broadband excitation) for the impulse responses simulated using the geometric room acoustic model. The result shown is for the 1kHz octave band. The error introduced by uneven excitation is the same order of magnitude as the difference limens. There are several of outliers in Figure 6-16, these are due to discontinuities in the response at the edges of the octave bands. When applying the music shaped spectral filter, small changes in the frequency response at the edge of the bands can cause large changes in the time response.

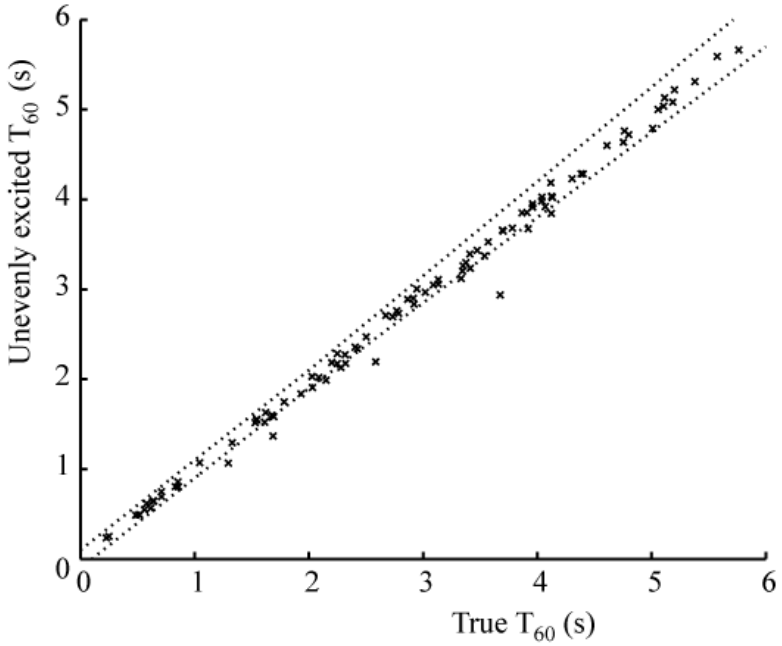


Figure 6-16. Unevenly excited reverberation time compared to the true, evenly excited value. The unevenly excited reverberation time has the same octave band frequency response as a typical piece of music. The dashed lines indicate limits derived from the perceptual difference limens after Kendrick et al. [73].

Figure 6-17 shows the EDT for an uneven excitation plotted against the EDT with even excitation. This shows that the uneven excitation has a more significant effect on EDT than Rt in the order of about two difference limens. This is because the early response is made up of a small number of discrete reflections, each may have differing frequency responses and depending on the frequency response of each reflection, the music filter will affect some reflections more than others; this can cause a significant change in the time response (and thus the EDT). The effect is less pronounced with the late response as the reflections are now well diffused and the effect of the music filter is more uniform along the late part of the RIR.

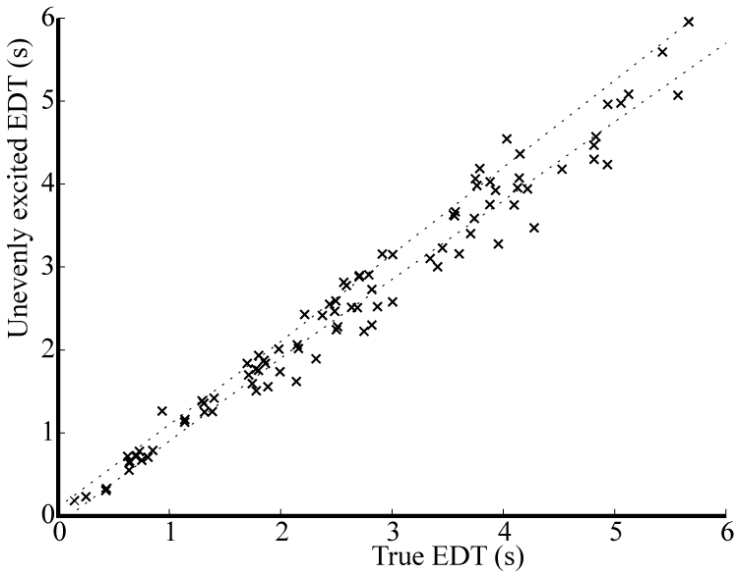


Figure 6-17. Unevenly excited EDT compared to the true, evenly excited value.

Figure 6-17 and Figure 6-16 show that it is necessary to carry out some pre-processing on the music signal to try and reduce the error introduced by the frequency response of the music.

6.8.3 The 12th octave filterbank envelope spectrum pre-processor

In order to compensate for the uneven spectra of music signals a new pre-processor was developed. This process is depicted in Figure 6-18. The first stage of the pre-processor breaks the signal down into one-twelfth octaves across the octave band being considered, where each one-twelfth octave corresponds to a note within an even-tempered, chromatic scale. The one-twelfth octave filtering used in the pre-processor helps to compensate for the uneven excitation. After the one-twelfth octave filters, the envelope of each of the twelve signals is detected using the Hilbert Transform [16] and a normalisation by the root mean square value in each one-twelfth octave carried out. This normalisation reduces the effect of uneven excitation with respect to frequency as demonstrated in Figure 6-16. The objective is to get a result closer to the one which would be obtained from an artificial test signal with even excitation with respect to frequency.

Previous work [2] with speech did not require this type of compensation as speech has a flatter frequency response across each octave band.

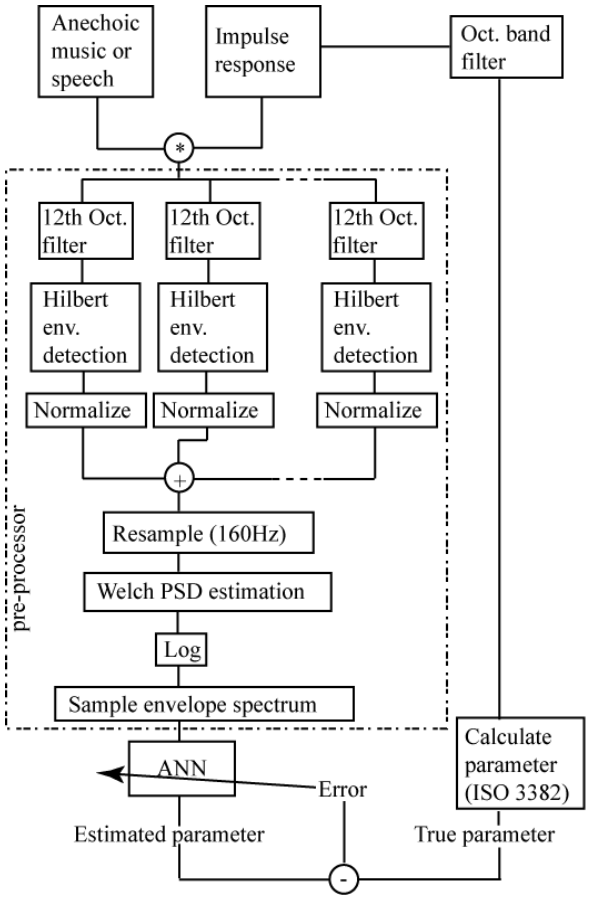


Figure 6-18. Schematic of 12th octave envelope spectrum method for estimating room acoustic parameters from music, after Kendrick et al. [73].

Instead of the filter bank method, a whitening filter, which would provide a flatter spectrum was considered, but as the signal-to-noise ratio in-between the 12th octave bands could be very low it was thought that this method would yield errors in the envelope spectrum, therefore the approach in Figure 6-18 was adopted.

6.8.4 ANN estimation results using music signals and the 12th octave envelope spectrum pre-processor

Applying the modified 12th octave envelope spectrum pre-processor to a series of anechoically recorded music recordings convolved with the room impulse response database, and training the ANN as previously describes yields the performance results listed in Table 6-4. This shows significant improvements in accuracy for every

parameter when compared with the octave band envelope spectrum results in Table 6-2. However, Rt accuracy still remains below that required for a measurement system. The performance for EDT, C_{80} and t_s are, however, more satisfactory, although these results indicate performance can vary significantly between different pieces of music.

	ANN performance (% within DL)			
Track number	Rt	EDT	C_{80}	t_s
1	32	62	72	83
2	31	61	72	81
3	36	57	65	77
4	49	85	94	98
5	32	68	82	86
6	31	58	56	75
7	29	46	60	67
8	22	32	40	42

Table 6-4. Results from training the ANN on music signals using the modified envelope spectrum pre-processor.

Even with the modified pre-processing, the unevenness of the power spectrum of the music across the octave band has a significant effect on the accuracy of the parameter estimations. Figure 6-16 illustrates this for reverberation time. Seven source signals are used comprising six pieces of music and one piece of speech. For each source signal, the ANN within the envelope spectrum method is trained. The simulated RIR database is used and the number of estimates within the DL rate is plotted against the spectral variance calculated across the octave band. The results show that as the excitation becomes more even across the octave band, in other words as the spectral variance decreases, the envelope spectrum method becomes more successful at estimating reverberation time. A similar trend occurs with the other parameters also.

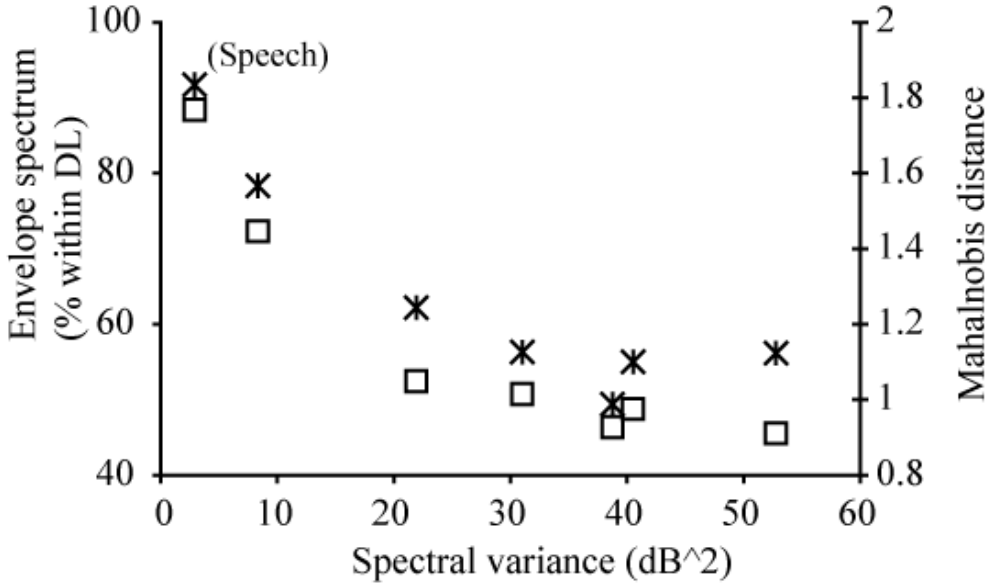


Figure 6-19. Variation of the Mahalanobis distance (□) and envelope spectrum method accuracy (*) with the variance of the music spectrum. 6 pieces of music and one piece of speech(EDT. after Kendrick et al. [73].

6.9 Real room impulse response parameter estimation

It is useful to validate the method's performance by presenting the trained ANN with signals convolved with real room impulse responses. This will give an indication of how well the system may perform in a real measurement task and give an indication as to how well the simulated RIR database represents reality. In this section an ANN is trained using the artificial database of room impulse responses and then presented with envelope spectra from music reverberated in real rooms. The modified 12^{th} octave envelope spectrum detector is used.

Figure 6-20 and Figure 6-21 show the error in the Rt and EDT as a function of their true value. In comparison to estimates from simulated RIRs (Section 6.8.4), the estimation of reverberation time gives a similar, poor level of accuracy. As previously mentioned, it is suggested that this arises because of the masking of the later parts of the impulse response by subsequent notes. The envelope spectrum method carries out an evaluation on the whole music passage, and in a piece of reverberated music the early decays of notes are going to be more prominent than later decay portions. Consequently, the envelope spectrum method struggles to accurately estimate reverberation times because

the information about late decay is lost in the vast amount of data from the whole music passage. As previously investigated in Section 6.6, companding can be carried out on the signal before it is fed to the envelope spectrum pre-processor. Companding biases the signal towards late decay and, as expected, this improves the reverberation time estimation but at the expense of accurate EDT estimation. This was not such a problem with speech, because speech has more periods of quiet.

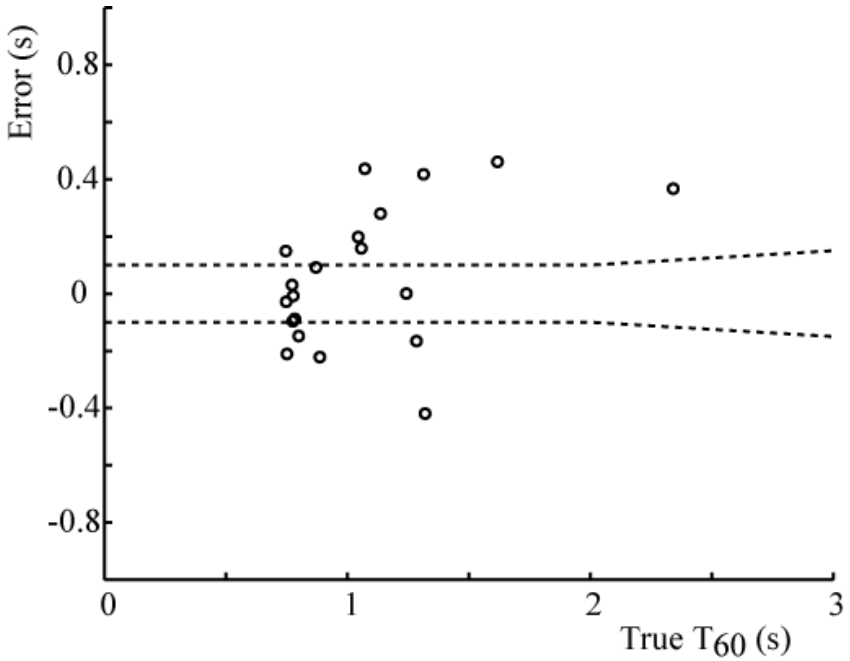


Figure 6-20. Error in Reverberation time estimation, using set of real room impulse response as test set, dotted lines indicate subjective difference limens.

In Figure 6-21 the results for EDT using music signals are presented, the estimation is quite successful where most results are within the subjective difference limens. This is comparable with the result from the simulated data set although there is a slight tendency for overestimation at low EDTs. This bias issue is even more apparent when evaluating the performance on C_{80} and t_s (Figure 6-22).

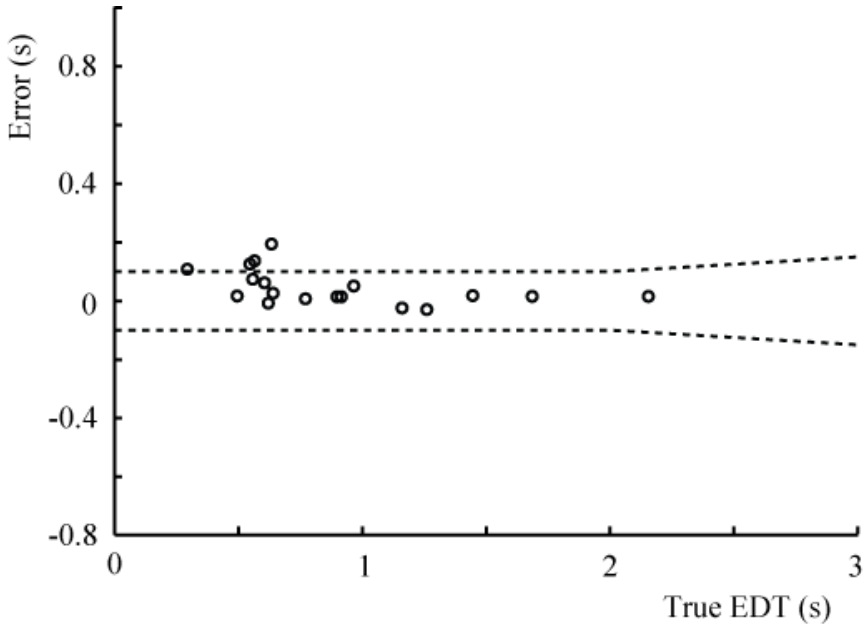


Figure 6-21. Error in EDT estimation, using set of real room impulse responses as test set.

For C_{80} and t_s a bias error is introduced, with the parameter values being underestimated for t_s and overestimated for C_{80}. The introduction of a bias with the envelope spectrum method estimation results is probably indicative of differences between the training dataset using simulated room impulses responses and the real room measurements. The loss in accuracy with the envelope spectrum method when estimating some parameters probably occurs because the simulated room impulse responses used to generate reverberated speech for training the ANN, are not completely representative of real room impulse responses. Consequently, the data used for training and testing have some significant statistical differences. The introduction of a bias as shown with some parameters and illustrated in Figure 6-22 is good evidence for this. As an ANN works to minimise the mean squared error, a well trained ANN should not generate a bias error unless something is wrong, such as the validation, test and training sets being different. It might be anticipated that, as the accuracy of geometric room acoustic models improves, this problem should disappear because the training set will better match reality.

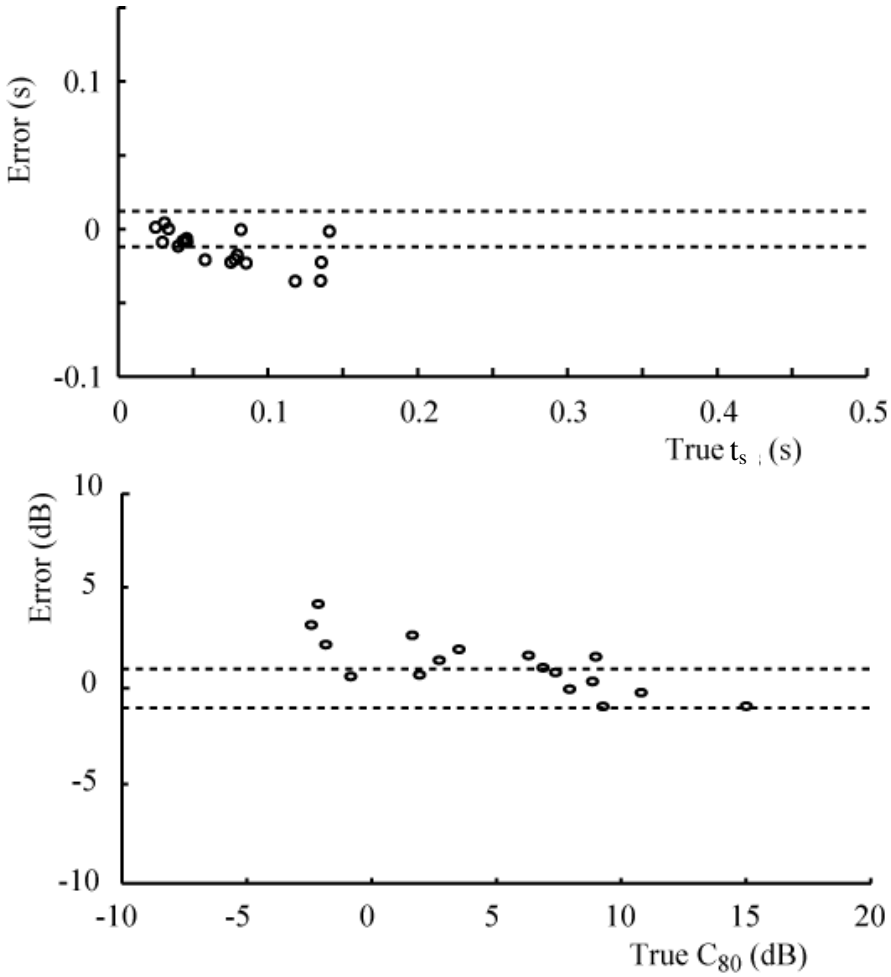


Figure 6-22. Error in t_s and C_{80} estimation, using set of real room impulse responses as test set, dotted lines indicate subjective difference limens.

One possible reason for the discrepancy between the real impulse response and the artificial RIR may be related to the fact that the impulse response is generated one octave band at a time. This means there is a sudden change in the response between adjacent octaves. This is not a feature in real rooms and with music the presence, or lack of, notes at the edges of octave bands can cause significant differences to the impulse response decay.

6.10 Companding music signals to increase Rt performance

As demonstrated on speech signals, companding can be used to increase the sensitivity of the envelope spectrum detector to changes in the envelope spectrum due to the late

reverberation. The same principles are applied to music signals in order to increase the accuracy of the reverberation time estimation. A μ-law compander is placed after each $1/12^{th}$ octave filter, this ensures that the late response is emphasised in the same manner for each narrow band prior to normalisation and recombination. To evaluate the performance a single piece of music is used and the Rt performance prior to and after the μ-law companding compared. A μ value of 150 is used and once again the quoted percentage is the average performance from 10 attempts to train the ANN. Table 6-5 compares the ANN estimation of Rt with and without companding and shows a small improvement in accuracy can be gained.

Rt Performance without companding (% within DL)	Rt Performance with companding (% within DL)
49 %	53%

Table 6-5. Companding music signals, comparison of the percentage of results within the DL with and without companding, results generated from a single piece of music, ANN training was repeated 10 times and the resulting performance indicator averaged.

Companding has been shown to work well for speech signals and much less well for music signals. The reason for the difference in improvement is due to the number of regions in the signal where late decay is available to be emphasised, within music there are less totally silent regions as notes tend to gradually decay away rather than cease abruptly. Additionally, in the presence of noise, companding can causes more problems as low level noise is amplified and this increases the noise level on the envelope spectrum.

6.11 Discussion

The envelope spectrum method has been modified so that in addition to speech signals, music signals can be utilised for the purpose of measuring acoustic parameters. Music signals are not ideal due to their uneven spectra. Acoustic parameters are usually measured using broadband signals and music does not excite all frequencies evenly. The signal power from music signals is strongly focused in narrow-bands centred on the equal tempered scale and the key used in the piece. This chapter demonstrated that a modified pre-processer can be used to improve accuracy. The modified pre-processor uses a bank of band-pass filters, where each filter is centred on a note in the equal

temperament scale. Each narrow band signal is then normalised before recombination. This effectively yields an envelope from a whitened version of the music signal. This methodology alone is insufficient to account for all variance in spectral content as music contains gaps in the spectrum where there is insufficient signal-to- noise ratio, no amount of whitening can compensate for this and as such it is an inherent limitation of the envelope spectrum method.

This method, used with one of the music signals trained and tested using simulated impulse responses can produce good estimates for EDT, t_s and C_{80} with up to 85%, 94% and 98% within the difference limen respectively. The Rt estimation accuracy is less successful with 49% being within the DL. Different pieces of music yield worse results. The ideal music signal is one that has plenty of short transient sounds and whose excitation is even across the octave band, in other words a piece of music which breaks traditional rules of western harmony and gives the same weight to all notes in the chromatic scale. Alternatively, averaging over different pieces of music can be effective, especially if the different pieces are in different musical keys. Another approach is to use pieces with lots of untuned percussion. The best music for this purpose also has large gaps between the notes so that the decay phases are prominent.

Additionally, the envelope spectra of music signals are not ideal for this machine learning approach because of the inherent patterns in music. There are gaps in the envelope spectrum of music signals, which is problematic especially when considering non-uniform decay curves. These require good estimates of the modulation level over a range of modulation frequencies for accurate estimations. In particular, low modulation frequencies are required for accurate Rt estimations and both speech and music lack the level of low frequency modulation to accurately estimate long Rts. A broad range of modulation frequencies are required when the RIRs have non-uniform decays. With speech and music the late reverberation is masked by subsequent notes and early reflections. If the late decay rate is different to the early decay rate the early reflections, which are dominant in the envelope spectrum, cause errors when estimating Rt. One way to counteract this is to find signals with more low modulation frequencies in their envelopes (longer gaps between notes or utterances). A second way is to use companding to emphasise quieter signal components and deemphasise louder components. This effectively emphasises the effect of the later reflections while

deemphasising the effect of early reflections. Care must be taken when companding as this can effectively decrease the signal to noise ratio of the signal. A third possible method would involve pre-processing the signal so that only regions of free decays are used to compute the input to the ANN. This will automatically minimise the direct sound components and maximise the amount of reverberant decay at the input to ANN. This would also minimise the amount of masking of late reverberation by direct sound and early reflections.

The ANN can also be trained to estimate the MTF from reverberated envelope spectra. This enables the decay curve to be estimated blindly from music and speech and all acoustic parameters can be calculated from this estimated decay curve. This negates the need to train one ANN for each acoustic parameter, although the accuracy of the parameters is significantly reduced when compared with the other approach. This is because the ANN is presented with the more difficult task of finding the optimal solution in much larger dimensional space.

In addition to the limitations placed on accuracy due to the uneven spectrum of music signals and the lack of modulation at certain points in the envelope spectrum, music envelope spectra differ considerably for different pieces of music. This means that signal independent estimation using static neural networks is not feasible. The envelope spectrum method has been shown to be very successful at estimating EDT for both simulated and real RIRs using speech and music, providing care is taken when choosing the piece of music (plenty of long gaps between notes and a relatively flat audio-spectrum). However the late reverberation time estimation is less successful and a bias is introduced for other early parameters such as C_{80} and t_s because of the difference between the geometrically simulated RIR database and real room responses.

In conclusion, this method demonstrates some success at parameter estimation with music signals but the limitations discovered direct the research towards seeking alternative methods. Alternative approaches are required to achieve signal independence and to improve the accuracy of the parameter estimations (in particular Rt). The limitations of this machine learning approach inspired the adoption of an alternative approach which is investigated in the next chapter. This alternative approach

utilises multi-decay rate model of sound decay within a maximum likelihood framework. Results from the present chapter have been published in [73, 74].

7 IMPROVED MAXIMUM LIKELIHOOD ESTIMATION OF ACOUSTIC PARAMETERS

In this chapter the development of an improved method for blindly estimating room acoustic parameters from speech and music is presented. The method uses a maximum likelihood framework to estimate parameters from decay phases automatically selected from speech and music signals. The method yields estimates for monaural parameters such as Rt, EDT, C_{80}, D, t_s, and binaural parameters such as ELEF and LG. The method is based on a method first reported by Ratnam *et al*. [3, 46]. The novelty centres on: 1) a multi-decay rate model that is capable of modelling non-diffuse and coupled room responses. 2) The ML formulation and optimisation of this new model on automatically selected reverberant decay phases. This combines *brute force* optimisation with a *gradient search* algorithm and utilises the symmetry of the likelihood function to reduce the number of function evaluations. 3) The ML estimations produce many decay curve estimates. This requires a decision framework to improve the blind selection, the 'best estimates' and to reduce the ambiguity between the early and late decay rate estimations. Several frameworks are presented with this aim, which attempt to minimise the influence of the tails of musical notes and utterances on the parameter estimates, while also accounting for the stochastic nature of the source signal.

7.1 Background

Ratnam *et al*. [3, 46] recently proposed a method to blindly estimate reverberation time (Rt) using a maximum likelihood estimation approach. A stochastic model of sound decay was used to produce ML estimates of Rt from decays following speech utterances. Although only one real speech example was presented and the accuracy was generally poor for realistic rooms, the results and the theoretical framework show the potential of the MLE approach for the blind estimation of room acoustic parameters.

A limiting aspect of the method was the use of an idealised single exponential decay as the decay curve model. Consequently, a diffuse field was assumed, even for parts of the decay where early reflections dominate. In real rooms, non-exponential decays are commonplace. Therefore, acoustic parameters other than Rt, such as EDT, are needed

for room assessment. The single exponential model in the MLE method makes the estimation of the decay curve inaccurate especially when the curve deviates from being uniform. As a result, even with idealised excitations such as bursts of white noise, errors exceed the perceptual difference limens for Rt. Despite its limitations, the method shows excellent potential for a blind method of acoustic parameter estimation as the major limitation is the simple model of sound decay used. By improving this model, so that it better represents what may occur in real rooms, an improvement in accuracy is expected. This chapter revisits and redesigns the method with the goal of improving accuracy.

The framework proposed by Ratnam and expanded upon in this chapter utilises the concept of likelihood. Likelihood is related to probability. Probability is the 'chance' of an event occurring where the parameters (such as mean and variance) describing the probability density function (PDF) of that event are known. Likelihood is the 'chance', given a recording of an event, that a particular set of parameters (that describe the PDF) were responsible for generating that data. Likelihood can be thought of as inverse probability, though its inventor, Fisher, later disputed its connection to inverse probability [75]. The development of the maximum likelihood method is perhaps responsible for the creation of a great many universal terms within statistics such as 'Statistic', 'Parameter', 'Likelihood' *(of course!)*, and perhaps even 'Estimation'. By adjusting the parameters within a likelihood function for a set of recorded data, there exists a maximum value for likelihood. These parameter values are known as the 'Maximum Likelihood Estimates'.

7.1.1 Determining reverberation time from decay phases

In this thesis, decay phases are defined as the silent portions in a signal in-between utterances or notes, where the predominant sounds are due to reverberation and echoes. The use of decay phases following speech utterances or stop-chords in music to estimate the reverberation time is not new (see chapter 3). Cremer and Müller suggested that the reverberation time could be estimated from the decay curve following a loud stop-chord from a Beethoven symphony [76]. Cox *et al.*'s method used a statistical machine which learnt how to determine reverberation parameters from the decay phases after separated speech utterances [40]. Several other methods for reverberation time estimation that

involve the fitting of energy decay models to segments of received signals have been suggested [41, 45 and 42]. The key to success and accuracy is to choose suitable statistical models and tools to quantify correctly the decay characteristics from the numerous available decay phases, whilst accounting for the presence of noises and source fluctuations.

The maximum likelihood method has several advantages over other 'stop-chord' methods. Utilising a realistic model of the process can be advantageous, the rationale being, that if the model chosen is representative of the physical process then the resulting estimate will be more realistic. Another advantage of the ML method is that the method is capable of providing estimates of the decay which are longer than the decay phase used to gain the estimate. In other words, the method is capable of using a limited time record and by using the sound decay model extrapolating it to predict the response for a longer time record. This is a particular advantage when Schroeder backwards integration is performed on the estimates.

Figure 7-1 shows examples of decay phases found in received speech signals in anechoic and reverberant conditions. By capturing the instances that are suitable for decay curve estimation and then fitting the falling edge with a suitable regression model, a blind estimation of the reverberation time can be made. The success of this method relies on two factors:

(1) The source signals, music or speech, should contain sufficient and relatively clean silent regions which are preceded by loud and sharp endings to utterances/notes. This is the intrinsic requirement of the approach. Fortunately, as discussed and will be empirically proven below, suitable signals are generally available in most speech or music sources.

(2) An effective statistical method to reconstruct or estimate the decay curve. The decay phases found in natural sound sources are far less uniform than those arising from the switching off of a white noise source. Due to the non-stationary nature of the sources, a rigorous statistical method must be adopted. Figure 7-1 shows an example of a decay phase and details how this changes when it is reverberated.

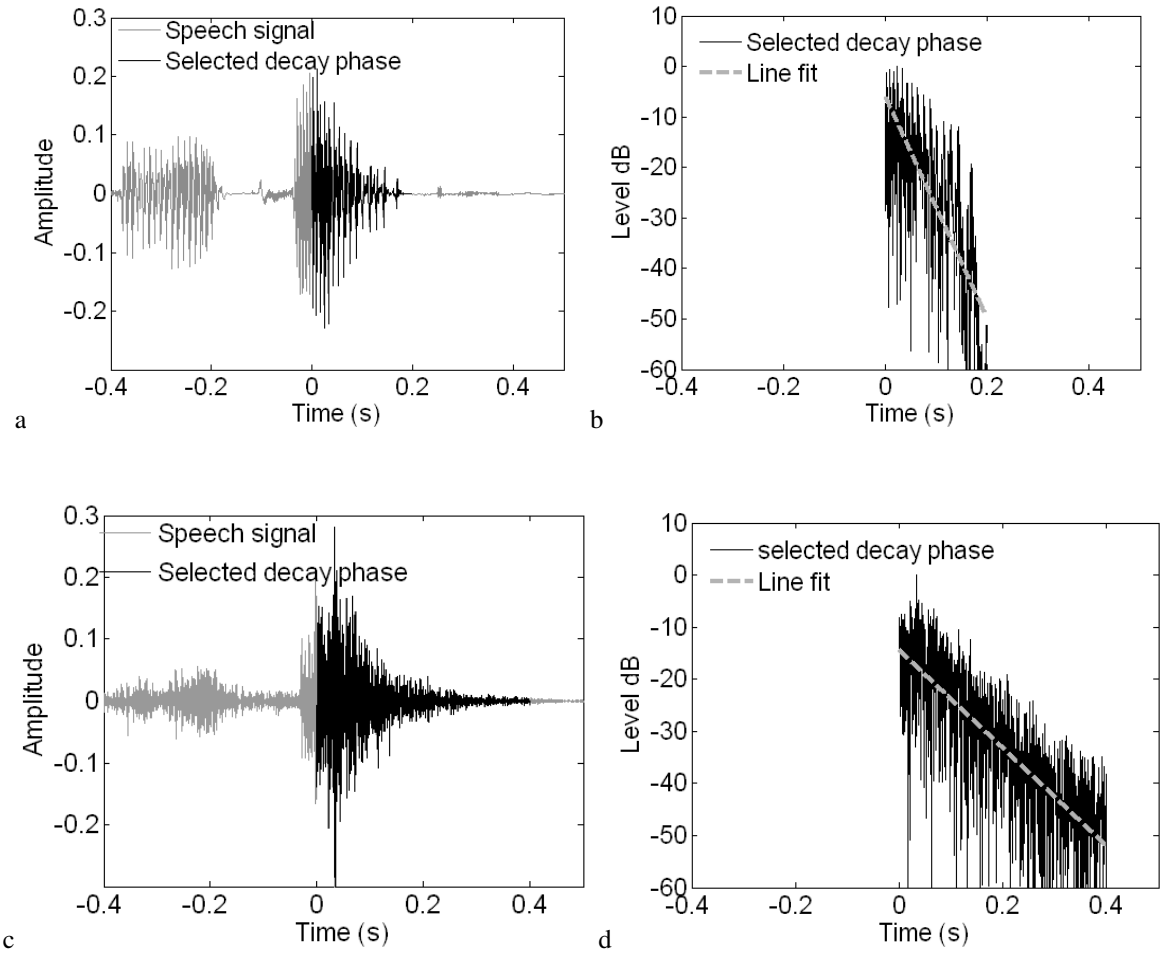


Figure 7-1. (a) and (b) the energy decay of an anechoic speech utterance saying: (dec)-'imal point'. The utterance itself has a decay time equivalent to a "reverberation time" of 0.3s. (c) and (d) depict the same section of speech but recorded in a space with a reverberation time of 0.63s, the resultant decay rate of the reverberant section is equivalent to a T_{30} of 0.64s.After Kendrick et al.[9].

7.1.2 The Ratnam method for blind estimation of Rt

The Ratnam method applied short rectangular windows to reverberated speech signals and performed a maximum likelihood estimation of the decay curve on each of these windows. This was performed on the whole received signal not just the decaying portion. Subsequent analysis of the estimated parameters enabled the method to identify the decaying portions of the sound. In order to correctly assign the true decay rate from these estimates, two possible approaches were suggested.

1) From the histogram distribution of MLE parameter estimates, as shown in Figure 7-2, an order statistics filter is employed where the true parameter is chosen from a certain percentile of the distribution. From this histogram the final Rt estimate is calculated by finding the point where the cumulative probability is equal to a certain value (0.3 in this case). The value 0.3 is chosen arbitrarily in this example to illustrate the procedure, in practice the value can be chosen based on prior knowledge regarding the proportion (in time) of signal compared with periods of silence. This is reasonable in the case speech as it is related to the speed of talking which is fairly consistent even between people, but for music this is particularly difficult.

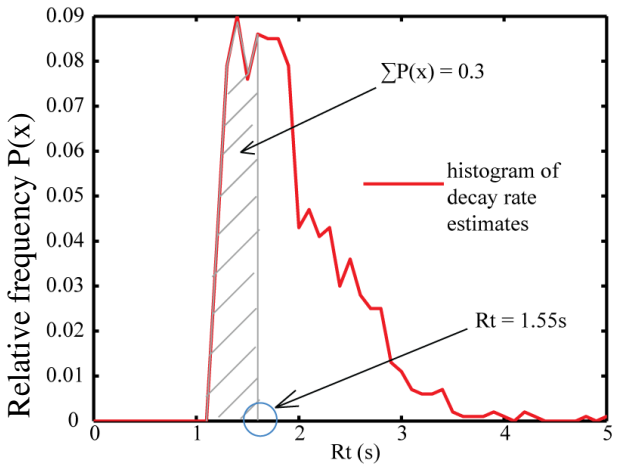


Figure 7-2. Histogram of ML decay curve estimates, showing order statistics method of choosing Rt estimate.

2) The distribution may be multi-modal due to varying rates of decay of the tails of the speech utterances. In this case the algorithm picked the peak corresponding to the lowest decay time. Figure 7-3 shows an example of a multi-modal distribution of Rt estimates, highlighting the first peak. A further extension of the work [46] used a quantised grid of parameter values, each of which represents a bin in the histogram. This improved the optimisation speed and resulted in a fast online version of the algorithm.

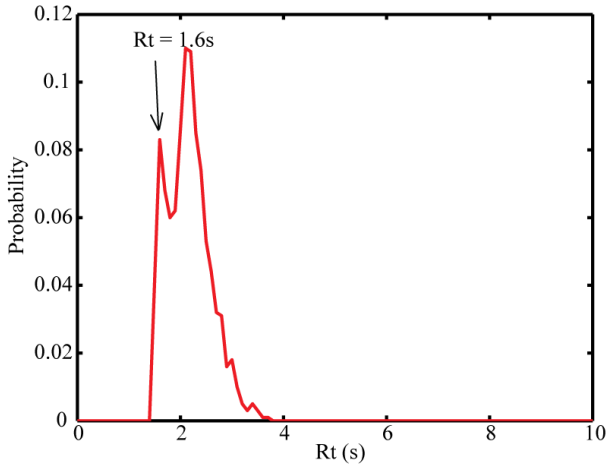


Figure 7-3. Histogram of ML decay rate estimates, showing how the first peak is selected as the best estimate.

This method has a number of limitations and addressing these has been the focus of the development of the maximum likelihood based algorithm:

- The room model assumes uniform decay. In reality this is rarely true. Often the early part of the sound field has been intentionally 'tuned' to improve the perception of the acoustics, or the space may be coupled with another room with very different properties causing non-uniform decay. This algorithm cannot differentiate between the early and late parts of the sound decay and this can cause the distribution of estimates to be skewed, making the task of selecting the best estimate very difficult.

- The distribution of decay rate estimates is signal dependent. In the case of speech, it is subject to factors such as the speed of talking and the length of gaps between words. Although the order statistics filter and the peak picking algorithms attempt to compensate for this, the method is not truly blind as the algorithms must be tuned according to the signal properties. This is because the histogram is not only dependant on the room response but also the source signal.

- When extending the method to other signals such as music, the previous problem becomes more apparent as the decay parameter distribution is much more heavily skewed due to the decays of the musical note.

- Performing MLE estimations for each short window means the method has a large computational overhead. A large number of these estimations are superfluous because they are performed on regions of the signal which are not decay phases.

In order to address these problems several of novel solutions are proposed and investigated:

1. New room models are developed which account for non-uniform decays of room responses.

2. New pre-processors are used to isolate regions of continuous decay to minimise the number of MLE stages (developed by *Zhang et al.* [50]).

3. The ambiguity between the early and late decay regions is reduced to enable accurate estimation of both.

4. More sophisticated selection processes for the resultant estimates are investigated to improve the blind nature of the algorithm.

7.2 An improved model for sound decay in a room for use in a maximum likelihood framework

This next section details the development of a multi-rate model of sound decay that approximates non-diffuse or coupled spaces.

7.2.1 Modelling the decay curve

Kuttruff [16] presents a description of the decay of sound energy in an enclosure derived from wave theory (from the Helmholtz equation). The derived function describes the room response as the superposition of modes, where modes are resonances of a room at specific frequencies. The overall response of the room can be thought of as the superposition of all modes of vibration for all frequencies. Each mode has a damping constant which is due to the particular dimensions and boundary conditions of

the room. It is these damping constants that are responsible for the rate of decay of the sound energy in the room. Kuttruff shows that when all of the damping constants are equal, the logarithmic decay curve is straight (i.e. the response is perfectly exponential). In real rooms, the damping constants are often very close to one another and therefore logarithmic sound decay curves often approximate straight lines. Kuttruff derives the response of a room to an impulse, as a sum of modal responses from a number of sinusoidal oscillations each with its own damping factor;

$$h(t) = \sum_n c_n e^{-\delta_n t} \cos(\omega_n - \varphi_n) \; for \; t>0 \qquad (7\text{-}1)$$

where n is the mode number, $h(t)$ is the room response, c_n is the modal amplitude, δ_n is the modal damping factor, ω_n is the modal frequency and φ_n is a phase shift. By squaring and then averaging with respect to time the cosine terms can be averaged out and the energy decay $h^2(t)$ is shown to be proportional to;

$$h^2(t) \propto \sum_n c_n e^{-2\delta_n t} \qquad (7\text{-}2)$$

As the damping factors δ_n in real rooms are often very close, the damping factor can be replaced, with minimal error, by the average damping factor $a = < \delta_n >$. This means that the sound energy decay can now be characterised by a single decaying exponential, (the c_n parameter can now be replaced by C, the mean of all c_n);

$$h^2(t) \propto Ce^{-2at} \qquad (7\text{-}3)$$

The above formulation is a model of the decay of sound energy in a room that is comparable to Ratnam's model. This simple model and cannot account for spaces with non-uniform decay rates, it is interesting to note that Kuttruff [16] presents a model of sound decay for N coupled rooms that is very similar to Equation (7-2).

$$h^2(t) \propto \sum_{i=1}^{N} \alpha_i e^{-2a_i t} \qquad (7\text{-}4)$$

The major difference with the coupled room model is that the parameter α_i is not restricted to positive values and the $a_{i,}$ values can no longer be replaced by their average as different rooms may have substantially different modal decay rates, thus the decay curve will be more complicated with differing rates of decay in different regions.

Equation (7-4) shows how the decay curve can be modelled as a sum of exponentials. The popular image source method [77] of acoustic simulation also shows that the decay curve can be modelled as a sum of exponentials, this is described in Appendix A.

7.2.2 Modelling the temporal pattern of received reflections

Thus far the model describes the decay of sound energy. The model must be expanded to include the fine structure of the room impulse response; the temporal pattern of received reflections. The image source method [77] models the impulse response of a room as the superposition of a series of delayed attenuated impulses. By assuming non-frequency dependant reflection characteristics for the walls, the response can be modelled as a sum of dirac delta functions, $\delta(t)$, with varying delays, t_n , and the amplitude of each reflection is A_n.

$$h(t) = \sum_{n} A_n \delta(t - t_n) \qquad (7\text{-}5)$$

A typical example of this pattern of reflections is shown in Figure 7-4. The density of reflections increases quadratically with time according to the following function which is derived from an image source model of a rectangular room. This function is also approximately valid for a room with an arbitrary shape.

$$\frac{dN_r}{dt} = 4\pi \frac{c^3 t^2}{V} \tag{7-6}$$

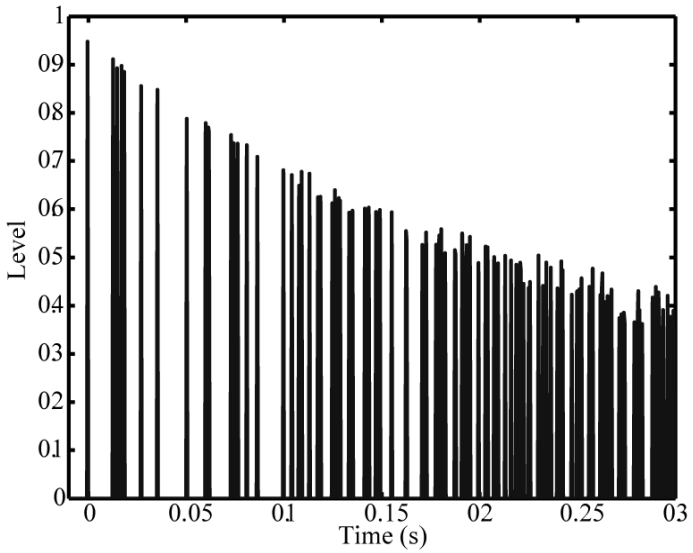


Figure 7-4. A room impulse response showing the increase in reflection density with time.

Polack [78] [79] developed a room impulse response model based upon the modulation of a stochastic process by an exponentially decaying curve. The stochastic process was used to represent the individual reflections, or the fine structure of the impulse response.

$$h(t) = g(t)e^{\frac{-t}{T}} \quad \text{for } t \geq 0 \tag{7-7}$$

where *g(t)* is a random Gaussian process modulated by an exponentially decaying function and *T* is related to the reverberation time. There is a time delay prior to this model becoming valid. The reflections can only be considered Gaussian-like if a sufficient number of discrete reflections overlap at any one time along the impulse response [79]. Polack described this limit as the point where 10 or more reflections overlap in a 24ms window. This 'mixing time' t_{mixing} is the location (in time) of the transition between early and late sound and is approximately given by [78];

$$t_{mixing} \approx \sqrt{V}\ (ms) \qquad (7\text{-}8)$$

where V is the volume of the room in m^3. In addition to this, room boundaries exhibit frequency dependant absorption; this causes the decay characteristics to vary with frequency. Additionally, with each boundary interaction the sound wave will be filtered, this means that later reflections typically contain progressively fewer high frequencies.

The use of Gaussian noise to represent the fine structure of the individual reflections produces a probabilistic model suitable for use in a maximum likelihood framework. The next section describes how the stochastic models of the RIR fine structure in Section 7.2.2 and the energy decay models in Section 7.2.1 are combined to realise the room impulse response model.

7.2.3 A new model of sound decay in a room for MLE

For the purposes of model simplicity and reliability of the estimation system, the temporal structure of reflections is simplified to a Gaussian noise source as described in the previous section and by Polack [78]. Therefore the room response is modelled without the transition between the early and late sound fields. The early temporal structure is lost and the gradual increase in reflection density between early and late sound-field will not be included in the model. This is deemed as an acceptable approximation as most acoustic parameters are related either to average rates of decay (Rt/EDT) or sums of energy within substantial time periods which contain many reflections (C_{80}, D etc). Therefore errors due to the lack of information about the fine temporal structure are presumed to be small enough to be acceptable. The time dependant filtering effect due to wall absorption will also not be included in the model as this would add substantial complexity to the model and it is thought that it will add little accuracy to the estimation of the energy decay curve. By yielding decay curve estimates for each octave band (assuming no time-dependant filtering effects in each octave band) the MLE method builds up an estimate that includes approximate time dependent filter effects as the octave bands decay at differing rates.

Using all this information the new room impulse response model is defined as a Gaussian noise source modulated (as described in Section 7.2.2) by an envelope which is a sum of exponentials (as described in Section 7.2.1);

$$h^2(t) = g(t)\sum_{i=1}^{M}\alpha_i e^{-a_i t} \qquad (7\text{-}9)$$

where $g(t)$ is the noise signal, M the number of exponentials in the model, α_i is the weight of each exponential and a_i is the decay constant for each exponential. This can also be rewritten in discrete form;

$$h[n] = g[n]\sum_{i=1}^{M}\alpha_i a_i^{\,n} \qquad (7\text{-}10)$$

where n is sample number from 0 to N. Figure 7-5. shows an example of a three exponential room model using equation (7-10), where the magnitude parameters, α_i and exponential parameters, a_i are; α_i=*[0.4983 0.214 0.6435] and* a_i=*[0.9997 0.9990 0.9993]*

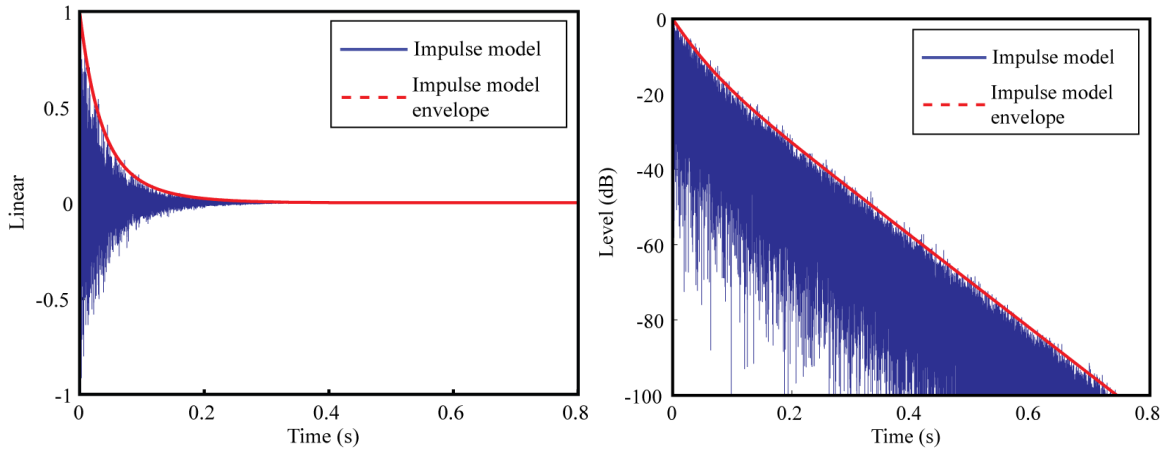


Figure 7-5. This figure shows an example of the RIR model generated from the above parameters. The plots show both the estimated envelope of the RIR (in red) and the RIR estimate (in blue)

To calculate reverberation time and the early decay, Schroeder backwards integration is used [10]. This is a method by which the average of all possible sound decays can be

computed and so it is a mathematically precise and robust definition of the decay curve. In this case, the sound decay is the result of exciting a room with a random noise signal, waiting until the level reaches a steady state and then switching it off.

$$\left\langle d^2(t) \right\rangle = \int_t^\infty h^2(t)dt \qquad (7\text{-}11)$$

where $<d^2(t)>$ represents the average of all possible decay curves. Figure 7-6 shows that the backwards integration of the squared RIR model and the squared envelope of the model are equivalent. This is because the squared noise signal $g^2(t)$ component has an expected value of 1 and therefore the expected squared RIR is simply the envelope squared.

$$E[env^2(t)g^2(t)] = env^2(t)E[g^2(t)] = env^2(t) \qquad (7\text{-}12)$$

Most acoustic parameters are calculated from sound energy levels in a given time window. Therefore $env^2(t)$ can be used to define the expected decay curve and from that, most of the acoustic parameters required. This removes fluctuations due to the random Gaussian noise and negates the need for averaging as demonstrated in Figure 7-6 (b).

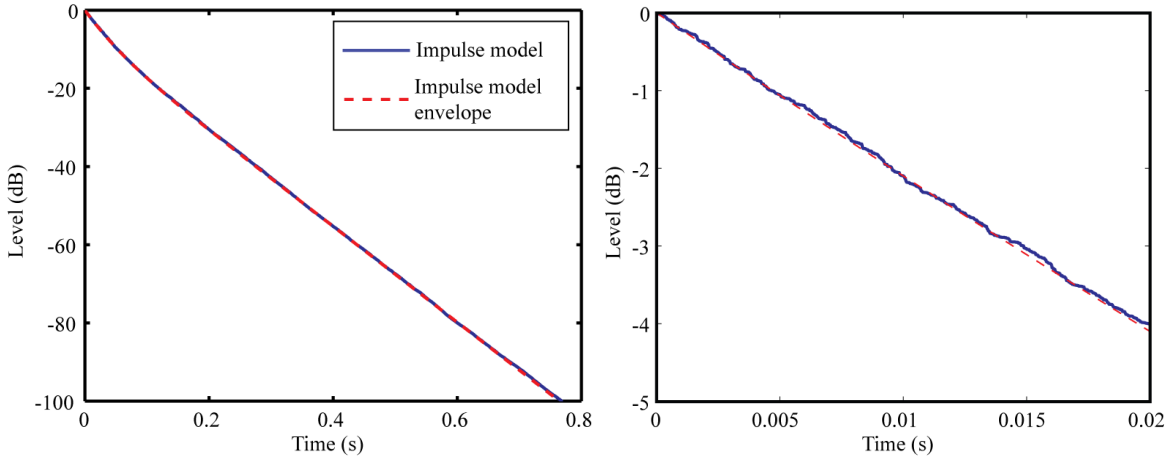


Figure 7-6. Schroeder curve of room impulse response models, the backwards integrated impulse response model and the backwards integrated RIR envelope are shown to be similar. a) shows the full dynamic range of the decay while b) shows a small section highlighting the random fluctuations in the impulse response model.

Equation *(7-13)* is a parameterisable model of sound decay where the complexity of the model comes from the number of exponential parameters that are used in the function *env*(*n*) that modulates a noise signal

$$h[n] = env[n]g[n] \tag{7-13}$$

where from Equation (7-10):

$$env[n] = \sum_{i=1}^{M} \alpha_i a_i^{\,n} \tag{7-14}$$

A common occurrence in reality is that there are two decay regions in the decay curve, such as in a coupled room or a space with non-uniform distribution of absorbing material. Pilot indicated that two decays were appropriate for the purposes of determining acoustic parameters with sufficient accuracy. A number of tests were run using three or more exponentials and it was found that often the accuracy of the acoustic parameters was worse due to the increased difficulty in optimisation due to the increased number of dimensions. To further simplify the model and reduce the number of parameters, a single weighting parameter is used

$$env[n] = \alpha a_1^{\,n} + (1-\alpha) a_2^{\,n} \tag{7-15}$$

where a_1 and a_2 represent two differing decay rates and α is a weighting factor, limited to the values between zero and unity, that changes the amount of influence each decay rate has, thus enabling the representation of an energy response with a non-uniform decay rate. Figure 7-7 shows this 'sum of two exponentials model' with different decay rates, where the factor α acts to define a knee point where the influence of the two decay rates cross over from a_1 to a_2 where a_1=*0.995* and a_2=*0.99*.

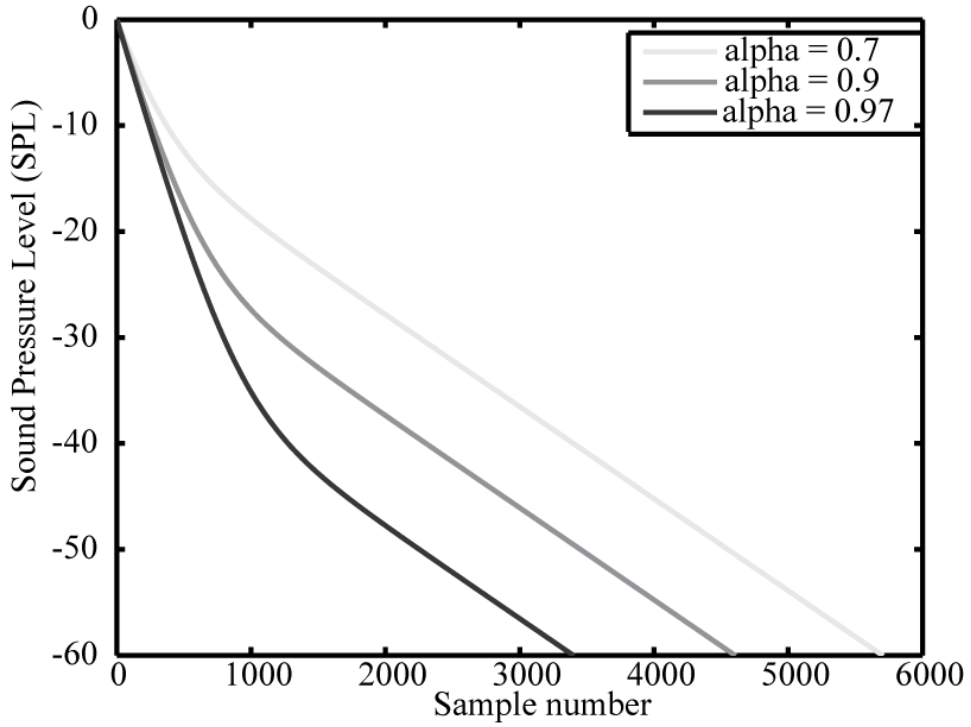


Figure 7-7. A dual exponential decay model can give non-uniform decay, where the faster decaying exponential dominates the early part of the decay curve, and the slower decaying exponential controls the late part of the decay. The decay envelopes for three different weighting factors as shown in the legend are illustrated.

7.2.4 The maximum likelihood formulation

Now that the parametric sound decay model has been defined, a method needs to be chosen to estimate the best model parameters from a given recording. Referring to Equation (7-13), a recording of the response of a room to an impulse source ($h_r[n]$) when divided by the envelope $env[n]$, can be approximately modelled simply by a random Gaussian variable

$$g[n] = \frac{h_r[n]}{env[n]} \tag{7-16}$$

The probability density function of a Gaussian variable $g[n]$, where μ is the mean and σ^2 the variance is [80];

$$f\left(g[n]\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{\left(-\frac{(g[n]-\mu)^2}{2\sigma^2}\right)} \tag{7-17}$$

The likelihood of a sequence of *N* independent, identically distributed Gaussian variables occurring is given by the conditional probability (multiplication) of each *f(g[n])* is [80];

$$L(g[n];\sigma,\mu)=\prod_{n=0}^{N-1}\frac{1}{\sqrt{2\pi}\sigma}e^{\left(-\frac{(g[n]-\mu)^2}{2\sigma^2}\right)} \tag{7-18}$$

where g[n] has been observed and recorded. Since μ = 0 it can be removed from the expression. Substituting in Equation (7-16);

$$L(h_r[n];\sigma,env[n])=\prod_{n=0}^{N-1}\frac{1}{\sqrt{2\pi}env(n)\sigma}e^{\left(-\frac{h_r[n]^2}{2\sigma^2 env[n]^2}\right)} \tag{7-19}$$

This can be rearranged to:

$$L(h_r[n];\sigma,env[n])=e^{\left(\sum_{n=0}^{N-1}\frac{-h_r[n]^2}{2env[n]^2\sigma^2}\right)}\left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\prod_{n=0}^{N-1}\frac{1}{env[n]} \tag{7-20}$$

Substituting *env*[*n*] for the new decay model in Equation *(7-15)* we get;

$$L(h_r[n];a_1,a_2,\alpha,\sigma)=e^{\left(\sum_{n=0}^{N-1}\frac{-h_r[n]^2}{2\left[\alpha{a_1}^n+(1-\alpha){a_2}^n\right]^2\sigma^2}\right)}\left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\prod_{n=0}^{N-1}\frac{1}{\left[\alpha{a_1}^n+(1-\alpha){a_2}^n\right]} \tag{7-21}$$

It is more convenient to work with the log of the likelihood function, as the product terms become sums. The log likelihood function is:

$$\ln L(h_r(n);\sigma,a_1,a_2,\alpha) = -\sum_{n=0}^{N-1}\frac{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]^{-2} h_r(n)^2}{2\sigma^2} - N/2\ln\left(2\pi\sigma^2\right) - \sum_{n=0}^{N-1}\ln\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right] \quad (7\text{-}22)$$

To find the maxima/minima of this function, score functions [81] are computed by differentiating the log likelihood function for each parameter;

$$\frac{\delta \ln L}{\delta\sigma} = -\sum_{n=0}^{N-1} - \frac{h_r(n)^2}{\left(\alpha a_1{}^n + (1-\alpha)a_2{}^n\right)^2\sigma^3} - N/\sigma \quad (7\text{-}23)$$

$$\frac{\delta \ln L}{\delta a_1} = -\sum_{n=0}^{N-1} - \frac{h_r(n)^2\alpha a_2{}^n n}{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]^3\sigma^2 a_1} - \sum_{n=0}^{N-1}\frac{\alpha a_1{}^n n}{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]a_1} \quad (7\text{-}24)$$

$$\frac{\delta \ln L}{\delta a_2} = -\sum_{n=0}^{N-1} - \frac{h_r(n)^2(1-\alpha)a_2{}^n n}{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]^3\sigma^2 a_2} - \sum_{n=0}^{N-1}\frac{(1-\alpha)a_2{}^n n}{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]a_2} \quad (7\text{-}25)$$

$$\frac{\delta \ln L}{\delta\alpha} = -\sum_{n=0}^{N-1} - \frac{h_r(n)^2\left(a_1^n - a_2^n\right)}{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]^3\sigma^2} - \sum_{n=0}^{N-1}\frac{\left(a_1^n - a_2^n\right)}{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]} \quad (7\text{-}26)$$

The solution when the score functions are zero provides the locations of the extrema of the function.

$$\begin{aligned}\frac{\delta \ln L}{\delta\sigma} &= 0\\ \frac{\delta \ln L}{\delta\alpha} &= 0\\ \frac{\delta \ln L}{\delta a_1} &= 0\\ \frac{\delta \ln L}{\delta a_2} &= 0\end{aligned} \quad (7\text{-}27)$$

For σ^2, the variance of the Gaussian processes, the equation can be solved directly and the following function can be used to evaluate the value of sigma at each iteration:

$$0 = -\sum_{n=0}^{N-1} - \frac{h_r(n)^2}{\left(\alpha a_1{}^n + (1-\alpha) a_2{}^n\right)^2 \sigma^3} - N/\sigma$$
$$\sigma = \sqrt{-\sum_{n=0}^{N-1} - \frac{h_r(n)^2}{N\left(\alpha a_1{}^n + (1-\alpha) a_2{}^n\right)^2}} \qquad (7\text{-}28)$$

However solutions to the other score functions cannot be derived analytically. The optimisation of the –ln(*L)* is computed numerically. Optimisation was introduced in chapter 5.1 and the various techniques adopted for finding the solution to this function are described in the next section.

7.2.5 Optimisation scheme

Once a decay phase has been recorded, the likelihood function must be optimised so that the most likely set of model parameters responsible for generating the decay are found. In the case of a likelihood function, this is the maximum function value for all possible parameters. Locating the maximum of a complicated, perhaps multi-dimensional, function is often a difficult problem as discussed in Section 5.1. Figure 7-8 shows the likelihood function, using a simulated room impulse response as the input, evaluated over a grid of parameters values, in this plot a_1 and a_2 are plotted against the log likelihood value. Normally this exhaustive evaluation of all possible values is very time consuming. It is presented here to allow the nature of the likelihood function to be better understood.

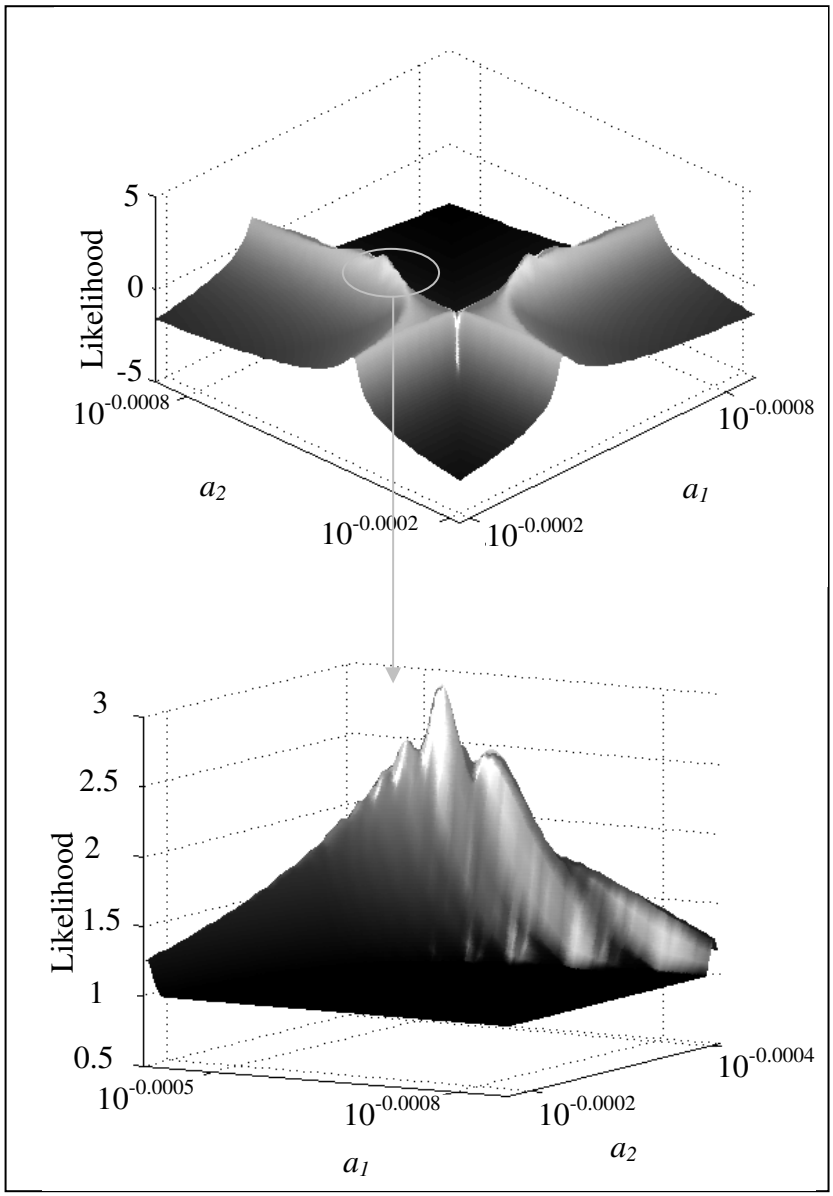


Figure 7-8. The likelihood function over a grid of exponential parameters a_2 and a_1. At each location the solution has been optimised with respect to the parameter; α. The maxima represent the local and global solutions. The bottom plot is a magnification of a section of the top plot as indicated.

It can be seen from Figure 7-8 that many local optima can exist. It is important to avoid these local optima to locate the global solution. Therefore the following multiple stage search method is employed as detailed in Figure 7-9. This is similar to the method employed by Apollo [82] for estimating the parameters of sums of exponentials contained in non-stationary noise. This technique uses a direct search method followed

by a gradient search method. It uses a large parameter range for the direct search, in order that the region the optimum solution exists in is found, before using a gradient search algorithm starting from that location to find the best solution.

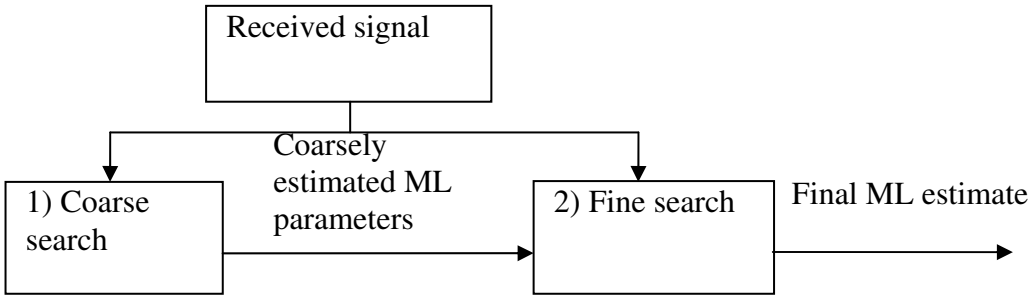


Figure 7-9. Minimisation overview. A two stage algorithm is employed. First a coarse search is applied to the likelihood function with the aim of finding the region where the global minimum is most likely to be present. Then these parameters are used as the starting point for a more detailed optimization.

The third parameter α is not shown in Figure 7-8. Parameter α controls the transition point where the rate of decay changes from being mainly dependant on one exponential to another. It turns out that the likelihood function for a varying α and fixed a_1 and a_2 is relatively smooth with a single maxima, this is common for all a_1 and a_2. An example of this is shown in Figure 7-10.

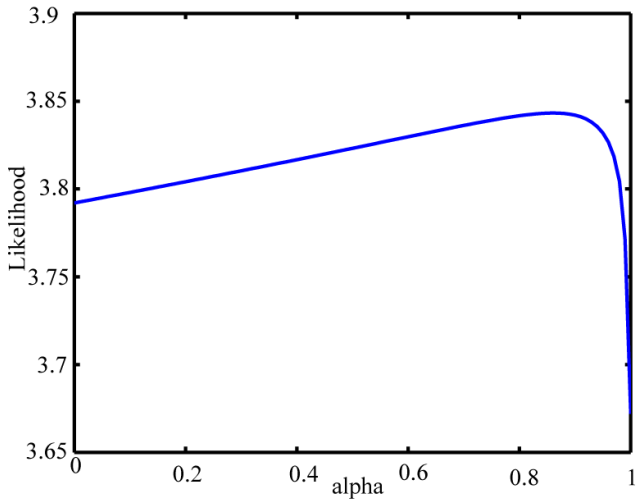


Figure 7-10. The likelihood across a grid of α parameters but with parameters a_2 and a_1 are fixed. This clearly shows the smoothness of the function and the single maximum for the α parameter.

Figure 7-8 has been generated utilising the smooth nature of the cost function in relationship to α. For every grid point, representing a value of a_1 and a_2, the likelihood function is optimised with respect to α as depicted in Figure 7-10.

Coarse search

First a coarse search attempts to locate the region where the global optimum is most likely located. Apollo [82] suggests two possible methods for the coarse search. The first possible method uses an exhaustive search grid and the other is to use an ANN which is trained to estimate ML parameters directly for received data. The grid search method was used not only for reasons of speed, simplicity and robustness but also because the inclusion of an ANN stage would require a period of training and the method would no longer be fully blind.

The room model function is symmetrical as exchanging the values of the exponential parameters (a_1 and a_2), and adjusting α accordingly, yields an identical decay curve. This symmetry reduces the search space considerably, and enables the grid to be optimised accordingly. The grid size optimisation entails the removal of redundant gird points, Figure 7-11 describes this process where almost half of the grid points are removed from the exhaustive search. Next, α is optimized at each location using the

same algorithm as the fine search as described in the next section. The log likelihood value is calculated for each location and the maximum located, only the shaded grid positions in Figure 7-11 are computed.

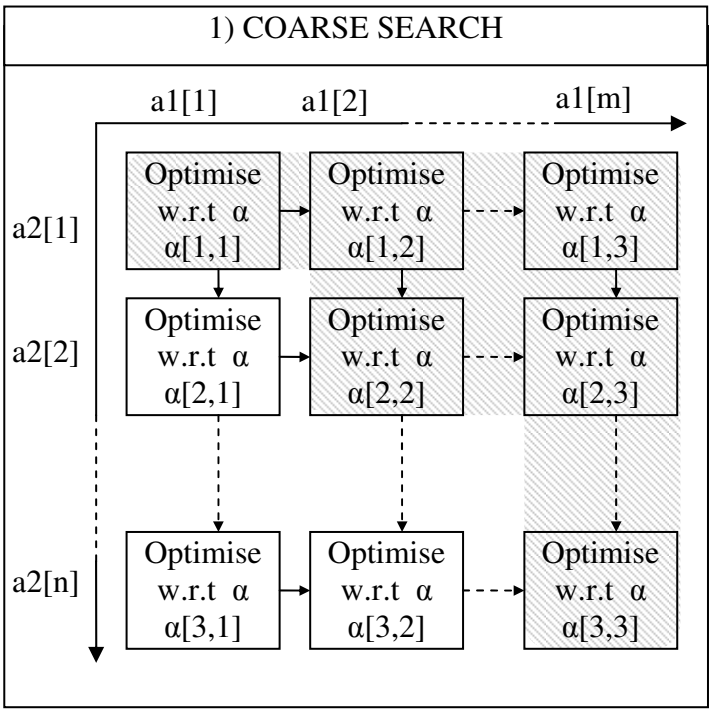


Figure 7-11. Coarse search method. Only the shaded cells are processed due to symmetry, the cost function is exhaustively computed at all highlighted grid points to reduce the chance of missing out on possible solutions in the coarse search. The sequential quadratic programming method [83] is used to minimize with respect to α at each grid position.

In this model, it is assumed that all parameters lie within certain bounds. This reduces the search size but also keeps the model realistic by constraining the range of a_1 and a_2 values. Although the decays at any point in time are a function of both a_1 and a_2, in general if the parameters differ, each parameter will dominate in a different region. The crossover between regions depends on α. Therefore, a_1 and a_2 are limited to rates of decay corresponding to Rt and EDT times between 0.05s and 100s. The reason for such a high upper value is to enable good fits to be made to noisy decay phases which may have very slowly decaying regions due to, for example, a high noise floor. Therefore a_1 and a_2 is limited to; $0.9954 \leq a \leq 0.99995$ for a 3kHz sampling frequency (Fs), where

$$a = e^{\frac{-6.91}{Rt \cdot Fs}} \qquad (7\text{-}29)$$

and α is limited to; $0 \le \alpha \le 1$.

Fine search

The coarse estimates are used as the starting point for the fine search. An optimisation method using non-linear programming known as Sequential Quadratic Programming (SQP) was used. The constrained minimization algorithm used was the MATLAB optimization toolbox implementation of SQP [83]. The algorithm approximates the likelihood function to a quadratic function and the constraints are replaced by linear approximations. A Hessian matrix approximation is computed at each iteration and this is used to compute the search direction for a line search procedure. This closely mimics Newton's method for unconstrained optimization which was used to optimize the likelihood equation in Ratman's method [3].

7.2.6 Model validation

As a preliminary evaluation of the method, the maximum likelihood estimations were performed directly on real and simulated room impulse responses using both the single and multi-exponential models. Results presented in Figure 7-12 show that using multiple exponentials in the estimation model significantly improves accuracy in comparison to single exponential models. The single decay model shows large estimation errors in the EDT. This is because the single exponential model cannot account for non-uniform decays. In general, an over-estimation is seen. The graphs also show the subjective difference limens for the decay parameters as dotted lines. These are used as indicators of the required accuracy of the estimation.

Figure 7-12 a) shows the Rt estimation accuracy to be very good for the 2-exponential model. Figure 7-12 b) shows that the accuracy of the EDT estimation accuracy for the 1-exponential estimation is very poor and while the 2-exponential model EDT accuracy is less than for Rt, it is still close to the difference limens. One reason for this lower accuracy with EDT is the lack of fine detail in the model to enable the proper modelling of strong individual reflections. The fine detail is represented by the function *g[n]* in

the model and over the Rt estimation range, the reflection density is higher which makes it less susceptible to inaccuracies generated by strong individual reflections.

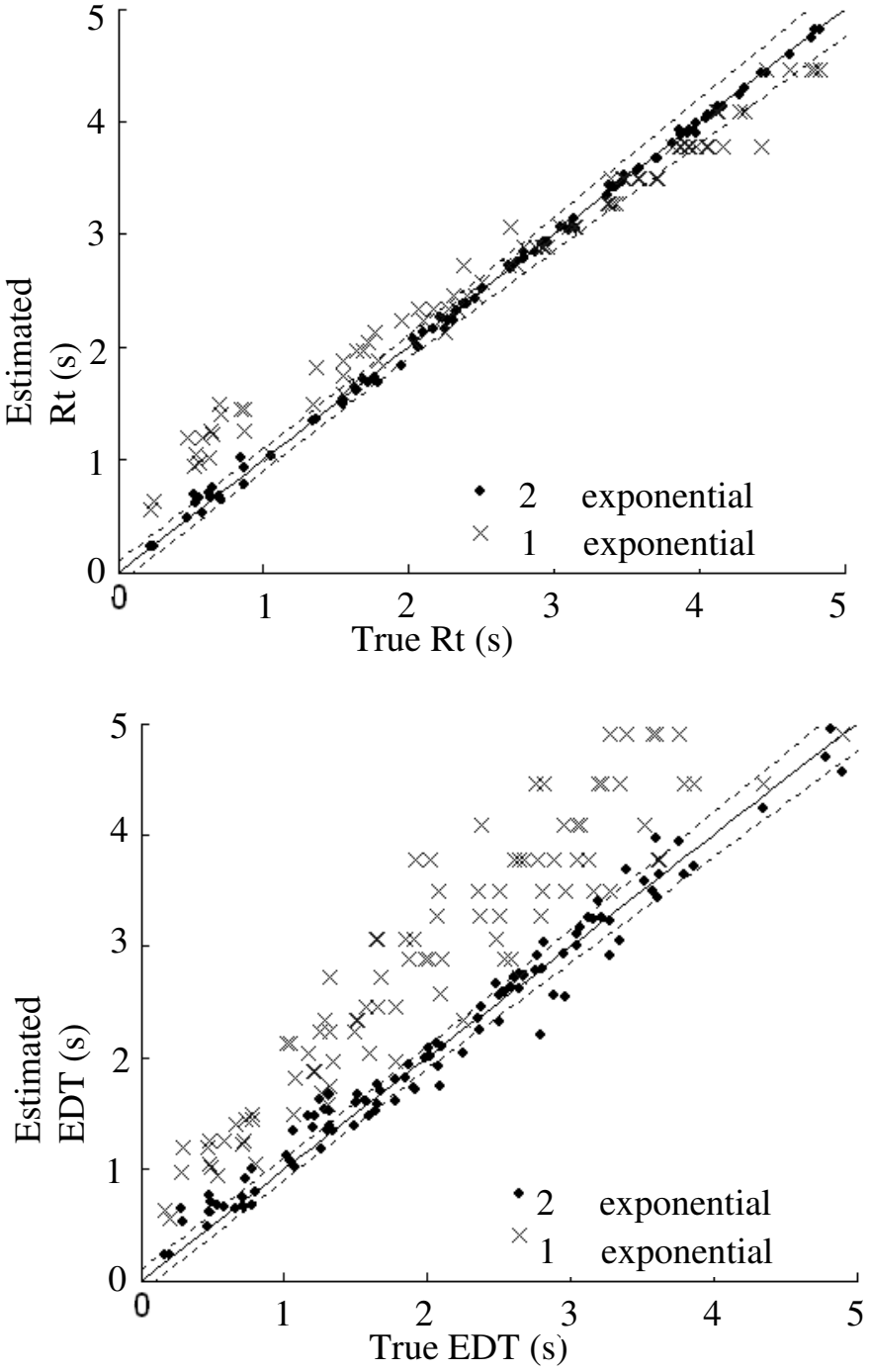


Figure 7-12. Comparison of dual and single exponential model estimations of Rt (a) and EDT (b). Maximum likelihood estimation performed directly on simulated room impulse responses. Dotted lines show perception difference limens.

Figure 7-13. shows the accuracy of the fit to the decay for an artificially generated room impulse response for a single and a dual exponential model. This particular room response has a large discrepancy between early and late decay rate and it shows how the dual-decay model is able to represent such room responses reasonably accurately, whereas the single decay model cannot.

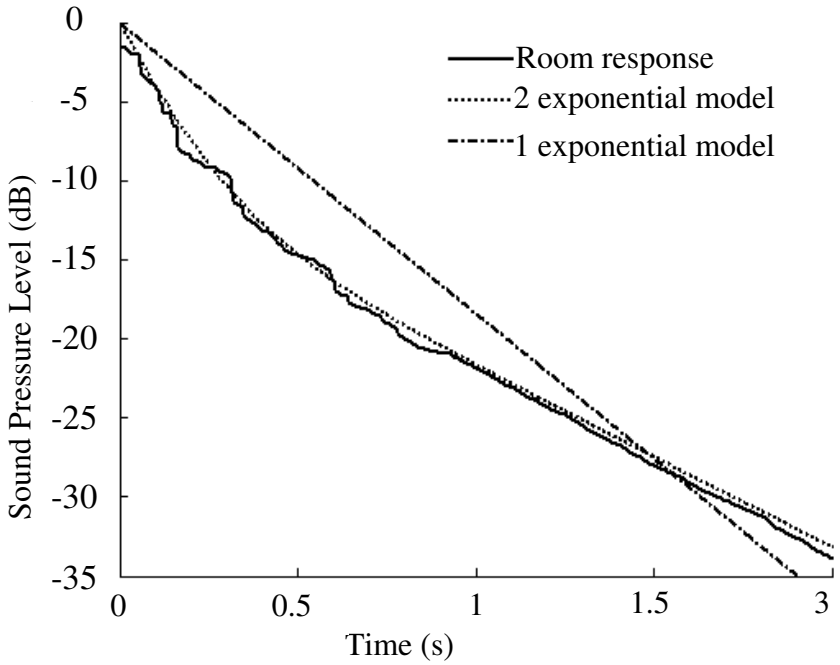


Figure 7-13. Comparison of the estimated decay curves for an artificial room response using the new dual exponential model and the single exponential model.

It was also found that the length of the data used as an input in the ML estimation had a significant effect on EDT and Rt accuracy. When the data length was sufficient to represent up to only 10dB of decay, EDT estimation was very accurate (while Rt estimates were, as to be expected, very poor). As the length is increased to include 35dB of decay, the EDT accuracy decreases slightly while the Rt accuracy increases. Increasing the data length past 35dB has little effect on the Rt estimate but decreases the EDT estimate still further. The reason for this is, for example with EDT, when the data length is increased past the -10dB point, the model complexities stop being used to account for fluctuations in the early decay and are used to model features outside the measurement range, thereby decreasing the quality of the fit to the first 10dB of decay. One way to counteract this problem is to increase the number of exponentials in the

model. This would greatly increase the dimensions of the problem and make optimisation more difficult and time consuming. However, it is expected that the selected decay phases from speech and music will produce a series of decays with a wide range of data lengths and dynamic ranges. Therefore when combining all of the ML estimates, the extra detail lost when performing single ML estimates will be regained as multiple estimates when different data record lengths are recombined by some form of averaging.

Figure 7-14 and Figure 7-15 show the best case estimation accuracy for clarity and centre time (C_{80} and t_s) performed directly on impulses. Accuracy is good but high clarity values are underestimated. This is due to the inability of the model to cope with very high direct to reverberant energy ratios. In these cases the early reflections are dominant so the Gaussian model contained in $g[n]$ is less valid. This problem also occurs when the room response has short centre time and EDT times. The overestimation at low clarity values is not as a result of the maximum likelihood fit. This is due to the finite length of the available data - the impulse response was limited to 2s in this case. When presented with shorter decay phases, this over estimation was not present (this is due to the limited model complexity). Once again, when the method has a large number of decay phases of different lengths from which to gain estimates, this problem of sub-optimal ML fitting will reduce and the parameter accuracy will increase.

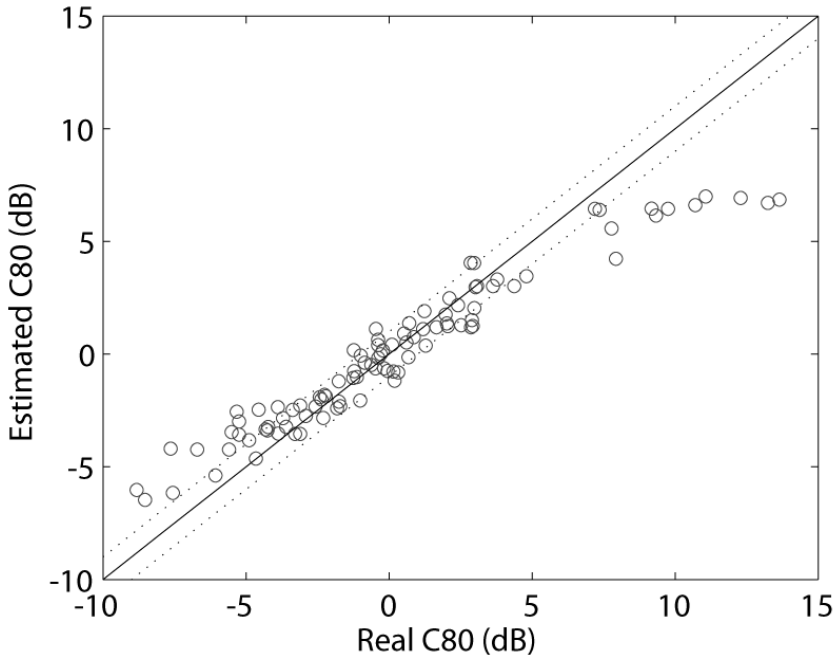
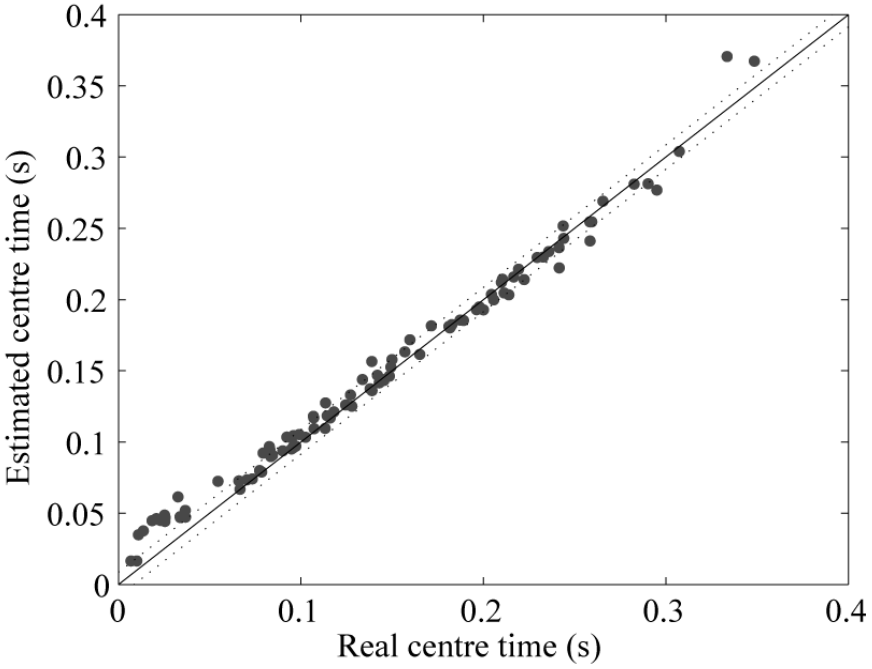


Figure 7-14. Comparison of the clarity estimated directly from impulse responses using MLE (2s length) vs. the true value.

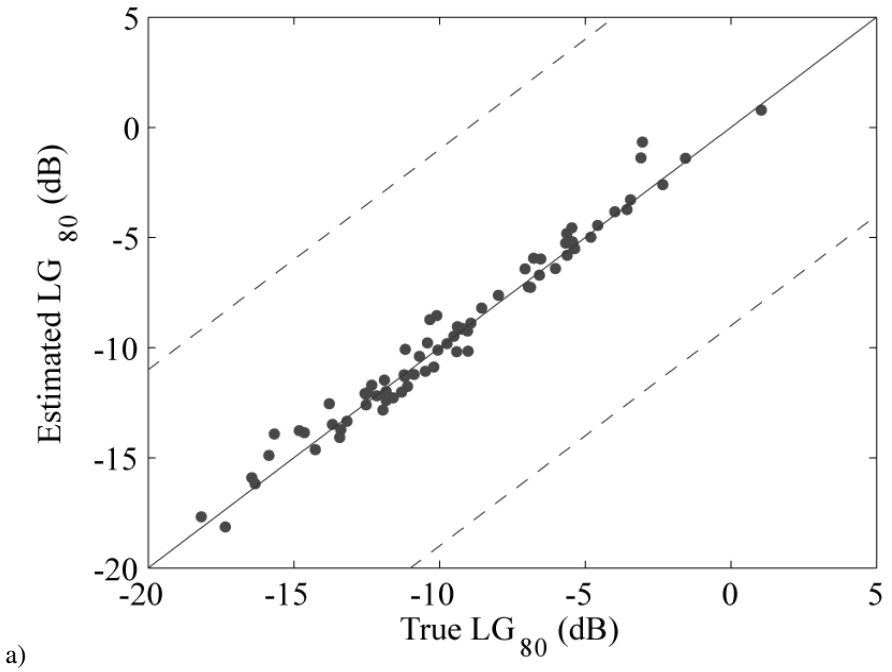


b)

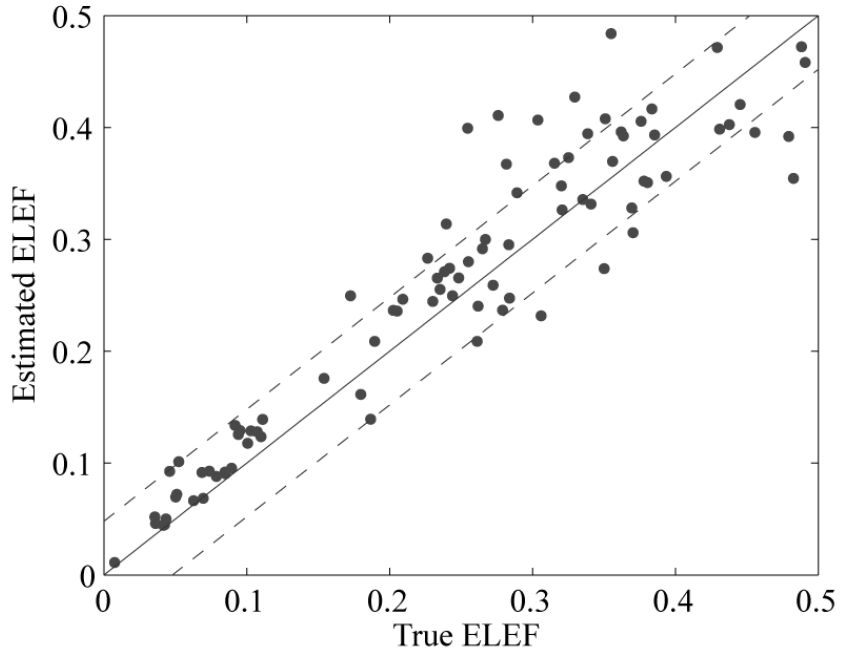
Figure 7-15. Comparison of the centre time estimated directly from impulse responses using MLE (2s length) vs. the true value

In addition to monaural parameters, tests were carried out using binaural parameters relating to spatial impression. The Early Lateral Energy Fraction (ELEF) and the late

lateral strength (LG) are two such parameters. ELEF is a ratio of early energy received laterally to energy received from all directions. LG is a measure of the strength of late lateral reflections. ELEF correlates well with the subjective descriptor, apparent source width (ASW), while LG correlates with the subjective descriptor of envelopment. To calculate these parameters, the room impulse responses are 'measured' in the room acoustic simulation software using omni-microphone and figure-of-8 microphone receiver models. Also required is the source level in anechoic conditions at 10m, this is measured by simulated anechoic conditions in CATT acoustic (see Chapter 4). Using the maximum likelihood estimates of the decay curve (performed directly on the RIRs), estimates for ELEF and LG are computed and compared with the true values yielding the results show in Figure 7-16.



a)



b)

Figure 7-16. Comparison of the estimated parameters a) Late lateral strength and b) early lateral energy fraction. (estimated directly on 2s length of artificial RIRs using omni and figure-of-8 microphones)

Figure 7-16 shows that the maximum likelihood estimation method is very suitable for the estimation of binaural parameters. It is particularly suitable for the estimation of the LG parameter, as it is well within the subjective difference limens. The accuracy is

lower for the early lateral energy fraction. This is because the ML method is less accurate when estimating the early sound field due to inherent limitations in the model. This is due to the inaccurate modelling of the temporal structure of the early reflections. Despite the inaccuracy for the ELEF, it is thought that the accuracy is sufficient to warrant further investigation.

7.3 Pre-processing the signal

Both speech signals and music signals contain suitable periods of free decay for the estimation of reverberation time. These regions are periods of silence between notes or utterances. An automatic algorithm developed by Zhang *et al.* [4] searches for these regions of free decay. This algorithm will be used to achieve the following:

- Reduce the number of decay phases to decrease the computational cost of the maximum likelihood optimisations.

- Yield decay phases of variable length exhibiting continuous regions of free decay. Variable decay phase length is desirable as the length of the gaps in-between utterances/notes is signal dependant and therefore the length must automatically adapt to this.

- Determine the optimum starting point of each decay phase to minimise the decaying tails of musical notes/utterance present in each decay phase.

- Provide estimates of the dynamic range of decay.

This algorithm is described in the following sections.

7.3.1 Envelope segmentation

The envelope segmentation algorithm has five stages:

1) Calculate the Hilbert envelope of reverberant signals, as described in Section 6.1.1.

2) Window the envelope using 0.5s overlapping (99.5%) rectangular windows and perform a least mean square straight line fit to the logarithm of each window and determine the gradient of the line. The long window length (0.5s) ensures that the algorithm will only yield decay phases with reasonable length. The high overlap is to enable accurate locating of the decay phases in time.

3) Remove all windows with a positive gradient as they indicate increasing rather than decaying sound.

4) Find regions in the envelope where consecutive windows have negative gradient and use the location of these regions to select decay phases from the original reverberant signal. Use the locations of the minimum and maximum magnitudes of the chosen decay phases as the start and end points.

5) Estimate dynamic range of decay and remove all estimates with very little dynamic range. It was found that using only decay phases with at least 25dB of dynamic range yields the best results.

The operation of this segmentation algorithm is demonstrated in Figure 7-17. The process of selecting of a single decay phase is presented. The first three plots (a, b and c) show a line fit to three consecutive windows (of the envelope), the final plot (d) shows how this information used to select a single decay phase.

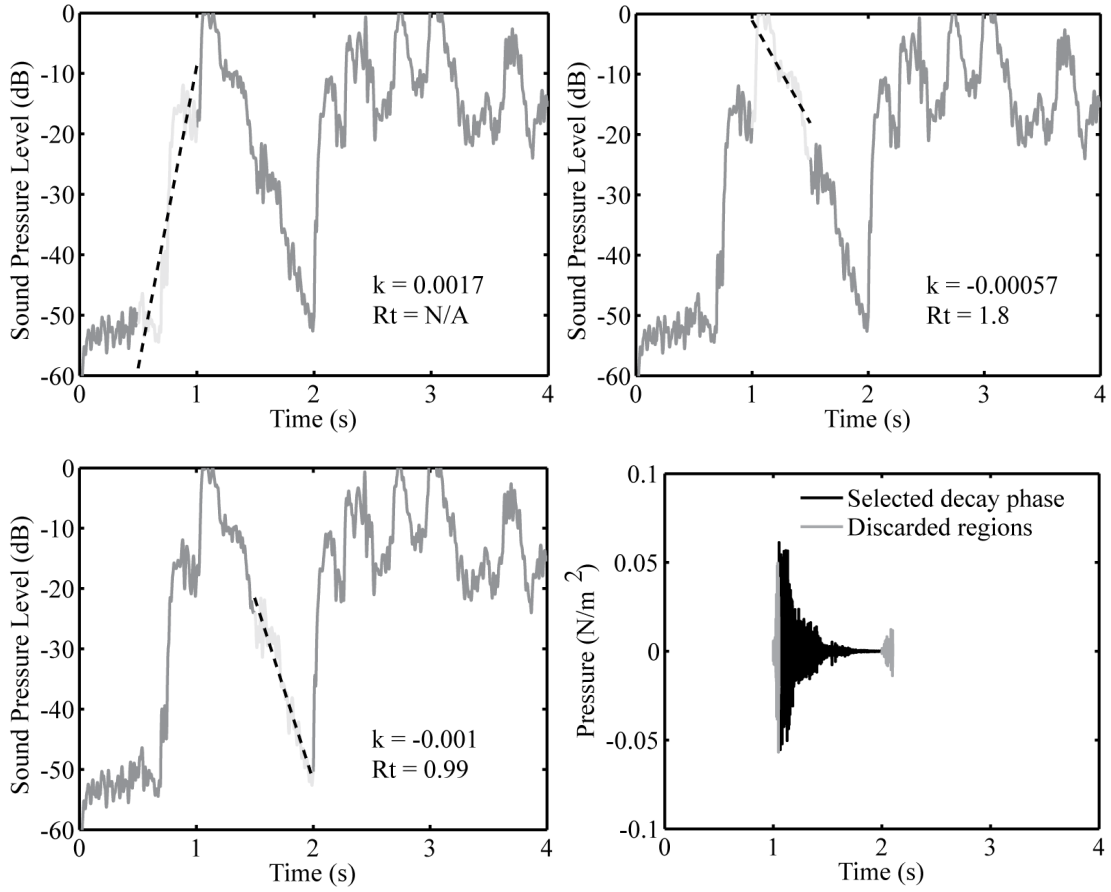


Figure 7-17. Segmentation with least mean square line fitting to identify free decay phases for further MLE. (a) Frame with positive k *indicating rise edge is discarded; (b) & (c) consecutive frames with negative* k *are retained to form the decay phase for MLE. (d) Selected decay phases by the heuristic method. After Kendrick et al.* [9]

The estimate for the dynamic range of each decay phase is gained from the maximum likelihood estimate of the decay curve. This means a computationally intensive optimisation must be carried out on every decay phase. A faster method to compute estimate of the dynamic range using a simple line fitting algorithm was considered, but the estimates were found to be less reliable.

An example of the result of the envelope segmentation method is shown in Figure 7-18, the blue signal is reverberated music and the red signal indicates regions selected by the envelope segmentation algorithm.

The optimum starting point for each decay phase is selected as the maximum absolute level. This is based on the assumption that the direct sound will generally always be at

a greater level than any reflections. This is true for most real rooms although there may be cases, for instance if the direct sound path is obscured, where this might not be true.

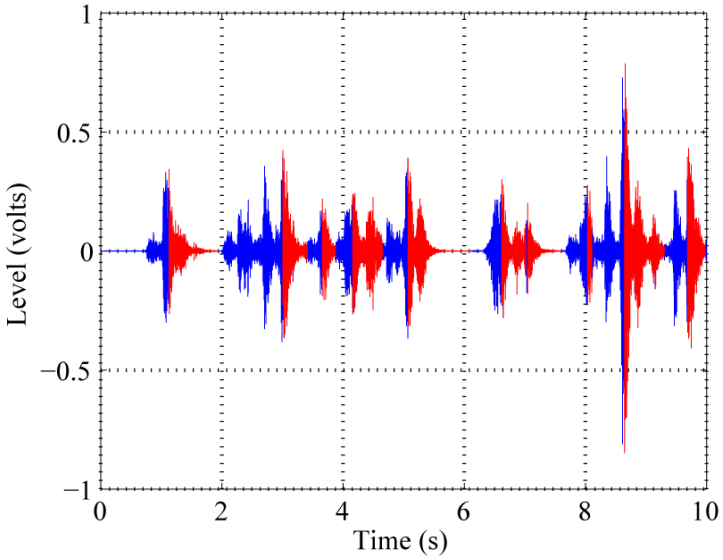


Figure 7-18. Selected decay phases for a 10s segment of reverberated music.

7.3.2 Number of available decay phases

The number of decay phases suitable for use in this ML estimation algorithm is a key factor that can dictate the possible accuracy. This will be investigated in more depth later in the thesis. Please refer to Section 8.3

7.4 Acoustic parameter estimation from ML decay phases

For each of the decay phases, MLE is performed to estimate the decay characteristics. Consequently, from each long segment of speech or music, a number of estimations of the sound decay are obtained. Next, it is necessary to obtain the best estimate of the decay and from it the room acoustic parameters. A simple averaging of the estimates is insufficient, as noise introduced into the estimates by low level reflections or by the natural decay of the source signal utterances or notes, always acts to slow the decay. In other words, the errors that are normally introduced into the measurement are biases and not random errors.

Previously [3] an "order statistics filter" was used to choose the final estimate from the frequency distribution of all the Rt estimations. This was done because signal segmentation was not carried out prior to the maximum likelihood estimation. These decay phases comprise both useful decay phases and regions which decay very slowly due to the presence of the tails of musical notes or utterances. In contrast, the new method developed in the current study employs the segmentation and selection pre-processing stage, described in Section 7.3, to speed processing and reduce problematic post-estimation decision-making. Reducing the number of maximum likelihood estimates also eases the uncertainty in assigning the correct decay properties. Typically, tens of useful decay phases can be found in a 90 second speech signal after the pre-selection, resulting in many sets of decay parameter estimates.

Two criteria are used in the current study to obtain the best estimate of the decay curve. Firstly, decay phases that do not satisfy the required dynamic range are ruled out. Secondly, the decay found in received signals is usually prolonged by any noise sources, such as future utterances or the intrinsic decay of the source, as shown in Figure 7-1, therefore this skewing of the decays must be accounted for. Three methods for selecting the best estimate were tested and evaluated and are listed below:

Method a) Determine the best acoustic parameter estimates by selecting the decay phase with the smallest Rt and extract parameter estimates from this decay phase. Find the decay phase with the smallest Rt from a number of signal segments (at least 1-3 minutes for narrative speech, 10-20 minutes for music) and then average the resulting acoustic parameters.

The decay phase with the smallest Rt may not yield the optimal result for other parameters especially the parameter $C_{80.}$ Therefore methods (**b** and **c)** are developed which attempt to estimate a decay curve that is not just optimal for Rt estimation but for all regions along the decay curve.

Method b) The likelihood function is expanded so rather than estimating one decay curve for each decay phase, a maximum likelihood estimation is performed using a number of decay phases to produce a single decay curve estimate.

Method c) Method b assumes all decay phases are clean and contain no tails of musical notes or utterances. This causes the estimated decay curve to decay slower than the actual decay curve and causes the parameter estimates to be skewed. Therefore a third method is developed that tries to account for this skew. All the decay curve estimates, as generated in method a, are used. The energy level in short overlapping rectangular windows is calculated along each decay curve estimate and for each window the decay curve with the minimum energy level in that window is used to reconstruct that section of the final decay curve. Once again the process is carried out to yield decay curve estimates for a length (segment) of signal (at least 1-3 minutes for narrative speech, 10-20 minutes for music) and then the median of all the decays from each segment is calculated sample by sample.

This next section will introduce and evaluate each of these methods and discuss the advantages and disadvantages of each.

7.4.1 Method a – minimum Rt

The parameters obtained from the maximum likelihood method are those for the envelope of the impulse response rather than the actual decay curve. Substituting these parameters back into the model and performing Schroeder's backward integration [10] yields the estimated decay curve. Reverberation parameters can then be extracted from the decay curve using the standard definitions given in ISO 3382 [8]. A short section (90s) of reverberant speech can yield a large number of decay curve estimates. The frequency distribution of Rt and EDT estimates (histogram) for each decay phase, identified by the envelope segmentation method (Chapter 7.3) in a 90s speech sample convolved with a simulated impulse response and filtered into the 1kHz octave band, is plotted in Figure 7-19.

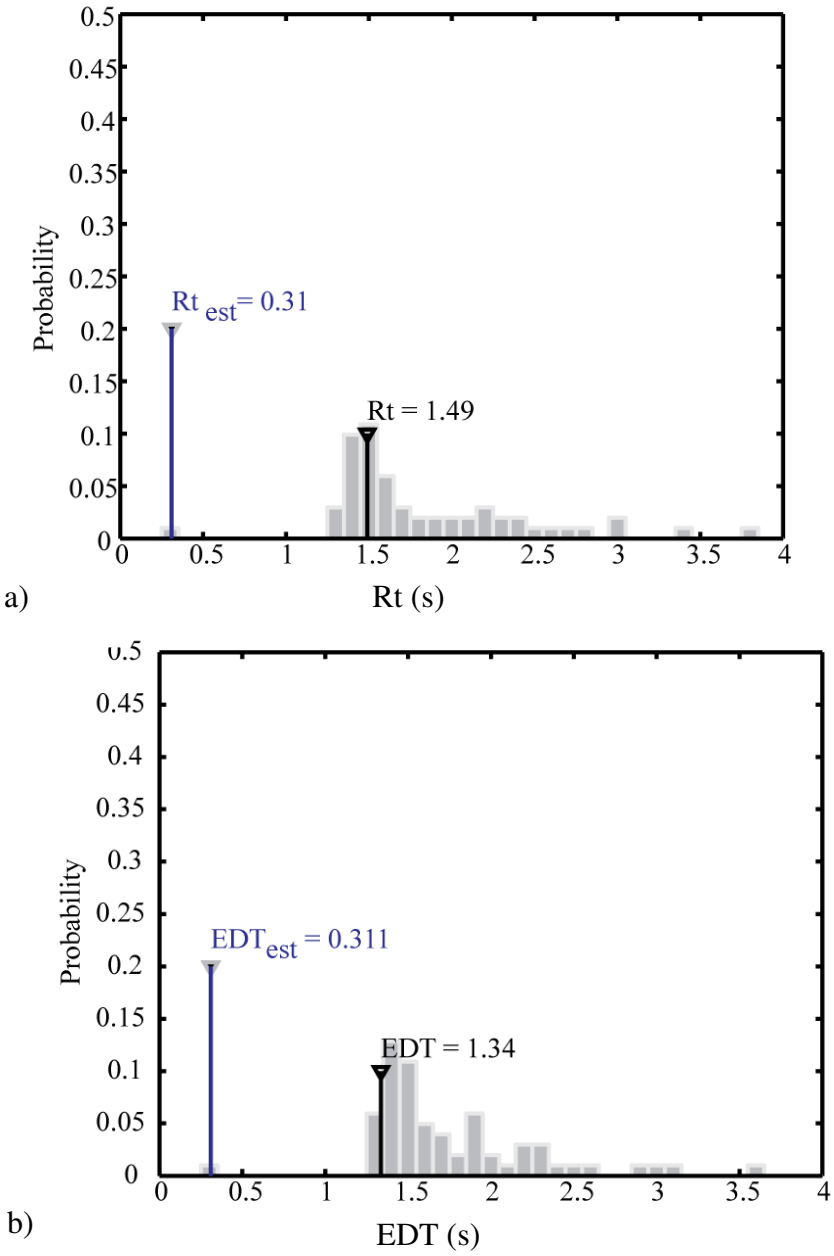


Figure 7-19. Probability distribution of maximum likelihood estimates of (a) Rt and (b) EDT, compared with actual Rt and EDT values. Minimum ML estimates RT_{EST} and EDT_{EST} are show on the plots. 90s speech stimulus for a single artificial room response.

What becomes apparent is that performing the estimation using all available decay phases can cause large underestimations in the decay rate. This is because the decay phase with the lowest decay rate may be very short and not represent a reasonable

length of the decay curve. As previously mentioned the ML estimates provide a reliable estimate for the dynamic range. Utilising this information by performing ML estimation using only decay phases with at least 25 dB of dynamic range, yields the frequency distributions shown in Figure 7-20;

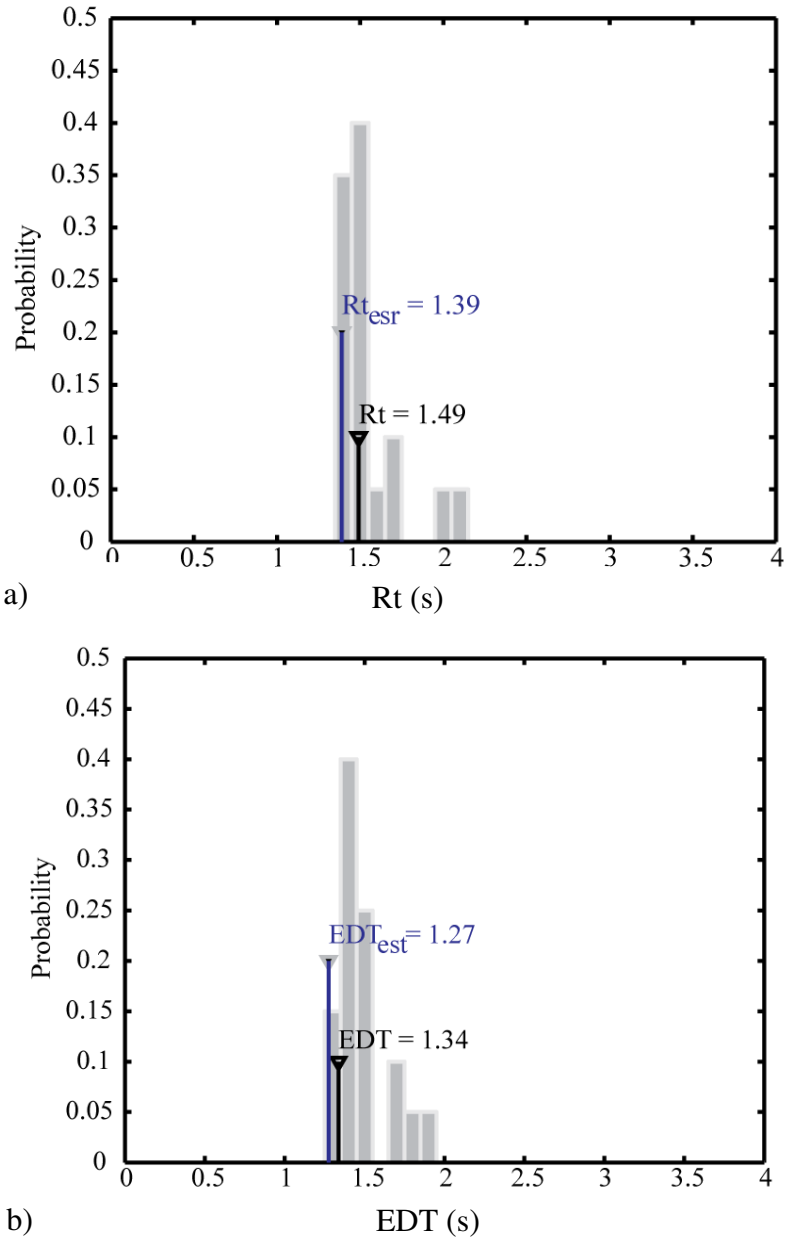
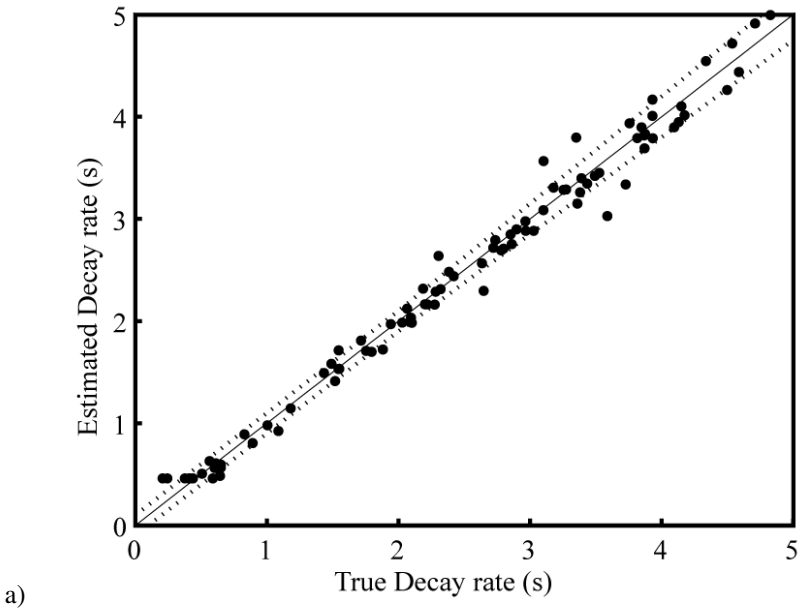


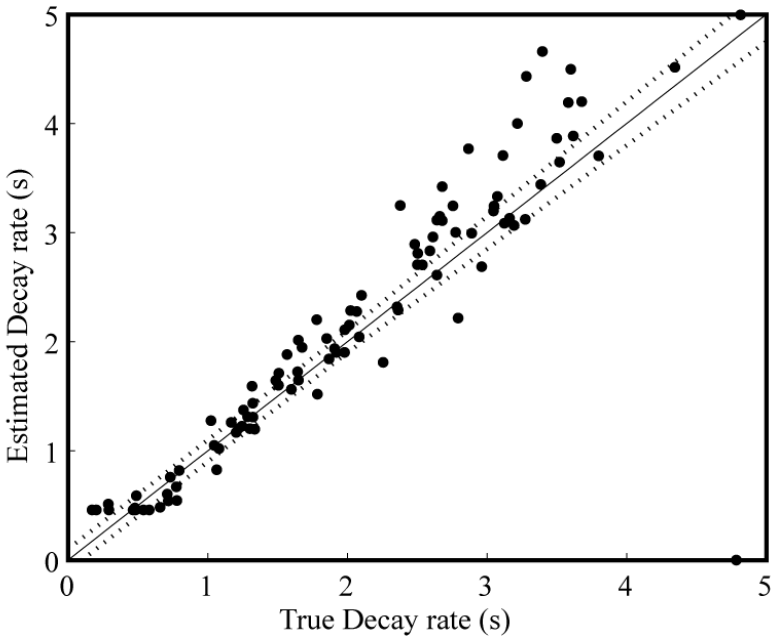
Figure 7-20. Probability distribution of maximum likelihood estimates with >25dB dynamic range of (a) Rt and (b) EDT, compared with actual Rt and EDT values . Minimum ML estimates RT_{EST} and EDT_{EST} are show on the plots. 90s speech stimulus for a single artificial room response.

For comparison purposes the same speech sample was also convolved with a real measured room response and the algorithm run again, a similar result is found for real room impulse responses.

To further investigate the accuracy of the system, a large number of RIRs were convolved with the same 90s speech sample. For each RIR the true acoustic parameter was calculated using the standard method. Each result is filtered into octave bands and these show the results for the 1kHz octave band. Figure 7-21 and Figure 7-22 show the comparison between the estimate and true decay parameters. For the results using simulated impulse responses, in Figure 7-21, the EDT estimation is more inaccurate than the reverberation time. For the real rooms, Figure 7-22, it appears that there is a tendency for slight underestimation for both parameters. The reverberation time estimation is also more reliable than the EDT estimation for real rooms. One reason for this is that speech utterances have a decay rate themselves and there is uncertainty as to where the true beginning of the room impulse response is. Slight underestimations in Rt and EDT can also be due to the overall variability in the dynamic range estimation. Decays with too little dynamic range will almost certainly give under-estimation as the early reflections more often than not decay faster than the late decays. This can be taken into account by performing multiple measurements using this procedure and averaging the resultant parameter values.

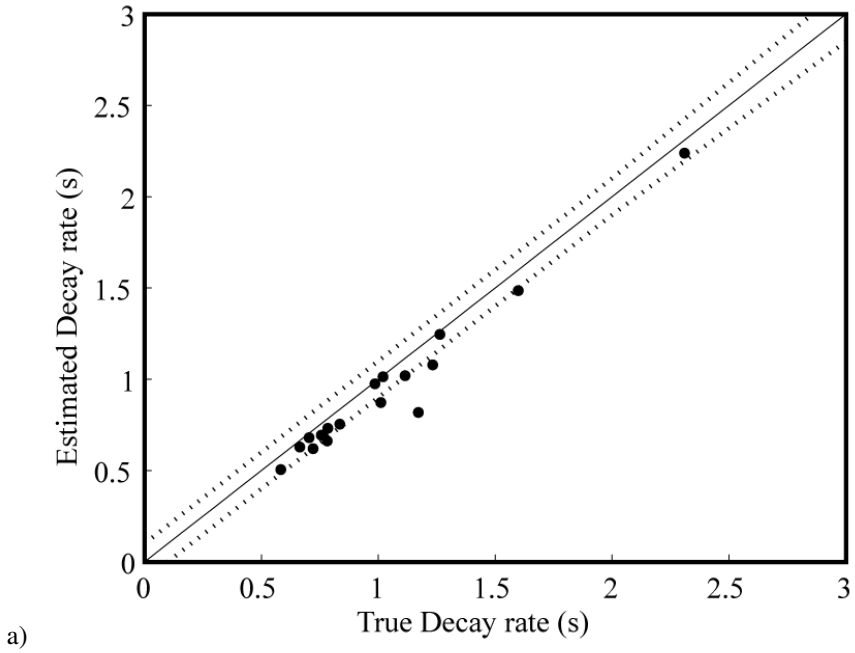


a)

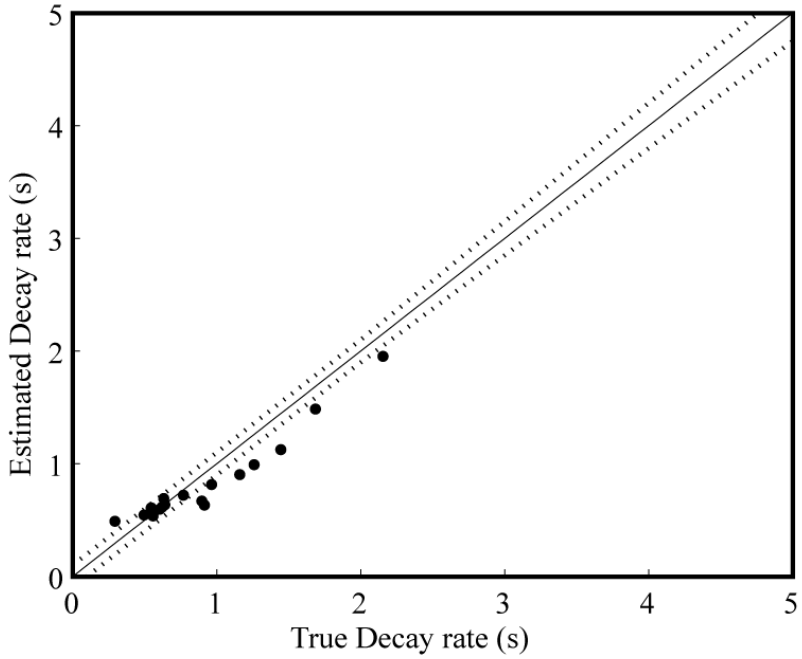


b)

Figure 7-21. Comparison of (a) reverberation time and (b) EDT, estimated using multi-decay maximum likelihood estimation and the true values. Artificially reverberated speech using simulated impulse responses (1 kHz octave band).



a)



b)

Figure 7-22. Comparison of (a) reverberation time and (b) EDT, estimated using multi-decay maximum likelihood estimation and the true value. Speech received in real rooms.

Figure 7-23 shows the estimations of C_{80} and t_s performed using this method. It should be noted that for clarity, rather than using the smallest estimate within a region (as for Rt), the largest value is used.

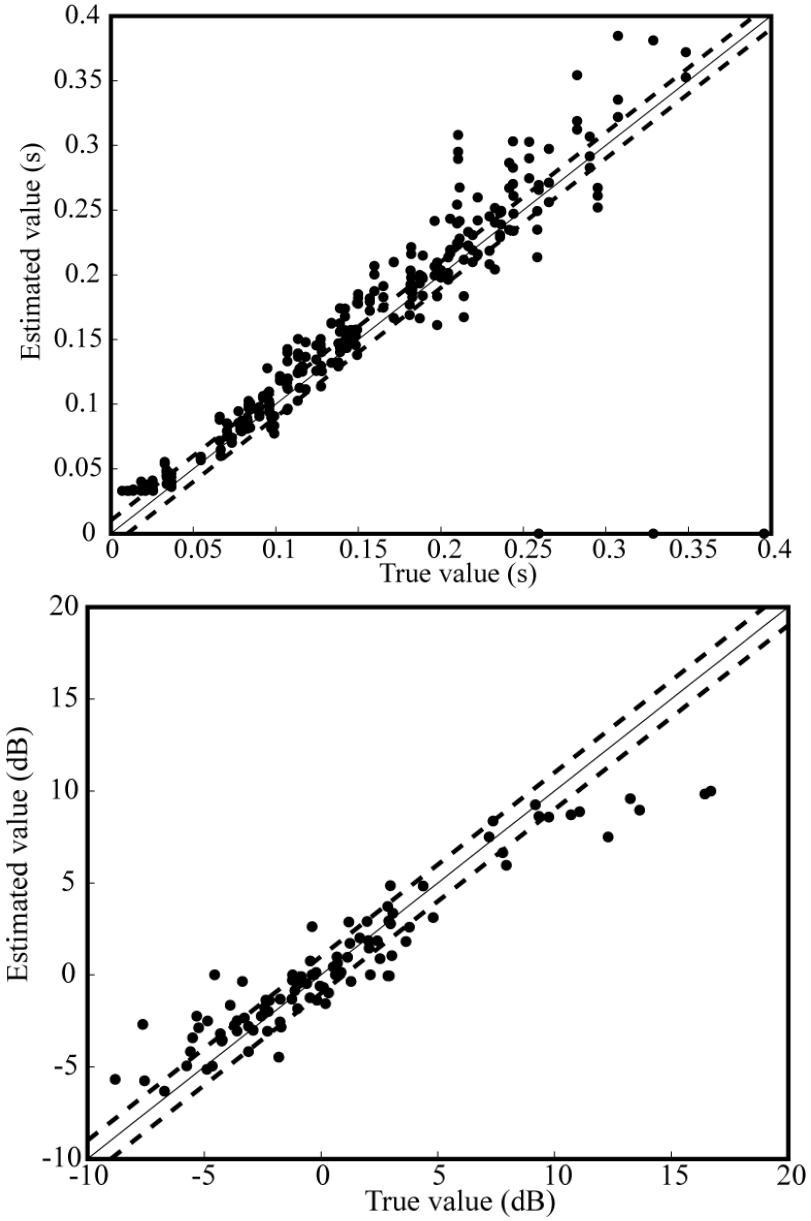


Figure 7-23. Comparison of (a) Centre time and (b) C_{80}, estimated using multi-decay maximum likelihood estimation and the true value. Speech received in simulated rooms.

Results that utilise this method have been published in [9]. It was found that by using a longer speech signal, yielding estimates from multiple segments, then averaging the parameters, the accuracy was improved. By using the median as an estimate of the average rather than the mean this avoided bias in the result from any outliers. It was also found that the distribution of estimates was not normal and using the median in this case is more appropriate.

A disadvantage of this method is that while good accuracy is gained for all of the parameters individually, it is likely that estimates were gleamed from different decay phases for different parameters. This means there is no single optimal decay curve estimate. A single optimal decay curve estimate is desirable, as it will facilitate the estimation of other parameters such as binaural parameters, it may also be helpful in the blind estimation of a room impulse response for auralisation purposes. Therefore methods b and c have been developed to enable the estimation of a single optimal decay curve estimate.

7.4.2 Method b – global ML estimate

It is preferable to yield an estimate from a number of decay phases to improve the accuracy in the face of stochastic variation of the excitation. Therefore a new likelihood function, that is the likelihood that a *single* set of model parameters generated from a *number* of decay phases, is used. The optimisation is performed across a number of decay phases, instead of averaging, which removes the problem of stochastic variability of the decays.

The likelihood (L) of two data sets, each being generated by a set of model parameters, is equal to;

$$L = L_1 (and) L_2 = L_1 L_2 \tag{7-30}$$

This can be extended to multiple (N) data sets:

$$L = \prod_{k=1}^{N} L_k \tag{7-31}$$

In the estimation of the room decay curve, the logarithm of the likelihood is used and as a result the log-likelihood (LL) for a number of decay phases is calculated as the total likelihood for all decay phases $y_k(n)$.

$$LL_T(y_k[n];\alpha,\sigma,a_1,a_2) = \sum_{k=1}^{N} LL_k \qquad (7\text{-}32)$$

At each cell in Figure 7-11**,** the function that is optimised is the sum of all likelihoods for all decay phases. Again it is necessary to consider only decay phases with at least 25dB of dynamic. The results for this method are shown in Figure 7-24 for real rooms and Figure 7-25 for artificial room responses.

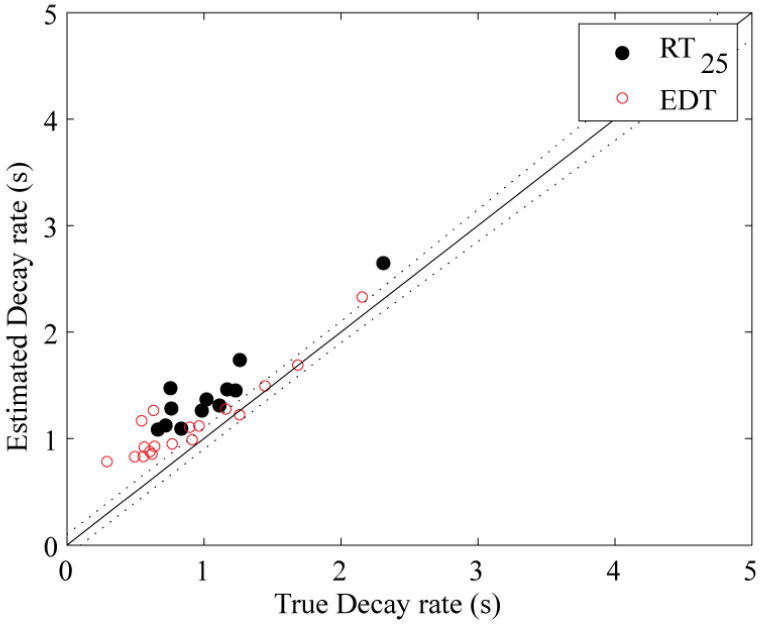


Figure 7-24. Comparison of reverberation time and EDT, estimated using all decay phases with >25dB dynamic range in a single maximum likelihood estimation and the true value. Speech received in real rooms.

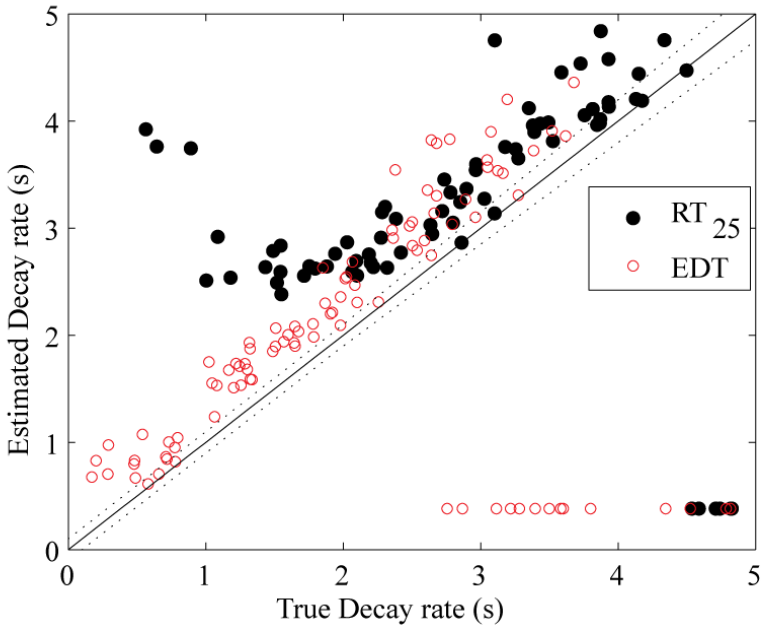


Figure 7-25. Comparison of reverberation time and EDT, estimated using all decay phases with >25dB dynamic range in a single maximum likelihood estimation and the true value. Speech received in artificial rooms.

Figure 7-24 and Figure 7-25 show a tendency for the overestimation of the decay rate, particularly when estimating the reverberation time. This is due to the fact that, while some fluctuation in the decay curves due to the variable signal excitation exists, generally the decay rates will be slow due to the presence of decaying tails of speech utterances. To empirically compensate for this a hybrid of method a and b was developed. The hybrid method selects the fastest decaying decay phases within a 45s window and uses selected decay phases from multiple 45s segments with which to perform the multi-decay maximum likelihood estimation. A longer source signal is required for this operation, so 180s of anechoic speech was used. Optimal decay phases are selected from each segment using method a) i.e. selecting the fastest decay phase with >25dB dynamic range, then all of the selected optimal decay phases from all segments were used in a multi-decay phase estimation (method b). The results for 180s of speech segmented into 4 segments using this method are shown in Figure 7-26.

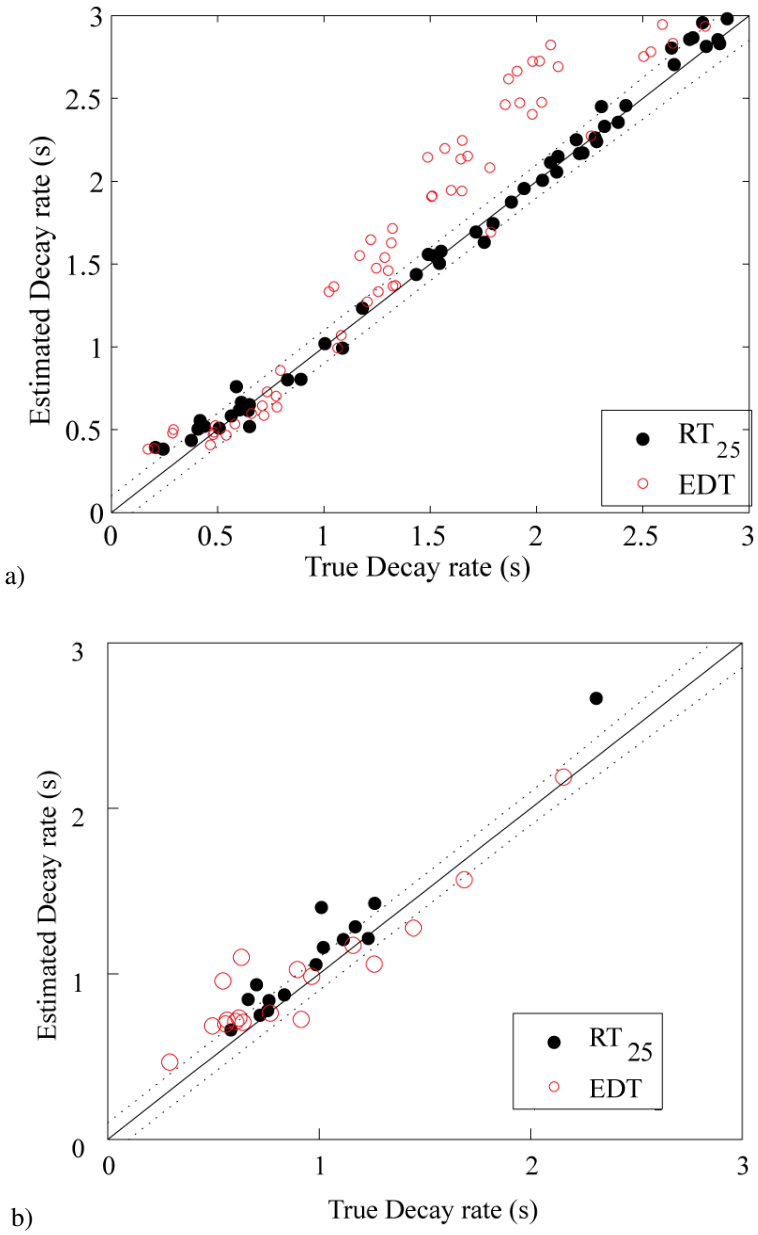


Figure 7-26. Comparison of reverberation time and EDT, estimated using 180s of speech, reverberated and segmented into 4 segments. Fastest decaying reponse with at least 25dB dynamic range selected and 4 decay phases used to estimate final decay curve. a) simulated rooms, b) real rooms

Figure 7-26 shows that an improvement in the accuracy is achieved by using the fastest four decay phases, particularly for the Rt, however over-estimation still prevails especially for the early decay time. This over-estimation is thought to be due to the assumed distribution of decay phases. The maximum likelihood method assumes that the decay phases will demonstrate a Gaussian distribution. In fact, because of the excitation of the signal, the distribution is heavily skewed. This means that the

maximum that is located by the ML method may not be in the correct location, this is analogous to finding the mean, rather than the 50% percentile (median) of a skewed distribution.

7.4.3 Method c – optimal estimation of the decay curve

This method was conceived due to the failings and limitations of the previous two. A method is required that produces a decay curve estimate from a number of ML estimates so that a single optimal decay curve is generated. The averaging of acoustic parameters from different segments of signal, as mentioned in Section 7.4.1 for method a, indicates that a long-term averaging approach is useful in achieving resilience to stochastic variance.

Method a used the shortest Rt as a descriptor of the best segment for estimation, this is problematic when estimating other parameters such as C_{80} and t_s. Which decay phases should be used to estimate these parameters, those with the fastest Rt or EDT? In answering this question it is defined what is required by the new method. The new method needs to:

- Estimate the optimal decay curve (rather than acoustic parameter directly). The optimal parameters can then easily be derived from the decay curve.

- Be rigorous to stochastic variations in signal excitation. Some form of long term averaging is required to ensure the optimal regions of decay are selected while avoiding the problems of stochastic variations in source excitation.

Such a method can be constructed by considering the decay phase energy. The k^{th} estimated decay phase $y_k[n]$ can be written as a function of a number of signal components:

$$y_k[n] = h[n] \otimes \left(\delta_k[n] + d_k[n]\right) + r_k[n] \qquad (7\text{-}33)$$

where $\delta_k[n]+d_k[n]$ represents the direct sound within the time frame. $h[n]$ is the RIR, and $r_k[n]$ represents reflections excited by signals occurring before the start of the selected decay phase. The direct sound is split into two components, $\delta_k[n]$ representing

ideal impulsive excitement at the start of the frame and a competing noise term $d_k[n]$ representing the subsequent decay of the musical note or speech utterance. Note that the decay phases from the ML estimation are normalised to the maximum absolute value and are therefore independent of excitation level. Calculating the signal energy based on Equation *(7-33)* yields a number of squared and cross terms. However, provided the energy is estimated over a sufficiently long time window, it may be assumed that the cross terms reduce to zero as variables within these cross-terms are uncorrelated and have a zero mean value. The signal energy is therefore approximately given by:

$$\sum_{n=0}^{N-1} {y^2}_k[n] = \sum_{n=0}^{N-1} h^2[n] + \sum_{n=0}^{N-1} \left({d^2}_k[n] \otimes h^2[n] \right) + \sum_{n=0}^{N-1} {r_k}^2[n] \tag{7-34}$$

The first term on the RHS of Equation *(7-36)* is constant so the energy in a decay phase only changes with d_k and r_k. Therefore, by finding the decay phase with the smallest energy for a given time range, this minimises d_k and r_k, hence finding the cleanest region of free decay and regions that are most likely due to impulsive excitation. This is an important result as it has implications for the extraction of acoustic parameters. The estimated decay must be treated like an impulse response and Schroeder backwards integration performed to yield the decay curve before yielding estimates for Rt and EDT.

Decay curve reconstruction

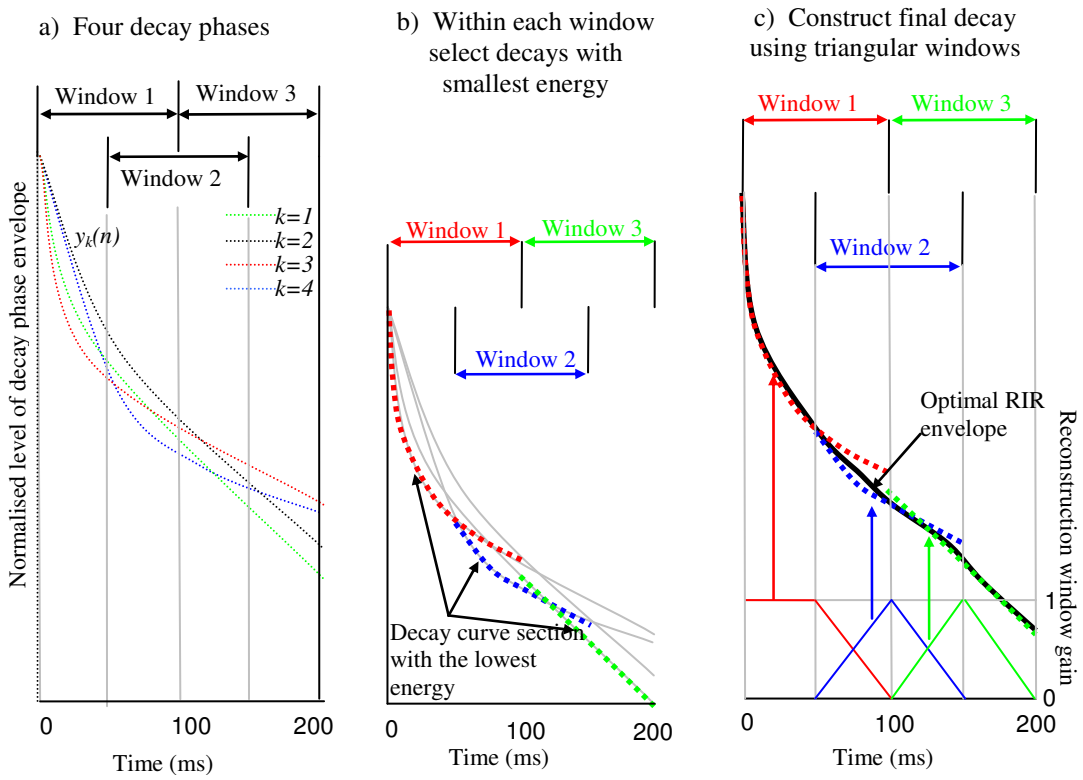


Figure 7-27. Optimal RIR envelope construction from a set of ML estimates. After Kendrick et al. [7].

There may be no single decay phase that is 'cleanest' for the full length of the RIR, so decay phase estimates are windowed using a 0.1s rectangular window using 50% overlap and the energy is calculated for each window (Figure 7-27.a). The decay curve with the least energy in each window is selected as the 'cleanest' portion for that range (Figure 7-27.b). The final decay curve estimate is constructed using triangular windows to smooth discontinuities in the decay curve (Figure 7-27.c). The Rt and EDT parameter accuracy was found to be very similar if not slightly superior to those achieved with method a), but the key result here is that all parameters are now calculated from a single, optimal, decay curve.

Long term averaging of the decay curve

It is assumed that within a certain length of speech or music, there exist sufficiently clean free decay phases that can accurately represent the RIR. Investigation indicates

that for most rooms one minute of a single person speaking or whole movements of orchestral music can be sufficient. Once a decay curve estimate has been produced a number of other factors are considered:

(1) Frequency content variability of sound source;

(2) Time variance of RIR, and

(3) Signal-to-noise ratio.

The limited frequency content of single decay phases (Factor 1) is a known source of error. The frequency content variability of the source excitation is a particular problem when using music signals. This is because the properties deviate quite considerably from broadband excitation. By averaging across a number of decay phases this problem is reduced as different decay phases will excite different frequencies and by averaging across decay phases the resulting estimate can assume a more broadband like excitation.

In real rooms the RIR is not time-invariant (Factor 2), in particular the properties change with temperature. If the rate of change of the RIR is slow enough this leads to the possibility of using the methodology to track the changes. This would work best with speech signals, over long periods many estimates could be made and any trends tracked. However this has not been formally tested.

Factor 3, S/N ratio, is a biasing factor. The presence of background noise causes over-estimation of the decay curve. Outliers in the decay estimation are common, due to noise, therefore the median is used rather than the mean. This operation is detailed in Figure 7-28 illustrating, amongst others, an outlying decay curve and showing that by using the median, the final estimation has not been overly biased by the outlier.

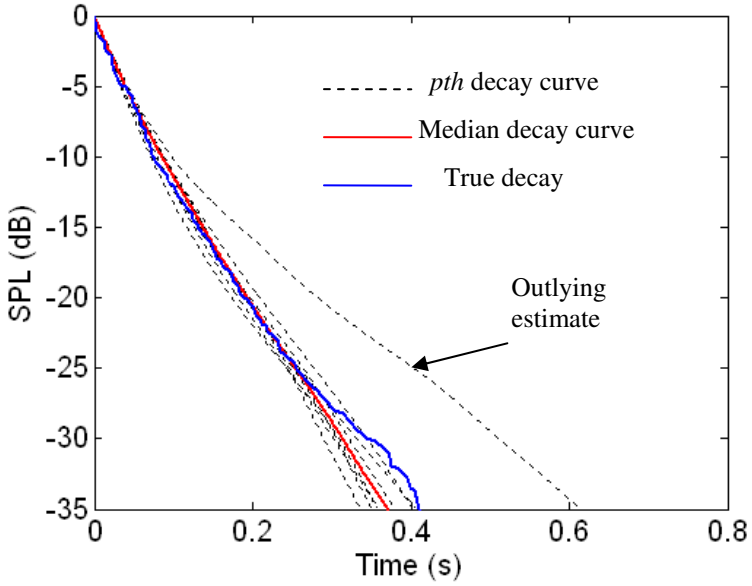


Figure 7-28. Optimal Schroeder curve estimation using eight, 60s segments of anechoic speech convolved with a real RIR. Each decay curve estimate has been backwards integrated to produce a Schroeder curve.[10]

7.5 Disscussion

This chapter has documented improvements to the maximum likelihood method for estimating acoustic parameters. An envelope following pre-processor selects portions of continuous decay and a maximum likelihood fit, using a dual-decay model of sound decay, is performed on each decay phase. The multi-decay model enables estimations to be performed in non-diffuse spaces and coupled rooms as the method yields good fits to non-uniform decays.

Once a large number of ML decay curve estimates are generated from all the selected decay phases, a final estimate needs to be produced for each acoustic parameter. Within this chapter three methods of achieving this were developed, the first simply selects that smallest Rt or EDT parameters from the distribution of estimates with sufficient dynamic range (25dB), this yields reasonable estimation accuracies for most parameters and averaging across many segments improves the robustness. Method b performs a single ML optimisation using all decay phases with sufficient dynamic range (25dB). This method generally resulted in bias errors because the decay phases contain residual

direct sounds and reverberant tails from previous utterances. The method attributes these residual components to reverberant decay and causes the decay rate to be underestimated. Therefore a final method, method c, was developed which also produces a ML likelihood decay curve from which all parameters can be calculated but can better account for the residual direct sound in the decay phases. To achieve this, the minimum energy is found over all estimated decay curves (with >25dB dynamic range) for all points in time. The final method is the most robust and versatile method achieving at least as accurate results as method a), while yielding a single optimal decay curve from which parameters are calculated, this single decay curve may also be useful for other applications (eg. auralisations).

The next chapter will evaluate the performance of method c on music and speech for both monaural and binaural parameters. Results utilising these methods have been published [7, 9, 73, 84].

8 APPLICATIONS OF THE MAXIMUM LIKELIHOOD ESTIMATION METHOD

This chapter describes the application of the ML parameter estimation methods developed in the previous chapter, to several simulations and real life recordings. Both speech and music signals are evaluated as test signals for the determination of room acoustic parameters. On the basis of these results, conclusions are drawn as to the performance of the ML methodology, its ability to cope with noise and the accuracy of the estimated parameters.

8.1 ML Acoustic parameter estimates from speech

In many acoustic spaces where the determination of parameters such as Rt is desirable, speech signals are often present because of the room's inherent purpose. Examples of such spaces include theatres, lecture halls and class rooms. To evaluate the performance of the ML method when using speech, a nine minute recording of anechoic speech (myself reading from an article in the anechoic chamber at The University of Salford) was convolved with a series of room impulse responses. A maximum likelihood estimation of the decay (method c) was performed on these reverberant decays and the result was used to generate a series of acoustic parameter estimates. The signal was first filtered into octave bands using filters conforming to IEC 1260:1995 [85]. Section 8.1.1 presents the results from the 1kHz octave band of parameter estimations of Rt, EDT, C_{80} and t_s. Section 8.1.2 describes the error in the parameter estimation over all octave bands. In discussing these results, the simulated room database is used for validation purposes. The simulated RIR database, due to the wide range of acoustic parameters it exhibits, represents a wider range of possible room responses than the real RIR database. Results from speech convolved with real room impulse responses have also been gained and similar results to the simulated RIR's are observed, these are included in Appendix F.1

8.1.1 ML Parameter estimates from speech convolved with simulated room impulse responses, 1kHz octave band results

9 minutes of reverberated speech was windowed using rectangular windows, with no overlap, into 1½ minutes segments. This produces seven decay curve estimates by calculating the minimum energy decay curve over each segments (method c). Then the median decay curve was calculated from these seven estimates. The optimal segment length for a signal (1 ½ mins in this case) is dependent on how often a gap in the speech occurred suitable to allow a good estimation of the reverberant decay. Ideal decay phases have both a long enough gap to reveal the room decay and are also preceded by an impulsive sound.

The results shown in Figure 8-1 to Figure 8-6 have been generated using speech convolved with 100 simulated impulse responses. The impulse responses were chosen at random from the database of responses, generated as described in Chapter 4. These figures show parameter estimates extracted from reverberated speech signals. These are plotted against actual parameters calculated directly from impulse responses. The figures also show the difference limens for each parameter as dotted lines.

Figure 8-1 shows that excellent accuracy can be achieved over a wide range of Rts. Most of the estimations are within the subjective difference limens. For reverberation times below about 0.4s some over-estimation of the Rt occurs. This happens because the ML method encounters the minimum decay time of the speech utterances, which then places a lower limit on the Rt estimation, below which overestimation is observed. The error also increases at long Rts, exhibiting a slight positive bias. This is because as the Rt increases the number of decay phases providing sufficient dynamic range decreases, making the method more prone to errors as the estimation is calculated from a smaller sample. In this case the overestimation is because the only available decays with sufficient dynamic range contain noise, slowing the rate of decay.

Overall this shows very promising results for reverberation time measurement using speech signals. Appendix F.1 shows the same results for real room impulse responses and these are also in good agreement with the simulated results.

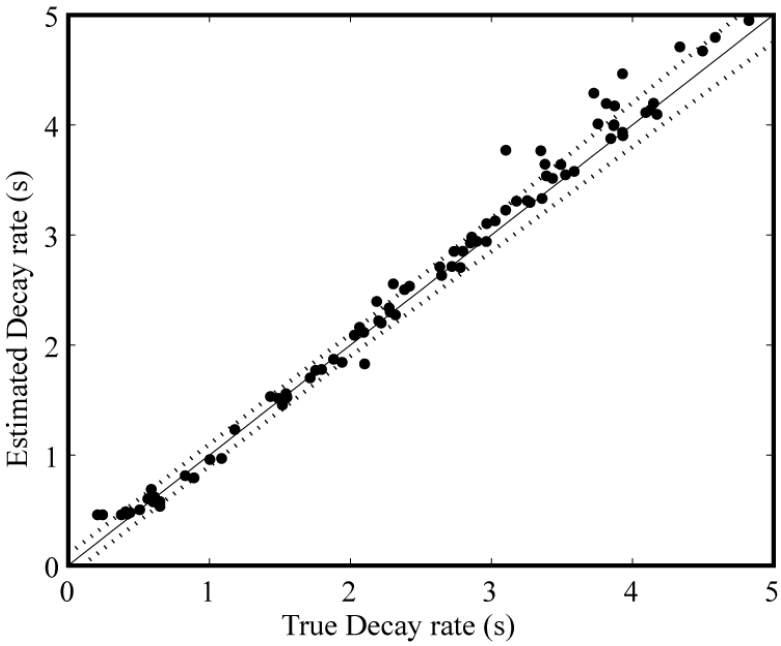


Figure 8-1. Rt estimated using ML method plotted against the true Rt for 100 simulated impulse responses and speech.

Figure 8-2 shows the result for EDT estimated by the ML method, vs the true EDT. The EDT accuracy is less than that for Rt. There are a number of contributing factors to this differing accuracy.

Firstly, referring back to Figure 7-12, this shows that even when performing ML estimations directly on impulse responses, the EDT accuracy is significantly lower than the Rt. This, as previously mentioned, is due to the limited complexity in the model. The model is less valid in the early region due to the non random nature of the early order reflections when compared with the later reflections which are more appropriately modelled as Gaussian noise.

Secondly, it is apparent that for EDT times above 2s there is a tendency for over-estimation. This can be explained by comparing the estimated EDT values with the true Rt values. The EDT estimates are generally found to be somewhere between the true EDT value and the true Rt value. Late reflections may be present in the decay phases due to previous sounds or utterances and if the early decay rate is slower than the later decay, these reflections will mask the early decay rate. It is common for RT to be greater than EDT in many rooms and it is certainly true for many of the simulated RIRs.

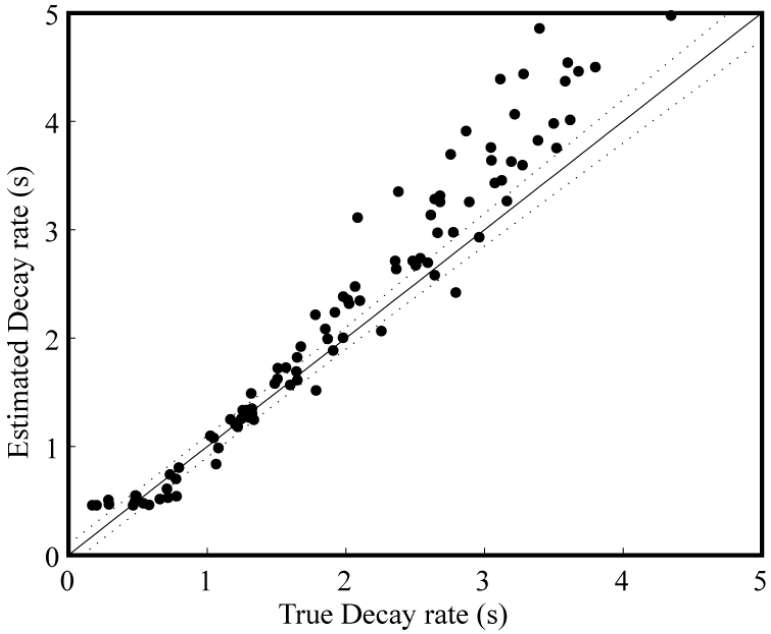


Figure 8-2. EDT estimated using ML method plotted against the true EDT for 100 simulated impulse responses and speech.

Each decay phase is neither the result of impulsive excitation nor a sustained signal being switched off, rather it is somewhere between the two (sustained refers to the signal being present for a period of time so that the level is constant prior to turning the signal off and recording the decay curve – aka interrupted noise method). The algorithm used (method c – Section 7.4.3) attempts to search for decay phases resulting from impulsive-like excitation rather than a sustained excitation. The accuracy of the recorded decay curve is affected by the availability of impulsive-like excitations and on how different the early and late decays are.

Figure 8-3 compares decay curves from impulsive and sustained excitation. Figure 8-3 shows that the early part of the decay from the sustained excitation does not decay as quickly as when excited impulsively, this is because slow late reverberation is masking the quickly decaying early reverberation. As in most RIRs, the late reverberation generally decays more slowly than the early which causes the EDT to be skewed towards the Rt value. As all the decay phases are the result of excitation that is in-between impulsive and sustained, this bias is prevalent especially when the decay is non-uniform.

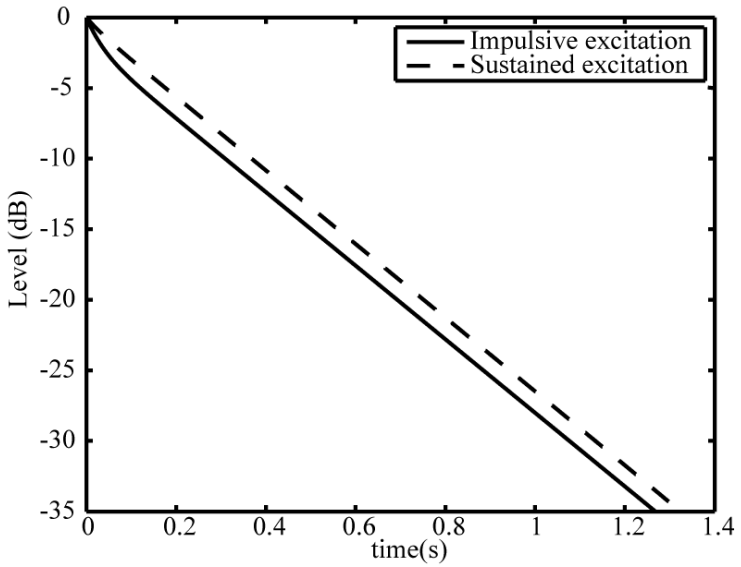


Figure 8-3. Comparison of the envelope of the reverberant decay when the RIR is excited using impulsive excitation and when the response is a sustained noise source switched off after a long period of time.

The algorithm automatically searches for the 'cleanest' decay phases, however, the lack of at least one suitable decay per segment of signal, will limit the overall accuracy of the method. A longer segment length can help improve the accuracy as it is more likely that the algorithm will find the cleanest decays. Figure 8-4 shows the EDT estimation using longer signal segments (3 minutes). The over-estimation at longer decay times has been reduced but the overall accuracy is roughly two difference limens. This accuracy can be increased by using longer recordings and increasing the number of segments.

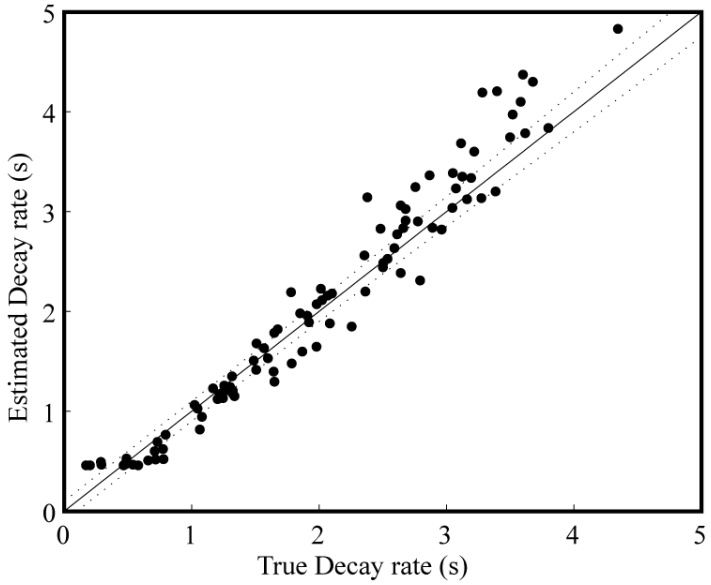


Figure 8-4. EDT estimated using ML method plotted against the true EDT. 100 RIRs were chosen at random from a database. Each impulse response was convolved with 9 minutes of speech. For the ML estimation, the reverberated signal was windowed into three, three minute segments.

Figure 8-5 and Figure 8-6 compare the true and estimated values of clarity (C_{80}) and centre time (t_s). The estimated values were obtained from reverberated speech signals, but using three minute sections of speech (rather than 1 ½ minute segments). These results show similar trends as seen for EDT and Rt. When C_{80} is large, i.e. in spaces with low reverberation times or high direct to reverberant ratio, the values are underestimated. Once again, the natural decay of the speech utterances prevents accurate estimation when the room decays are short. Centre time estimates are accurate above 0.03s, the overestimation at these low t_s is due to the rate of reverberant decay being less than or comparable to the fastest decaying speech utterance. The clarity appears to have a trend for overestimation, this is due to the ambiguity between late and early decay rate, where the late reverberation masks the early reflections. This causes an increase in the energy in the first 80ms when compared with the later energy and therefore an increase in the clarity index.

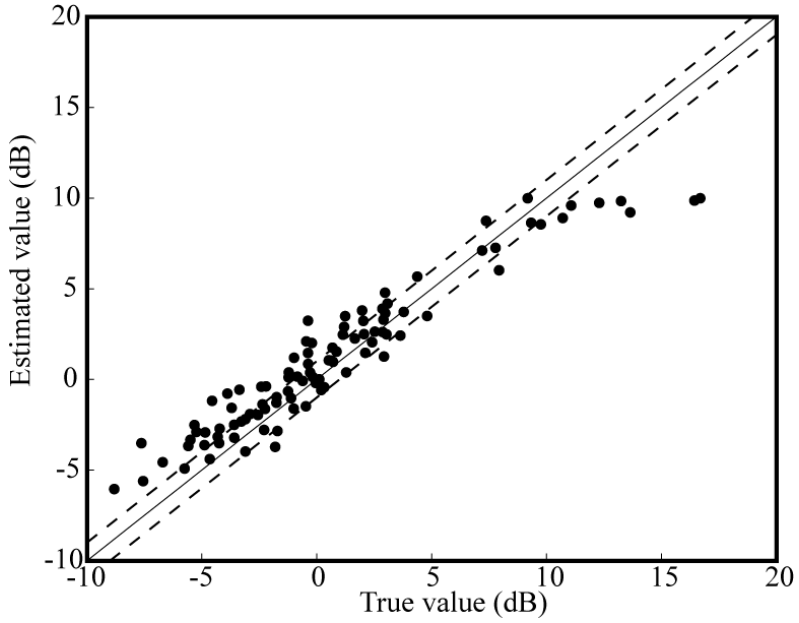


Figure 8-5. C_{80} estimated using ML method plotted against the true C_{80}. 100 impulse responses were chosen at random from a database. Each impulse response was convolved with 9 minutes of speech. For the ML estimation, the reverberated signal was windowed into three, 3 minute segments.

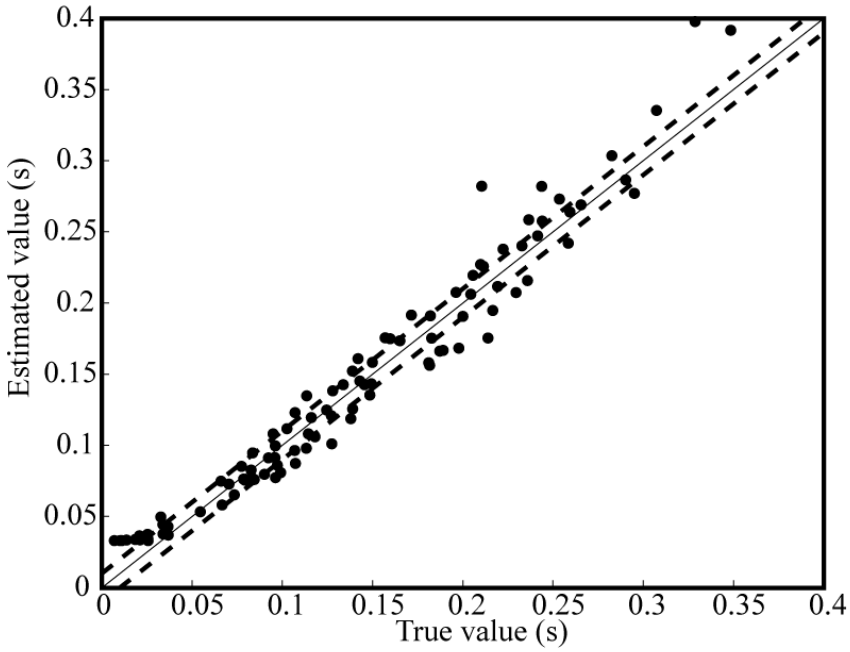


Figure 8-6. t_s estimated using ML method plotted against the true t_s. 100 impulse responses were chosen at random from a database. Each impulse response was convolved with 9 minutes of speech. For the ML estimation, the reverberated signal was windowed into three, 3 minute segments.

8.1.2 Results from all octave bands

In addition to the results from the 1kHz octave band results presented in Section 8.1.1, the method's performance was evaluated in other octaves bands (centre frequencies from 63 to 8000 Hz). For the lower octave bands (≤500 Hz), it was found that keeping the sampling frequency at 3 kHz was beneficial to the maximum likelihood algorithm Although in the lower octave bands a lower sampling frequency increases computational efficiency, it was found that by having fewer samples, the ML estimation accuracy decreased. Because the ML method uses a statistical room model more data equates to a better fit.

For the higher octave bands, sampling rates were chosen to be sufficiently above the upper cut-off frequency of the octave band filter so that aliasing artefacts were not problematic, see Table 8-1. This had to be balanced against the large increase in computation time that a higher sampling frequency caused for the algorithm. For these reasons, the following sampling frequencies for each octave band were chosen.

Octave Band Centre frequency (Hz)	Sampling frequency (Fs)
63	3000
125	3000
250	3000
500	3000
1000	3000
2000	6000
4000	12000
8000	24000

Table 8-1. Sampling frequencies chosen for each octave band

Due to the large number of results, the performance data are presented in a series of tables and plots showing error performance of each of the parameters. Two indications of accuracy are used, the percentage of estimates within the difference limens and the average absolute error. 100 artificial room impulse responses (Table 8-2, Table 8-3 and Figure 8-7) and 18 real room impulse (Appendix Table F-3 and Table F-4) responses were used. The trends for the artificial and real room data are similar, and so only the simulations are presented in the main part of the thesis because they cover a wider range of acoustic parameters. Results from the real rooms (Appendix F.1) appear initially to

produce a lower error than the simulated rooms, however, this is because the real room impulses represent a small cross section of possible responses (i.e. much shorter Rts) and when comparing like with like, the results are similar.

	Octave band (Hz)							
Parameter	**63**	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt (%)	15	25	47	71	79	72	85	93
EDT (%)	3	16	37	35	51	55	55	67
C_{80} (%)	24	33	40	37	56	48	66	41
Ts (%)	9	24	29	47	48	49	65	69

Table 8-2. Summary of the percentage of the predicted room impulse response parameters that are within one difference limens. Estimates are based on an ML analysis of simulated room responses convolved with 9 mins of anechoic speech split into 3 segments. Results are shown for all octave bands.

	Octave band (Hz)							
Parameter	**63**	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(±s)	1.3	0.8	0.4	0.3	0.3	0.2	0.2	0.2
EDT(±s)	1.1	0.8	0.6	0.6	0.4	0.3	0.3	0.3
C_{80}(±dB)	6.2	4.5	3.8	5.2	5.4	4.5	3.1	3.6
Ts(±s)	0.08	0.05	0.05	0.03	0.03	0.03	0.02	0.02

Table 8-3. Summary of average absolute error in the predicted room impulse response parameters. Estimates are based on an ML analysis of simulated room responses convolved with 9 mins of anechoic speech split into 3 segments. Results are shown for all octave bands.

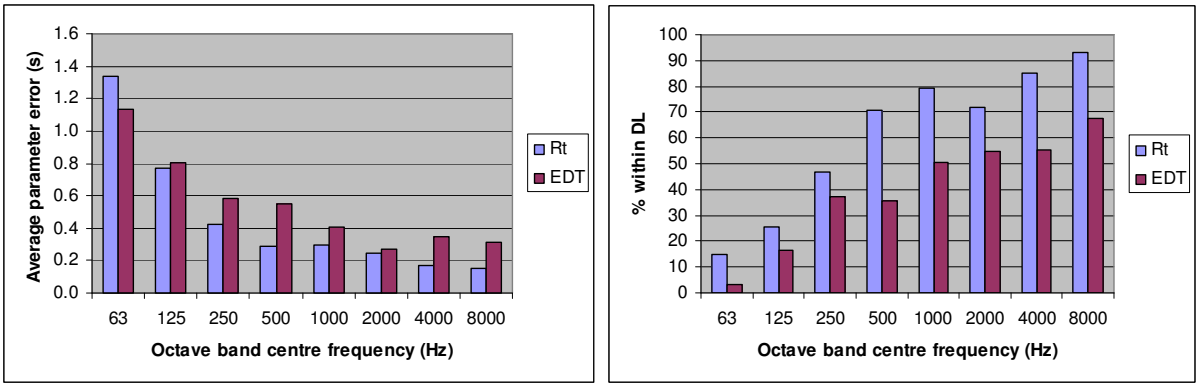


Figure 8-7. The average error for Rt and EDT as a function of octave band (left plot) and the percentage of Rt and EDT predicted parameters within the subjective DL (right plot) using simulated impulse responses convolved with 9 mins of anechoic speech windowed into three, 3 minute segments.

In Figure 8-7, the Rt results show low average error in the bands 250Hz to 8000Hz, where the average parameter error is between ±0.4s and ±0.2s. When Rt was estimated in the octave bands 500Hz to 8000 Hz the percentage within the subjective DL varied between 71 to 93 %, however, in the 250Hz band, the fraction within the DL is 47%. This indicates that speech can be used to reliably estimate Rt in the octave bands between 500Hz and 8000Hz, and additionally, fairly reasonable results can also be achieved at 250Hz. The decrease in accuracy at lower frequencies is due to a number of factors. Firstly, the ML model fits are generally found to be quite poor because at lower frequencies the excitation signal becomes less noise like, due to the harmonic nature of speech signals. At lower frequencies excitation can often be from single frequencies at (the fundamental harmonic of the speech), the resultant decay curve is not representative of the octave band response. Secondly, the reverberation times are generally much longer at lower frequencies due to the lower amount of absorption and the ML method is less accurate generally at longer Rts. The opposite effect is seen at higher frequencies as the Rt times are generally much shorter, on closer inspection; part of the reason for the good accuracy at high frequencies, is because most of the decay rates are generally faster. It should also be noted that the simulated room model is not realistic at lower frequencies, i.e. in the 63Hz-125Hz bands.

When using speech, the accuracy of the EDT estimation is not as good as for Rt. However, the method provides a useful approximate EDT estimate. Reasonable estimates are achieved in the bands from 1000Hz to 8000Hz with an accuracy of ±0.3s or ±0.4s or a fraction within the DL between 51 and 67 %. However, in the octave bands below 1000Hz the errors become too large. Results for clarity and centre time are similar in terms of % within the DL to EDT.

8.2 ML parameters from music

The ML method has been shown to perform well using speech signals, thus its application to music signals is now explored. A number of anechoic orchestral music pieces have been placed in sequence, creating a 40 minute signal. The signal is convolved with real and artificial room impulse responses and the parameter estimates within the 1 kHz octave band are presented in this section and the real room results are

in Appendix F.1. The estimation errors for all parameters in each octave band are presented in Section 8.2.2.

The music used in this evaluation of the ML estimation method was anechoic orchestral recordings from [72] and [86]. All pieces used in this evaluation were between twp and five minutes long; with varying styles, speed and number of instruments, including full orchestral and chamber music arrangements. For more information please refer to Appendix D.

8.2.1 ML parameter estimates from music convolved with simulated room impulse responses, 1 kHz octave band results

A music signal is convolved with the database of 100 simulated room impulse responses. The ML estimation algorithm (method c) is applied to each case and acoustic parameters estimated. As observed with speech signals, the real room responses produced similar results when compared with similar artificial responses, but as the simulated RIR database covers a wider range of acoustic parameters only the simulated results are presented here, the real room results are presented in Appendix F.3.

The resulting estimation accuracy for Rt, using eight, five minute segments of music, is depicted in Figure 8-8. These results show comparable accuracy to those achieved using speech with a tendency to underestimate in a small number of cases. As for speech, at low Rts there is a small tendency for overestimation. This is because the rate of decay of the reverberation is starting to become comparable to the release of the musical note (the rate of decay of the anechoic note).

The reason for the underestimation of a number of the results when compared with speech is due to the non-broadband-like excitation of the music. Music only excites certain portions of the frequency band, depending on the particular notes being played. Due to this increased stochastic variability in the excitation, the resulting ML estimated decay phases also have greater stochastic variability compared with those estimated from speech. The ML algorithm chooses the fastest decaying phases and therefore selects estimates from the lower bound of this variability; this is the cause of the underestimation bias. This phenomenon was also encountered when the music

spectrum filter was applied to the simulated RIRs in Section 6.8.2, but the error is more pronounced here as the decay phases are often the result of single note excitation and thus the spectrum deviates even further from the desired broadband response.

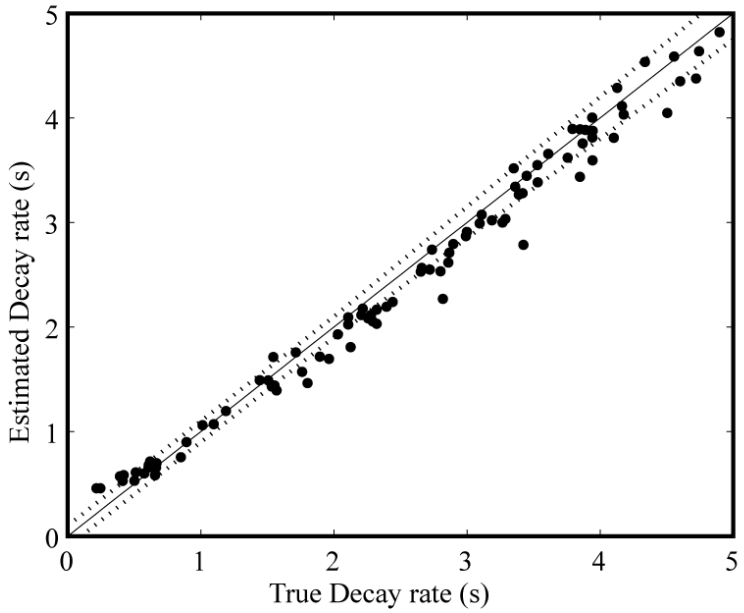


Figure 8-8. Comparison of estimated and true reverberation time. Estimates were obtained from the application of the ML method to simulated impulse responses convolved with 40 mins of anechoic music windowed into eight, five minute segments.

By decreasing the segment length (and therefore increasing the number of averages that are used to calculate the decay curve), the number of selected decay phases from which the median is estimated is increased. Assuming that the spectrum of each decay phase varies randomly, by computing the response using more decay phases, the resulting response is more representative of a broadband response and the tendency for underestimation is reduced. This is shown in Figure 8-9 where 20, two minute segments are used. Figure 8-9 also highlights another problem, the shorter segment length increases the number of averages when computing the decay estimate but there is less constraint on the algorithm to select the cleanest possible decay. Therefore the over-estimation at short Rts has been increased as more sub-optimal decay phases containing significant residual tails of musical notes are used in the averaging. This overestimation is even more problematic for the other early reflection based parameters (EDT C_{80} etc.) as the increased residual tails of musical notes bias the early part of the ML estimates and these parameters are particularly sensitive to errors in this region.

The solution to this problem is to increase the overall length of the recording to increase the number of averages, or to use music which has more pauses after short notes suitable for ML estimation.

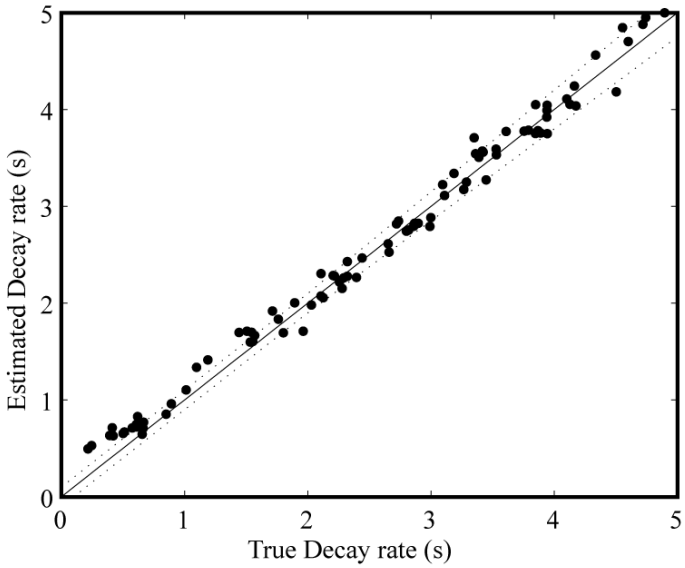


Figure 8-9. Comparison of estimated and true reverberation time. Estimates were obtained from the application of the ML method to simulated impulse responses convolved with 40 mins of anechoic music windowed into 20, two minute segments, results presented.

Figure 8-10 shows the EDT results. As can be seen there is a tendency for an over-estimation of EDT, due to two factors: The decay of the musical notes being included with the decay of the room in the ML estimation, and the positive bias that is introduced when the excitation is not impulsive (as explained for speech signals). Once again the EDT accuracy is below that for Rt, again this is because the EDT is very sensitive to changes in the early part of the response and the early part is more heavily influenced by persisting tails of musical notes and the ambiguity between impulsive and sustained excitation.

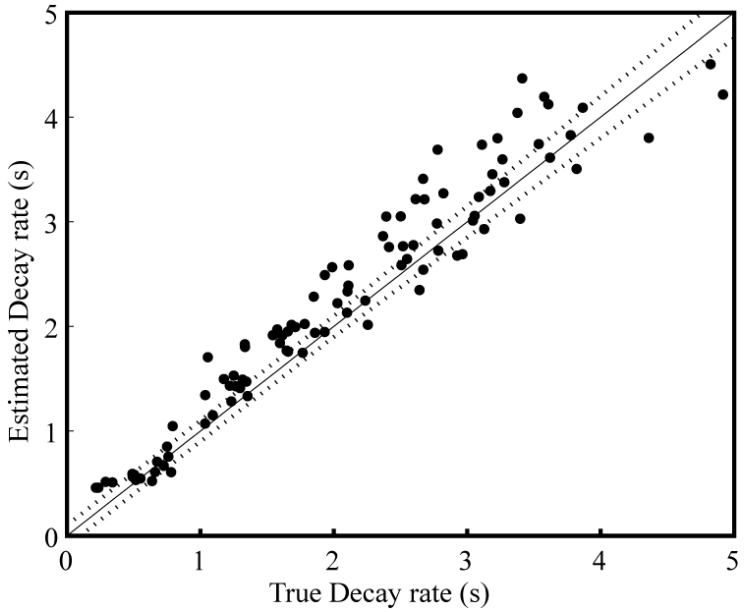


Figure 8-10. Comparison of estimated and true EDT. Estimates were obtained from the application of the MP method to simulated impulse responses convolved with 40 mins of anechoic music windowed into eight, five minute segments.

Section 8.1 demonstrated that, for speech, increasing the segment length improves the prospects of finding a clean decay phase. A disadvantage of the longer segment length is that with a fixed length of recorded audio, the number of averages is decreases. Figure 8-11 shows that, by increasing the segment length to 10 minutes (thereby increasing the chance of finding 'cleaner' decay phases) the over-estimation trend is removed. Although the overestimation trend is removed, Figure 8-11 shows a significant variation in the parameter estimates, which is due to the small number of decay curves used in computing the median (only four segments!).

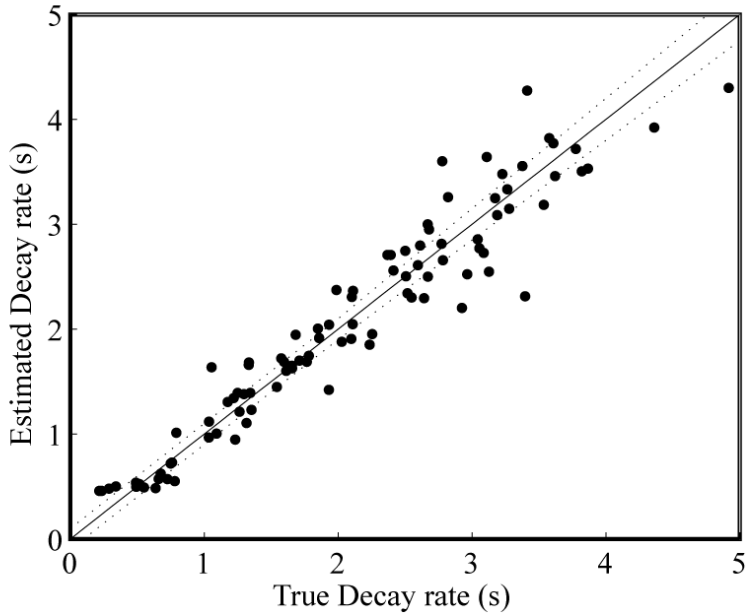


Figure 8-11. Comparison of estimated and true EDT. Estimates were obtained from the application of the MP method to simulated impulse responses convolved with 40 minutes of anechoic music windowed into four, ten minute segments.

Figure 8-12 and Figure 8-13 show the results for C_{80} and t_s using four ten minute segments. The under estimation at high C_{80} and low t_s values is similar to the features seen with speech, and is thought to arise because decay of the anechoic musical notes causes inaccuracies in the estimations. To improve the accuracy for EDT C_{80} and t_s requires a longer length of signal. The variation in the estimates is due to the small number of decay phases used to compute the estimate and the highly stochastic nature of the source signal. It is postulated that a recording of two hours would be sufficient to ensure plenty of suitable decay phases.

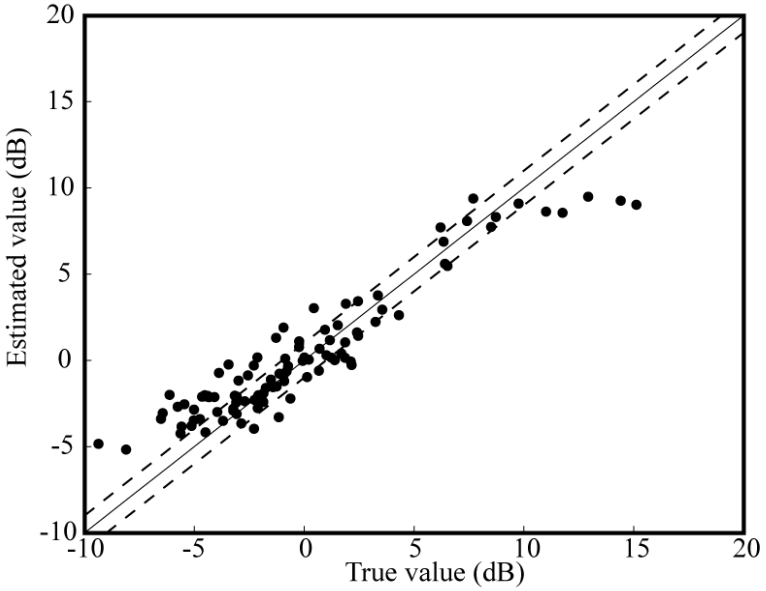


Figure 8-12. Comparison of estimated and true C_{80}. Estimates were obtained from the application of the ML method to simulated impulse responses convolved with 40 mins of anechoic music windowed into 10 minute segments.

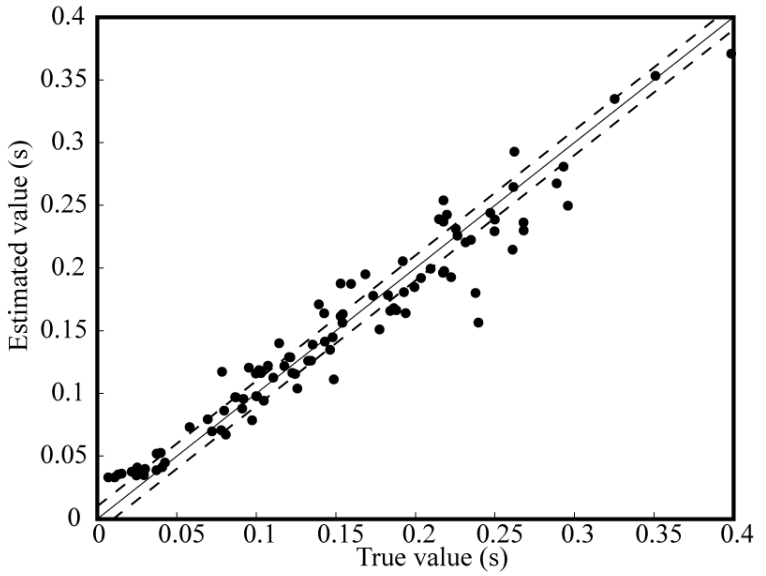


Figure 8-13. Comparison of estimated and true t_s. Estimates were obtained from the application of the ML method to simulated impulse responses convolved with 40 mins of anechoic music windowed into 10 minute segments

8.2.2 Results from all octave bands (music)

In order to further quantify the performance of the ML algorithm on music signals, the performance of the parameter estimations are analysed across more octave bands and

presented in Table 8-4, Table 8-5. Summary of the percentage of the parameter estimates that are within a difference limens of the true value . Estimates are based on an ML analysis of real room responses convolved with 40 minutes of anechoic music split into eight, five minute segments for Rt, and four, ten minute segments for EDT, C_{80} and $t_{s.}$

and Figure 8-14.. As discussed for speech signals the sampling frequencies were set for each band as described previously.

	Octave band (Hz)							
Parameter	**63**	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(%)	3	1	7	25	58	58	46	16
EDT(%)	1	2	10	26	38	50	54	39
C_{80}(%)	5	10	31	48	45	50	59	45
Ts(%)	7	2	14	32	37	46	62	59

Table 8-4. Summary of the percentage of the parameter estimates that are within a difference limens of the true value . Estimates are based on an ML analysis of simulated room responses convolved with 40 minutes of anechoic music split into eight, five minute segments for Rt, and four, ten minute segments for EDT, C_{80} and $t_{s.}$

	Octave band (Hz)							
Parameter	**63**	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(±s)	2.8	3.4	2.3	1.2	0.3	0.3	0.3	0.5
EDT(±s)	2.1	2.1	1.9	1.5	0.6	0.4	0.3	0.4
C_{80}(±dB)	6.1	4.8	4.5	4.9	5.1	4.5	3.1	4.2
Ts(±s)	0.13	0.14	0.11	0.08	0.04	0.03	0.02	0.02

Table 8-5. Summary of the percentage of the parameter estimates that are within a difference limens of the true value . Estimates are based on an ML analysis of real room responses convolved with 40 minutes of anechoic music split into eight, five minute segments for Rt, and four, ten minute segments for EDT, C_{80} and $t_{s.}$

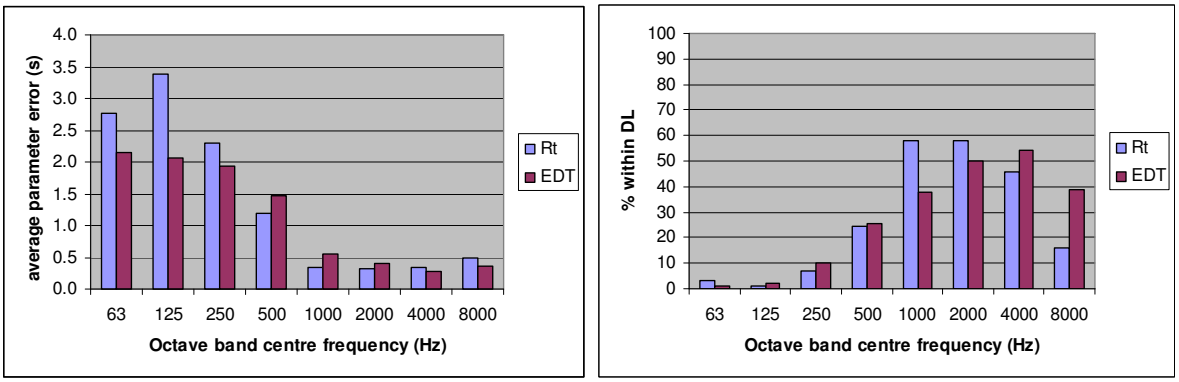


Figure 8-14. The average error for Rt and EDT as a function of octave band (left hand side) and the percentage of Rt and EDT predicted parameters within the subjective DL (right hand side). The analysis is based on the application of the ML method to simulated impulse responses convolved with 40 mins of anechoic music segmented into eight, five min segments for Rt and four, ten minute segment for EDT.

Table 8-4, Table 8-5. Summary of the percentage of the parameter estimates that are within a difference limens of the true value . Estimates are based on an ML analysis of real room responses convolved with 40 minutes of anechoic music split into eight, five minute segments for Rt, and four, ten minute segments for EDT, C_{80} and $t_{s.}$

and Figure 8-14. Show that for all parameters, useful results were achieved in the octave bands 1000 – 8000Hz. Rt in particular showed good performance, with the accuracy varying between ±0.3s and ±0.5s. In the 8000Hz band the estimations of Rt are worse than for EDT. By comparing the average spectrum from a five second portion of music with the spectrum from a silent region in the same signal, the signal-to-noise ratio at different frequencies can be examined. Figure 8-15 compares these two spectra and shows that the signal to noise ratio decreases with frequency. The relatively high noise floor in the upper frequency bands causes the late part of the estimated decay phases to decay more slowly. This slowing of the decay curves causes Rt to be overestimated. The EDT is less sensitive to errors caused by a decreased signal in noise ratio because less dynamic range is required to calculate the EDT. Rt results, in the 8000Hz octave band show a trend for overestimation while the EDT results are reasonably accurate, supporting the suggested theory.

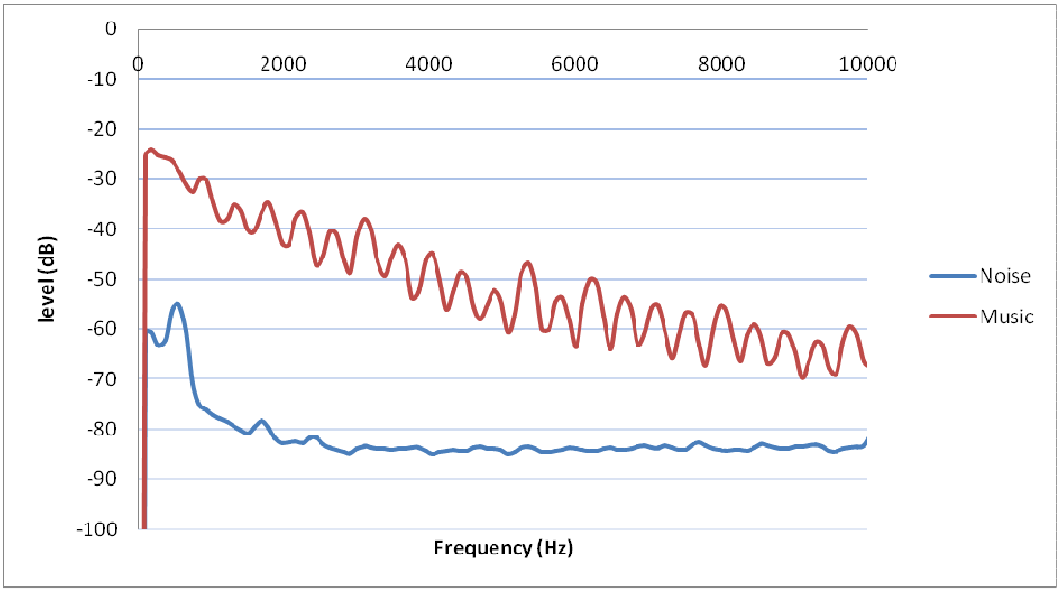


Figure 8-15. Noise vs. signal level for anechoic recordings

All parameters show large errors in the bands below 1000Hz (as shown in Table 8-4 and Table 8-5. Summary of the percentage of the parameter estimates that are within a difference limens of the true value . Estimates are based on an ML analysis of real room responses convolved with 40 minutes of anechoic music split into eight, five minute segments for Rt, and four, ten minute segments for EDT, C_{80} and $t_{s.}$

). This is due to the harmonic nature of musical instruments giving rise to sparse excitation in frequency, and causing the ML fit to be poor. In addition, the decay curve is not representative of broadband excitation. Both these factors cause errors in the parameter estimation. In the upper octave bands, much of the signal content is made up of higher order harmonics whose fundamental frequency is outside (below) the octave band frequency range being analysed. Higher octave bands will contain a larger number of harmonics and thus a more broadband like excitation as harmonics are linearly spaced.

Deterioration in performance in the lower octave bands also occurred with speech signals; however, the deterioration takes place at higher frequencies with music signals. The reasons for this can be explained by comparing the fundamental frequencies associated with speech and music notes/utterances. The fundamental frequencies of speech vary between 85-155Hz for adult males and 165-255Hz for adult females [87]. The fundamental frequencies of musical notes are not restricted to a specific range, but for example a low C_1 has a fundamental of 32.70Hz and a high C_6 has a fundamental of 1045Hz. This means that for music signals, excitation by very narrow-band signals can occur in higher octave bands than speech.

A second more intractable problem arises from a particular feature of many musical instruments. The decay of the envelope of a musical note is frequency dependant. Often lower frequencies partials (partials are members of a harmonic series) have a slower rate of decay than higher frequencies. This is particularly true of plucked string instruments like violins and harpsichords or percussive instruments such as drums or piano instruments. This slower rate of decay at lower frequencies can mask some of the decay due to reverberation. Results for clarity and centre time show similar trends to EDT and the accuracy is similar, in terms of % within the DL. Once again useful results are gained in the bands 1000-8000Hz. Below 1000Hz the accuracy begins to deteriorate because of the further deviation from broadband-like excitation.

8.3 Suitability of signals for ML estimation of decay curve

In order to study the suitability of a signal for the ML estimation acoustic parameters it is helpful to know how regions of decay there are with at least 25dB of dynamic range in each of the anechoic signals. Table 8-6 which indicates how many regions of decay with at least 25dB of dynamic range each of the anechoic signals contain (1 kHz octave band). This can indicate how suitable a signal is for estimation using the MLE method. Figure 8-16 shows the ML estimated decay phases superimposed over the reverberant speech envelope (RT=1.65s). Table 8-6 also indicates the percentage of silence within each signal (% of time the normalised signal is lower than -40dB), this, combined with the spectral variance, is a good indicator as to the suitability of the signal for parameter estimation using the envelope spectrum method.

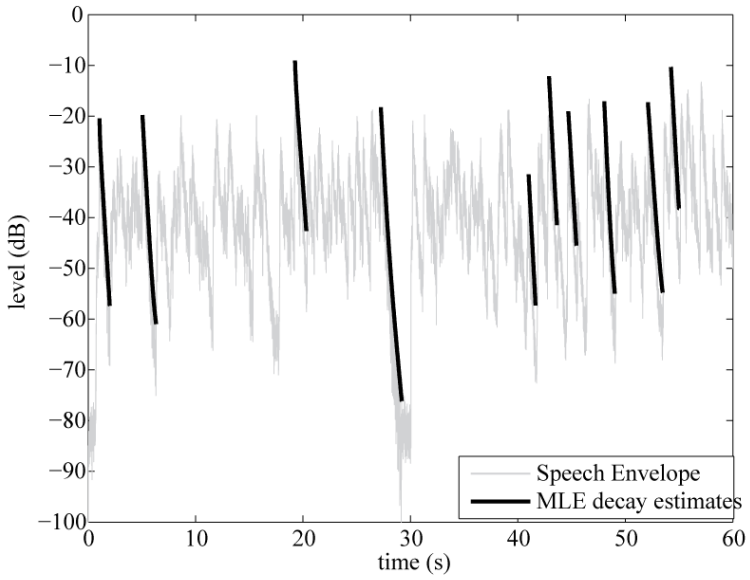


Figure 8-16. The selected decay phase estimated from 60s of narrated speech. after Kendrick et al. [73].

Signal	Average number decay phases / minute	Amount of quiet in signal (%)
Music track 1	4.18	0.19
Music track 2	4.27	2.22
Music track 3	4.76	0.86
Music track 4	12.04	6.36
Music track 5	3.17	0.53
Music track 6	4.96	0.17
Speech	37.35	24.92

Table 8-6. Average number of decay phases per minute exhibiting at least 25dB of decay, and percentage of quiet in each anechoic signal. Percentage of quiet is calculated by computing the percentage of (non-overlapping) 0.05s length windows in each signal with energy 40dB less than the maximum energy, after Kendrick et al. [73].

Table 8-6 indicates that music track 4 and speech contains the highest number of decay phases / minute and also the highest percentage of silence in the signal. This correlates with the results which found that when performing the ML estimation of parameters on individual pieces of music, speech and track 4 produced the most accurate results while the accuracy was much lower (and quite similar) for all the other signals.

8.4 Real room measurements

To validate the ML method, a number of recordings of real orchestras were made in situ. Two halls were measured, 'The Atrium' at Bradford University and the Haden Freeman Concert Hall at the Royal Northern College of Music. In both cases a rehearsal was recorded where the audience area was semi-occupied. Up to two hours of recordings were made for each location. The control measurements were made using repeated snare drum hits and hand claps. The reason for this type of control measurement was due to the logistical difficulties encountered when making in situ measurements using standard measurement methods (swept sine waves, MLS sequences or interrupted noise). These reasons, ironically, were the main reasons for developing a blind acoustic parameter estimation method in the first place.

It was decided that impulsive excitation using the musicians themselves would be sufficiently accurate, provided the parameters were averaged from a number of

impulses. ISO 3382 recommends that if impulse excitation is to be used then at least 45dB SNR is required for Rt_{35} although if measuring Rt_{20} then 35dB is sufficient. During the first measurement at 'The Atrium' at Bradford University, repeated snare drum hits were used, however this was found to give poor signal-to-noise ratios. This was due to the multifunction nature of the space (people wandering through the area) and the fact that a hail storm occurred during the measurements! Therefore during the second experiment at the Royal Northern College of Music, the whole orchestra was asked to clap simultaneously to provide a louder source. As the musicians were very skilled, it was thought that by the conductor visually counting them in prior to each clap, it would be sufficiently simultaneous.

8.4.1 Distributed RIRs

In an orchestra the players are distributed over a large area, and therefore measurements obtained where the whole orchestra clapped simultaneously are very different than those obtained from single sources. The distributed measurements include the relative delays between the sound sources due to the distribution of the musicians and the particular location of the listener. This is an interesting problem for measuring acoustic parameters, as most are calculated using single source to receiver RIRs, while the actual sound heard by the listener is the superposition of a number of different sounds affected by a number of different RIRs. Information regarding the effect of the distribution of sound sources is generally only accounted for by crude averaging of parameters. This does not account for the varying delays in receiving the direct sound that occurs across the different sources, and could quite conceivably have a large impact on aspects such as perceived clarity, because the instruments can be spread over a wide area, causing the direct sound to be received over a range of time delays. To illustrate the distributed RIR phenomena consider a line of musicians 20m in length, where a listener is seated 15m away perpendicular to one end of the line, here the listener experiences delays in receiving the direct sound ranging from 43.7ms to 72.9ms.

Standard acoustic parameters do not account for this particular feature of distributed sources, however when using the ML method the decay curve estimated is from the distributed source and is the result of the superposition of numerous RIRs. Figure 8-17 shows two examples of a distributed impulse response recorded in the RNCM (the

musicians clapped simultaneously – brought in visually via their conductor). This was recorded about five metres off the central axis of a concert hall at a distance of roughly 25m from a full size orchestra. A notable feature of these impulses is the build up of sound at the start of the RIR due to sound from the closest sources arriving first, and as more direct sounds from more distant sources arrive and mingle with early reflections from closer sources, the density and level of the sound increases. The level then begins to decrease as the direct sound from all the sources arrives and the early and late reflections from all RIRs mixes and decays away.

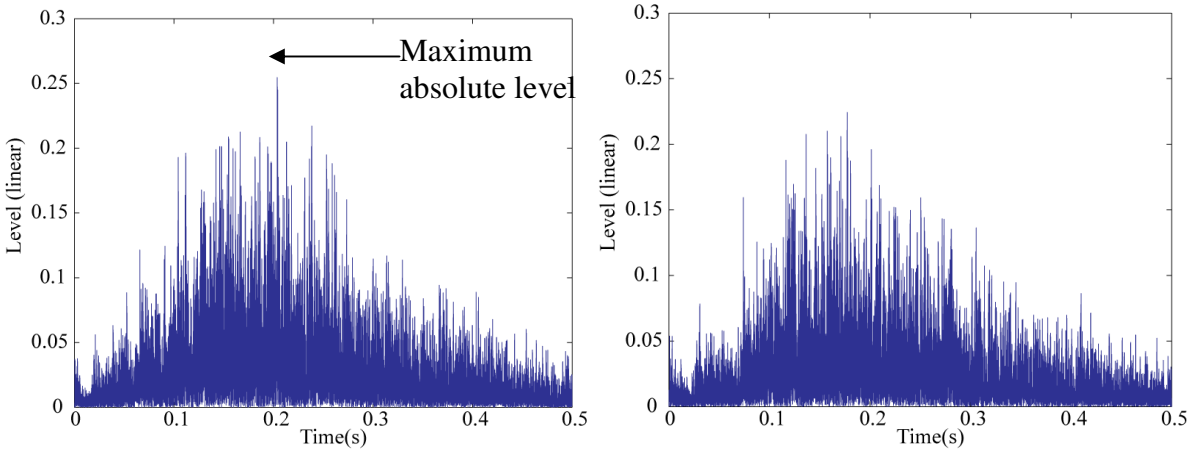


Figure 8-17. Examples of the 'distributed source' impulse responses. Impulse responses recorded by orchestra members clapping simultaneously.

Calculating parameters from a distributed impulse response is not standard practice, and further research is required to determine the correct treatment of these distributed RIRs in relation to acoustic parameters but this is beyond the scope of this thesis. Here a simple treatment of these RIRs is performed in order to simply compare the properties of the distributed impulse with ML estimated parameters. In this treatment the maximum absolute level is selected as the starting point of the impulse and the acoustic parameters then calculated using standard methods. This starting point is illustrated in Figure 8-17. The reason for this treatment is that it is thought that the initial quiet sounds will have little subjective effect while effecting significant changes in acoustic parameters (such EDT or C_{80}). Additionally, this is how the ML method identifies the start of each decay curve, making the comparison between true and estimated parameters more appropriate.

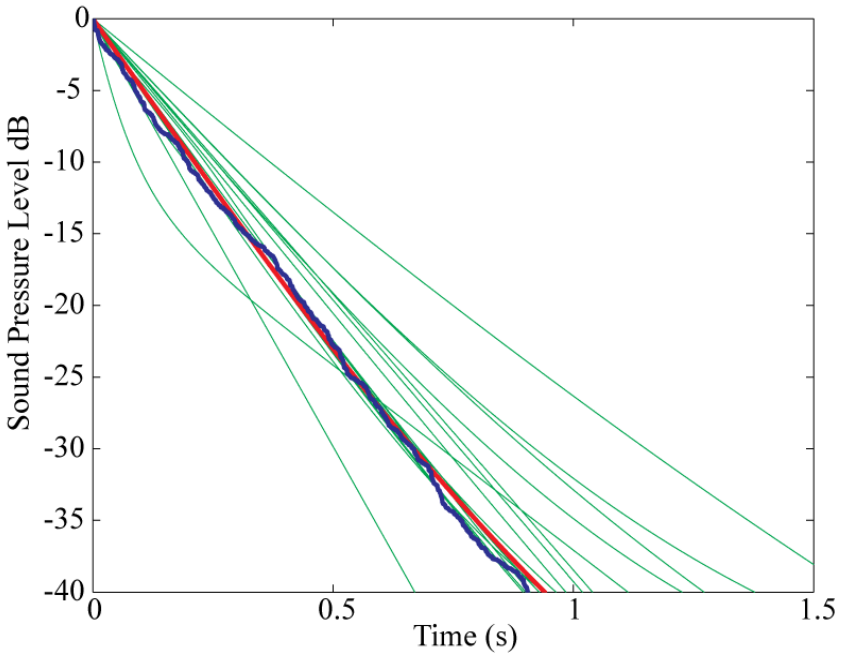
8.4.2 'The Atrium' - Bradford University

The Atrium is a large space with a domed plastic roof. The space is coupled to a number of hallways and corridors. The Bradford University Orchestra was rehearsing for a concert later that day. Two omnidirectional measurement-quality microphones were set up on stands. Microphone 2 was positioned on the 3rd from last row in the right hand side of the audience. Microphone 1 was positioned in the 3rd from the front row on the left hand side of the audience. The control impulse responses were produced

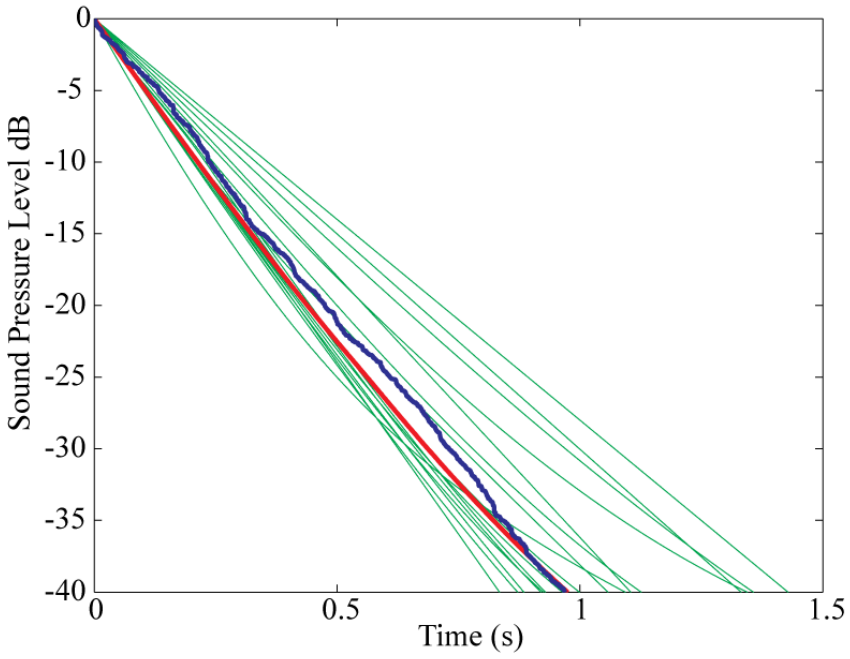
using repeated snare hits, with the drum being located roughly in the centre of the orchestra. During the recording, the orchestra was rehearsing the piece 'Night on the Bare Mountain' by Rimsky-Korsakov, after which the wind group rehearsed. The whole rehearsal was recorded and provided about two hours of material from which to perform the estimation. It should be noted that during the rehearsals very high levels of background noise were produced by a hail storm falling on the roof. Recordings were made at 48 kHz sampling frequency onto an Apple Macbook computer running Audacity.

Figure 8-18. Photographs of the Atrium at the University of Bradford.

The recording contained long gaps where the musicians were discussing or setting up, so these were removed taking care not to remove any useful segments or provide any artificially fast regions of decay which may cause estimation problems. This provided roughly 90 minutes of usable material. This was down-sampled to a 3kHz sampling rate and filtered into the 1kHz octave band, the recording was segmented into 22, four minute segments, ensuring each overlapped by at least two seconds to ensure no decay phase would be lost by the segmentation process. The maximum likelihood estimation was performed on the resulting segments. The estimated decay curves are plotted in Figure 8-19, which show quite promising results from the algorithm. The blue line represents the control impulse, calculated as the average decay curve from the four snare drum strikes.



1)



2)

Figure 8-19. Decay curve estimates for the Atrium at the University of Bradford. Green lines are individual segment estimates, red is the optimal decay curve estimate and blue is the averaged true decay curve from four snare hits. Measured at two microphone positions (one and two).

Table 8-.7 compares acoustic parameter estimation using the ML method with the actual parameter averaged over four recorded impulse responses (snare hits). Results show

that for both microphone positions results are within, or close to, the subjective difference limens.

Mic position		Rt (s)	EDT (s)	C_{80} (dB)	t_s (s)
1	**Estimate**	3.2	2.6	-2.6	0.24
1	**Actual**	3.1	2.8	-2.9	0.23
2	**Estimate**	3.1	2.6	-2.5	0.19
2	**Actual**	3.1	2.9	-2.9	0.20

Table 8-.7 Comparisons of estimated and true acoustic parameters for 'The Atrium' at the University of Bradford. Estimated parameters were obtained by the ML method while true parameters were obtained from repeated snare hits (parameters were averaged across four responses).

Confidence limits for the estimated parameters can be computed to indicate the possible level of accuracy that has been achieved. To compute the confidence a bootstrap method is used to compute upper and lower confidence limits for the decay curve by looking at the sample-by-sample distribution of the level over all the curves used in the estimation of the median decay curve (Section 8.6 details how to calculate these confidence limits). The upper and lower 95% confidence limits of the decay curve were used to compute limits on each estimated parameter. These limits, calculated for each parameter, are not confidence limits in the strictest sense, as they are computed indirectly rather than by evaluating the variance of the parameters directly. However they still provide useful quantitative measure of the accuracy of the decay curve with respect to a particular parameter.

It is useful to use these confidence limits to compare the accuracy of the control RIR (snare drum excitation) with the ML estimation. The confidence limits for both the control, (four snare drum hits) and the ML parameter estimates (mic position 1) are presented in Table 8-8 and Table 8-9. It can be seen that both the ML parameter estimates and the estimates using the snare drum excitation have similar confidence limits for most parameters (eg 2.8-3.3s for ML estimated Rt and 2.9-3.3s for the snare excited RIR). This suggests that the ML method produces similar accuracy compared with using snare drum excitation averaged over four hits therefore it is difficult to gleam any further information as to the accuracy of the ML method.

	Rt	EDT	C_{80}	t_s
Upper 95% limit	3.3	3.1	-2.6	0.24
Median estimate	3.2	2.6	-2.6	0.24
Lower 95% limit	2.8	2.5	-3.9	0.19

Table 8-8. Confidence limits on parameter estimates for the acoustic characteristics of the Atrium in Bradford using the ML method.

	Rt	EDT	C_{80}	t_s
Upper 95% limit	3.3	3.1	-1.8	0.27
Median estimate	3.1	2.8	-2.9	0.23
Lower 95% limit	2.9	2.5	-3.7	0.19

Table 8-9. Confidence limits on parameters computed from the snare drum excited RIR in the Atrium in Bradford, computed from 4 snare drum hits.

8.4.3 The Royal Northern College of Music

The Haden Freeman Concert Hall is a hexagonal purpose built concert hall with a seating capacity of 466. The orchestra was rehearsing 'The Planets' by Holst. Recordings were made in two locations using the same equipment as the previous recording. Microphone 1 was located in the central set of seats slightly right of centre (facing the stage). Microphone 2 was located in the middle of the left hand side set of seats.

Figure 8-20. The Haden Freeman Concert Hall, at the Royal Northern College of Music.

Ninety minutes of the rehearsals were recorded and a series of control impulses were also recorded by asking the first violinist to clap five times (with long pauses in-

between each clap). Results from the estimation (using 10 minute segments) are presented in Figure 8-21 and Table 8-10. There were some problems with position 2 which was found to have excessive noise on the signal and often clipping occurred, so information obtained from position 2 was considered to be unreliable and discarded. No such problem occurred at position 1.

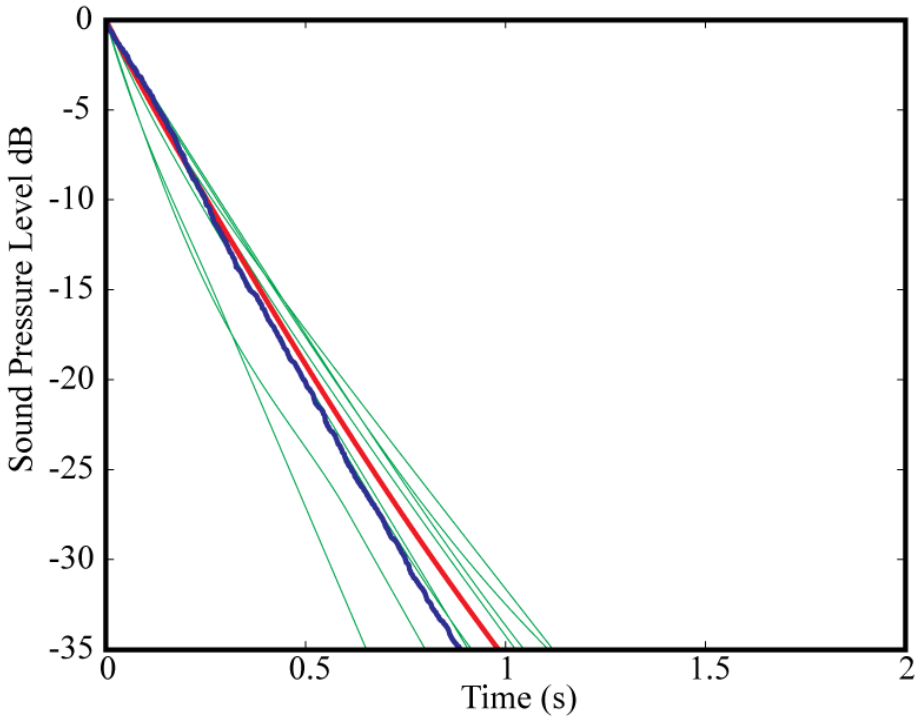


Figure 8-21. Decay curve estimates for the Haden Freeman Concert Hall. Green lines are individual segment estimates, red is the optimal decay curve estimate and blue is the averaged true decay curve from 4 snare hits. 8, 10 minute segments were used.

Table 8-10. compares the acoustic parameter estimation using the ML method with the actual parameter averaged over four recorded impulse responses computed for both the single hand clap and the multiple simultaneous hand claps. To create the multiple simultaneous hand claps, the whole orchestra was asked to clap simultaneously 5 times with long gaps in between each clap. To achieve this, the conductor counted everybody in (visually using the baton) to yield simultaneous claps. These impulses provided the response to multiple simultaneous sound sources. The results from this table are discussed in the next paragraphs.

	Rt (s)	**EDT (s)**	**C_{80} (dB)**	**t_s (s)**
Estimate	1.6	1.5	1.5	0.38
Actual (single clap)	1.5	1.3	1.8	0.44
Actual (simultaneous clap)	1.5	1.4	0.7	0.38

Table 8-10. Comparisons of estimated and true room response parameters for the Haden Freeman Concert Hall. Estimated parameters were obtained by the ML method, while true parameters were obtained from hand claps. Actual parameters were calculated from both the 1st violinist's hand clap but also the whole orchestra clapping simultaneously (parameters were average across a number of responses).

Comparing the ML estimated result with the parameters obtained from the single hand clap in Table 8-10. shows that results are generally very good for most parameters and are within, or very close to being within, the DL. There is a tendency for the overestimation of EDT which can be explained by examining the results for the 'distributed source' RIR as discussed in the following paragraphs.

The accuracy of the ML estimated parameters are detailed in Table 8-11 and Table 8-12. This shows that the variability in the control RIR (from five repeated single hand claps) is lower when compared with the control measurements using repeated snare hits in the Bradford atrium. It is thought that this is due to a combination of the noise level being lower at the RNCM concert hall (the atrium is an open space and is not purpose built concert hall and the recording was during a hail storm!) and the higher degree of variability in the excitation between snare hits when compared with hand claps. The method indicates that the control RIR produces quite accurate results and therefore it is reasonable to treat the results as a true measure of room response. It can be seen that the ML Rt measure lies within the 95% confidence bounds for the control measure and is therefore treated as un-biased and accurate. The ML estimate for EDT however, does not lie within the 95% confidence bounds for the control EDT measure. This is an indication of a biasing factor that is affecting the ML estimate, however, the ML EDT estimate is within the 95% confidence bounds for the parameter calculated from the distributed source. With the exception of clarity, all of the ML estimated parameters are closer to the distributed source parameters than the single clap parameters. This suggests that, as expected the ML decay curve estimate is more representative of distributed excitation than excitation from a single source.

	Rt	EDT	C_{80}	t_s
Upper 95% limit	1.8	1.8	1.8	0.47
Median estimate	1.6	1.5	1.5	0.38
Lower 95% limit	1.6	1.2	1.2	0.33

Table 8-11. Confidence limits on parameter estimates for the acoustic characteristics of the Haden Freeman Concert Hall in RNCM using the ML method.

	Rt	EDT	C_{80}	t_s
Upper 95% limit	1.6 (1.5)	1.4 (1.5)	3.1 (1.0)	0.51 (0.39)
Median estimate	1.5 (1.5)	1.3 (1.4)	1.8 (0.7)	0.44 (0.38)
Lower 95% limit	1.5 (1.4)	1.1 (1.3)	0.6 (-0.1)	0.37 (0.34)

Table 8-12. Confidence limits on parameters computed from the Haden Freeman Concert Hall RIR (computed from 5 single claps). The estimate and confidence limits for the distributed source RIR are presented in brackets.

Table 8-10., Table 8-11 and Table 8-12 shed some light on one possible reason for the overestimation of the EDT when comparing the distributed source RIR parameters with the single clap parameters. The distributed source RIR, after it is truncated so that the start of the response is the maximum absolute level, is now a mixture of the late decay of the claps that arrived prior to the selected starting point and the earlier reflections of those arriving at or after the selected starting point. Therefore the rate of decay that is seen in this distributed impulse response is a mixture of both the early and late decay rates of single source impulse responses. This is seen in the value for the distributed Rt (1.5s) and the single source EDT (1.5s). This may also account somewhat for the ML method EDT estimate being 1.4s rather than 1.3s (compared with the single source results in Table 8-10.), one explanation is that the distributed nature of the orchestra means that unless a single instrument is playing there will always be some 'interference' between the early and late decay rates of different paths.

Parameters from a distributed source would be more meaningful, but current knowledge does not allow parameters to be calculated directly from a distributed source RIR measure. The author is not aware of any research investigating the effect of source distribution on perception which is then utilised to either validate the use of existing acoustic parameters, such as reverberation time or clarity, or define new ones. This is an interesting area of further research and may have ramifications on the influence of orchestral layout on the perception of the room acoustics.

Additionally a method may be possible, utilising multiple microphone recordings of orchestras, where Blind Source Separation (BSS) algorithms may be used to separate the individual instrument signals from a whole orchestra recording. The ML algorithm may then be applied to these individual instrument signals (or groups of closely positioned instruments) to gain acoustic parameter estimations for different source-receiver paths, averaging of multiple parameters should gain results similar to the standard method in ISO 3382 [8].

8.5 Maximum likelihood estimation of spatial impression

It is desirable to estimate parameters that characterise the 'spaciousness' of a room. As already mentioned in Section 2.3.5, a number of parameters have been developed to quantify spaciousness. Such parameters include, Early Lateral Energy Fraction (ELEF eq. (2-9)), Late Lateral Strength (LG eq. (2-11)) and the Inter-Aural Cross-correlation Coefficient (IACC eq. (2-10)). IACC is not suitable for estimation using the ML method as the method depends heavily on the temporal structure of the impulse response to produce a correlation value between the two signals at the ears. The ML method assumes a uniform temporal structure of received reflections and therefore the ML estimation of IACC using this room model will yield meaningless results. However, both the ELEF and LG measures calculate their values using energy summations over specific time intervals. This makes their values less sensitive to deviations within the temporal pattern of reflections and therefore should be more suitable for ML estimation as indicated in Section 7.2.6.

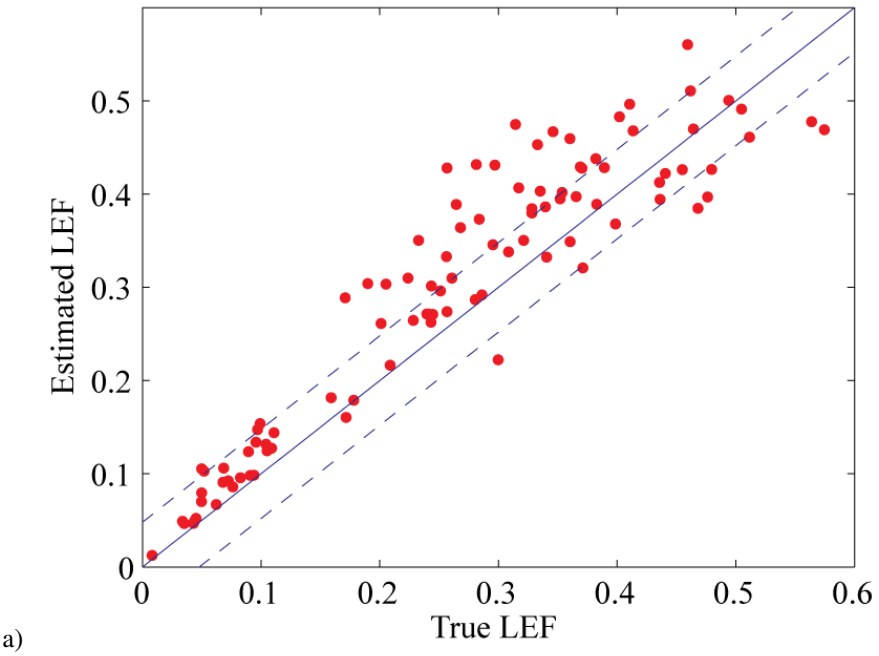
8.5.1 Level calibration

LG is dependent on the source level, the ML method loses this information as the estimated decays are all normalised. This normalisation is required so that final decay estimations can be extracted from a series of decay phases of varying level. To compute the LG, both the estimated and real impulses response need to be scaled to a level related to the average energy received at the microphone. To do this, the normalised RIR for each channel is multiplied by the total rms level of the received reverberant

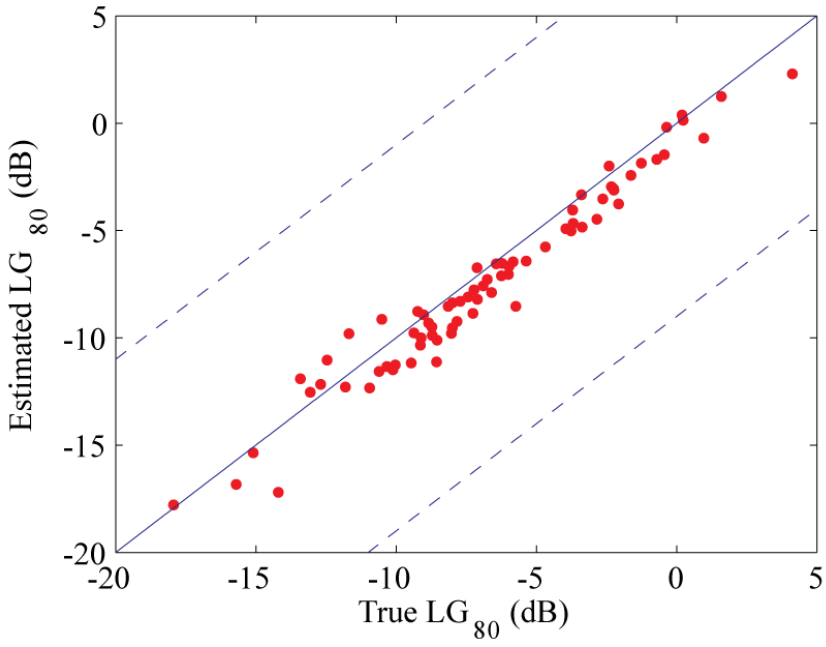
signal for that channel. This is done for both real and estimated impulses response for both the omni and figure-of-eight microphone pickup patterns.

8.5.2 Application of ML to the estimation of ELEF and LG

A series of simulated room geometries and room impulse responses are generated in the same manner described in Chapter 4. However, in addition to the omni-microphone model, a figure-of-8 microphone is placed in the same location with the null pointing towards the sound source. An 180s sample of anechoic speech is convolved with both RIRs to produce the test material. Each reverberated signal is windowed into four segments of equal length. The maximum likelihood (method c) is used, yielding an optimal decay estimate from each of the segments and calculating the median decay estimate. This yields a decay estimate from both the omni and figure-of-8 microphone signals. In calculating the optimal decay curve the actual sound level is lost as the method requires the normalisation of the decay curve. This is not a problem as each decay curve is normalised after estimation to the total rms level of the reverberated speech (as described in the previous section). The results from these analyses are presented in Figure 8-22.



a)



b)

Figure 8-22. Maximum likelihood estimates for a) Early Lateral Energy Fraction (ELEF) and b)Late Lateral Reflection Strength LG.

Figure 8-22 shows that the estimation accuracy for LG is good over a wide range of parameter values. The estimation accuracy for the Early Lateral Energy Fraction is less accurate and often exceeds the difference limens, particularly for high ELEF values. Part of the reason for this is that the LG DL are much wider than the ELEF, because

human perception is more sensitive to changes in apparent source width than to changes in envelopment. Consequently, the accuracy requirement for LG is much less than for ELEF. This is also coupled with the fact that LG is calculated using only the late part of the decay curve and the ML method is known to be more accurate at estimating this than the early part, for which the ELEF requires an accurate estimate. This is consistent with the estimation accuracy for early and late decay rates (Rt and EDT). As previously discussed, the detail that is lost when the decay curve is estimated, due to either the presence of decay tails of utterances/musical notes or the short comings of the model in the early region, causes inaccuracies in parameters that are particularly dependant on aspects of the early sound-field.

8.6 Determining the accuracy of the parameter estimation

Confidence limits for the decay curve have been utilised in the previous sections. This section explains in more detail how these limits are calculated.

Once a blind estimation is performed, it is desirable to have a blind measure of the accuracy of the parameter estimate. To achieve this, the distribution of level for each sample over all decay curve estimates may be examined. The more widely spread the estimates, then the more uncertainty in the final estimate. The measure must be calculable from the available data (i.e. blind). This blind indication of accuracy can enable further optimisation of the algorithm to yield more accurate results.

The decay curve estimate is generated by finding the sample-by-sample median of a number of decays from different segments of the signal. By examining the sample-by-sample distribution of levels over all these decay curves it is possible to define confidence limits for the decay curve. It is proposed that by tuning parameters in the ML algorithm, such as the segment length, it may be possible to optimise the algorithm to yield more accurate results.

Confidence limits can be derived for decay curves and from these limits confidence intervals for each of the parameter estimates are calculated. When computing the mean of a population, confidence intervals on the mean value can be computed using a one-

sample t-test[1]. Computing the confidence interval for the median value is not quite as simple. Percentile based confidence intervals can be computed using the boot-strap method. Boot-strap uncertainty estimates are computed using sub-samples of the data set that can have a sample size equal to or less than the number of available samples. A large number of sub-sample sets are generated and the median computed for each. The distribution of median values, computed from the cumulative distribution function, can then be used to find the 5% and 95% percentiles, and hence the 90% confidence limits for the decay curve.

A long (>$1^1/_2$hr) recording is windowed into four min segments, which provides a good length of data to yield accurate estimations. The properties of the source signal can have an impact on the estimation accuracy, for example slow legato pieces do not provide as many opportunities for estimation as staccato pieces. However, the method is quite rigorous in its selection procedure, as if no suitable decay phases are available the method will yield no estimates; the method has a certain level of quality control built-in. Figure 8-23 shows the distribution of the estimation energy decay curve from the Bradford Atrium results (Section 8.4.2). For each sample, a non-parametric probability density function is fitted to the level value of all the ML decay curve estimates for each sample. The distribution fit is achieved using the kernel smoothing method in the Matlab statistics toolbox, ksdensity.

[1] http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm

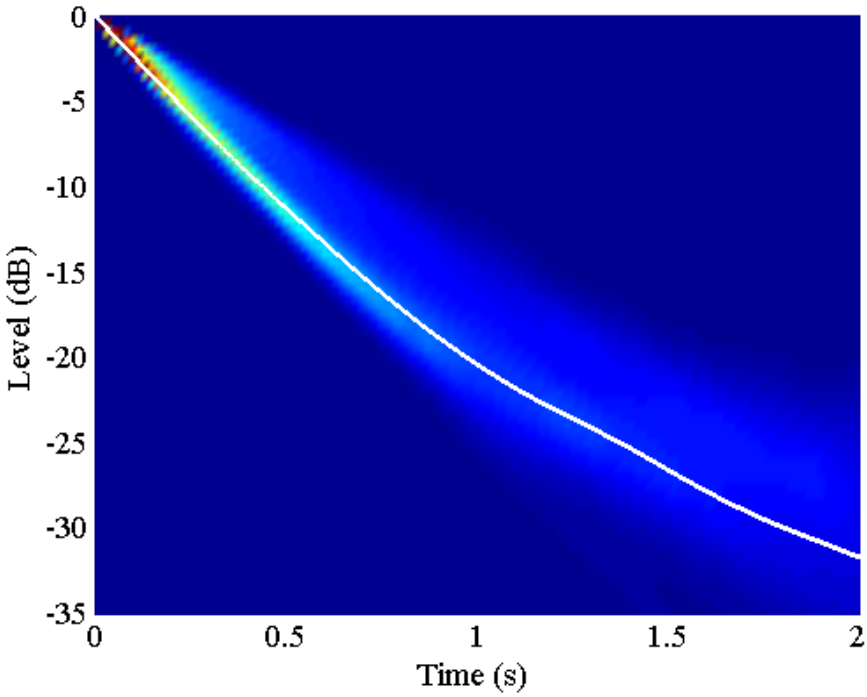


Figure 8-23. Approximate distribution of decay curve estimates, the white line is the median decay curve

Using the boot-strap method the confidence limits for each sample value of the impulse response envelope model can be computed across all the estimates. The 90% confidence intervals for the decay curve can then be estimated. The resultant 90% confidence limits are shown in Figure 8-24. Figure 8-24 also shows the confidence limits for the snare drum based estimate of the decay curve. The RIR confidence limits were computed using the same bootstrap method and show that the ML estimated decay is within the confidence limits for the snare drum estimated decay curve down to approximate -25dB.

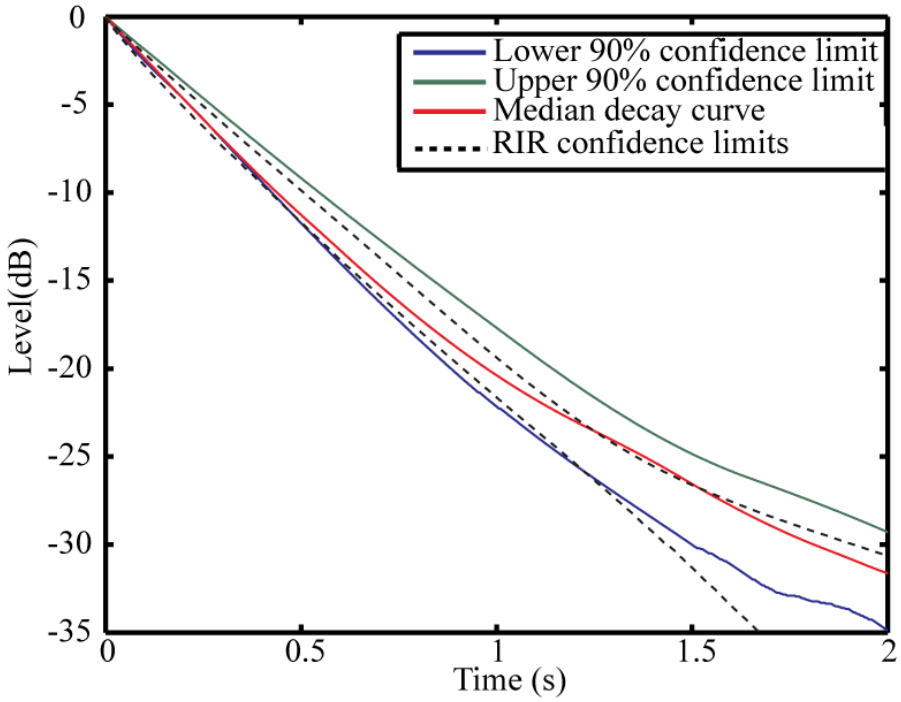


Figure 8-24. 90% confidence bounds for median decay curve estimation

In addition to these decay curve bounds, a set of confidence limits can be computed on the acoustic parameters themselves. This gives a rating of the 'quality' of the parameter estimate. Table 8-13 indicates the quality of the acoustic parameters estimated at the Atrium in Bradford. These parameters can be used to tune the estimation algorithm, by altering the segment length and available data until the confidence limits are contracted.

	Rt	EDT	C_{80}	$\mathbf{t_s}$
Upper 90% limit	3.7	2.98	-2.4	0.21
Median estimate	3.1	2.60	-2.5	0.19
Lower 90% limit	2.9	2.50	-4.1	0.17

Table 8-13. Confidence limits on parameter estimates for the acoustic characteristics of the Atrium in Bradford using the ML method.

Using the recording from the Royal Northern College of Music, a series of parameter estimates and confidence bounds were generated using the same 1½ hr recording but varying the segment length and number. This yielded the results in Table 8-14;

Segment Length (s)		Rt (s)	EDT (s)	C_{80} (dB)	t_s (s)
4	*Lower Limit*	*1.57*	*1.49*	*-0.64*	*0.10*
	Median	**1.86**	**1.63**	**0.23**	**0.12**
	Upper Limit	*2.08*	*1.77*	*0.49*	*0.13*
6	*Lower Limit*	*1.51*	*1.45*	*-0.27*	*0.10*
	Median	**1.68**	**1.59**	**0.34**	**0.11**
	Upper Limit	*1.81*	*1.68*	*0.51*	*0.12*
8	*Lower Limit*	*1.65*	*1.40*	*-0.15*	*0.09*
	Median	**1.69**	**1.53**	**0.50**	**0.11**
	Upper Limit	*1.82*	*1.73*	*1.56*	*0.12*
10	*Lower Limit*	*1.59*	*0.99*	*-0.14*	*0.07*
	Median	**1.55**	**1.49**	**0.60**	**0.11**
	Upper Limit	*1.83*	*1.74*	*3.33*	*0.12*

Table 8-14. 90 % confidence limits on parameter estimates, estimated using the bootstrap method using 5000 boot-straps on data from the RNCM. These give an indication as to how the quality of the estimation varies with segment length

The 90% confidence interval for acoustic parameters estimates can be used as a measure of the confidence there is in the decay curve estimate. Figure 8-25 shows the 90% confidence limits expressed as a fraction of the actual parameter. The graph indicates that most confidence can be placed in the result when the segment length is six minutes.

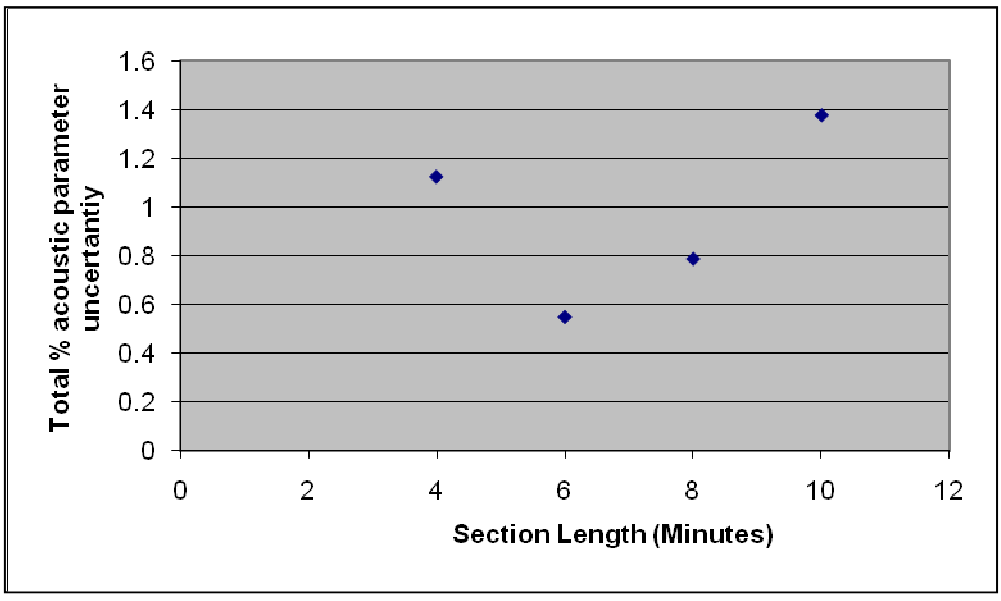


Figure 8-25. The 90% confidence limits expressed as a fraction of the actual parameter and averaged over all parameters, as a function of the segment length.

The parameter confidence interval span (i.e. upper minus lower limit) provides a measure of the quality of the estimate. This value can be optimised by re-running the ML algorithm with varying segment lengths, or increasing or decreasing the overall number of segments. This measure of parameter quality is very useful as it is blindly computed using only the outputs from the ML algorithm.

8.7 Performance of the ML method with interfering noise

Noise is a problem for many acoustic measurement systems and must be taken into account when proposing a new one. Often, noise will be present on a recording which is either the result of electrical or acoustical interference. Electrical noise can be due to the performance limits of microphones, electromagnetic interference or limitations of the recording device (bit depth/aliasing). Acoustic noise comes from unwanted noise sources within the space such as noise from an air-conditioning system or the audience. Electrical noise will be a particular problem in low level recordings such as when using speech where the signal to noise ratio is quite small. Both types of noise will have a significant impact on the measurement system accuracy. During measurements, electronic noise can be removed by selection of appropriate equipment. For occupied measurements, the noise generated by air conditioning and the audience is not going to be so simply removed.

To investigate the effect of noise on the ML parameter estimation system, an experiment was carried using 100 synthetic impulse responses convolved with 180s of anechoic speech. The result was split into four segments and method c was used to estimate the decay curve and acoustic parameters. The simulation was run multiple times, each time a different level of random white noise was added to the reverberant signal. This simulates stationary acoustic noise, such as an air-conditioning fan. Adding noise after the convolution operation of the speech with the room impulse response is considered to have an equivalent effect, but the signal-to-noise ratio is easier to control.

The experiment was run at a number of signal-to-noise ratios. Initially, the system simply failed to produce an estimate unless the signal-to-noise ratio was above 25dB. This is consistent with the algorithm design and shows that the decision making algorithm for acceptable decay phases is working correctly. In order to further test the algorithm, the decision making process for valid decay phases were altered slightly, with a number of levels of acceptability for the estimated signal-to-noise ratio defined. If no 25dB decay phases exist then the acceptable limit was changed to 20dB, if no 20dB decay phases were available the limit was then change to 15dB, then to 10dB. This yielded the following results in Figure 8-26 for Rt and EDT.

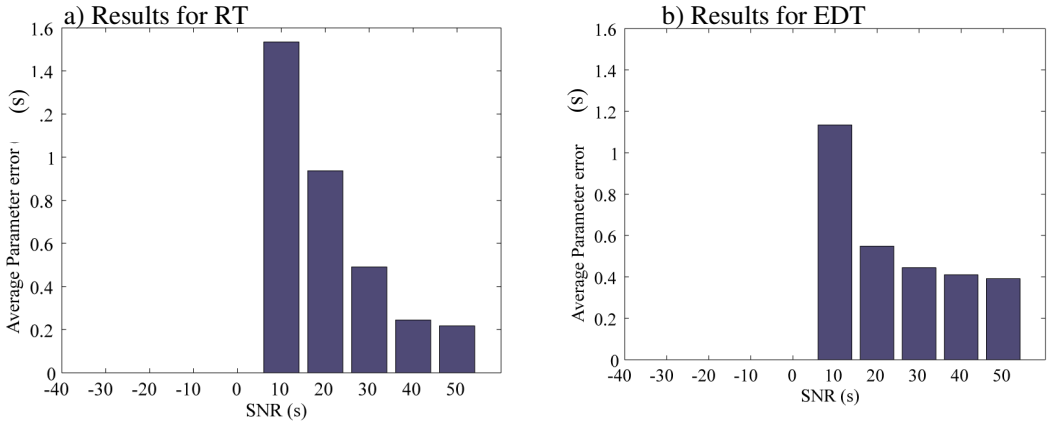


Figure 8-26. Effect of change of signal to noise ratio on parameter estimation accuracy using speech as excitation signal a) result for RT, b) result for EDT.

The results show that EDT remains quite accurate down to 20dB signal-to-noise ratio but once it reaches 10dB the accuracy is compromised. This is quite intuitive as EDT is a measure that uses the first 10dB of decay to compute the early rate of decay. On the other hand the Rt accuracy becomes compromised between an signal-to-noise ratio of 20 and 30 dB. Once again this is an intuitive result as the Rt uses the range from -5dB to -25dB to compute the rate of decay. The algorithm produced no results for very low signal to noise ratios as the algorithm successfully identified the low signal-to-noise ratio condition and did not yield any estimates. Figure 8-27 shows the effect SNR ratio has on the estimation accuracy of C_{80} and t_s.

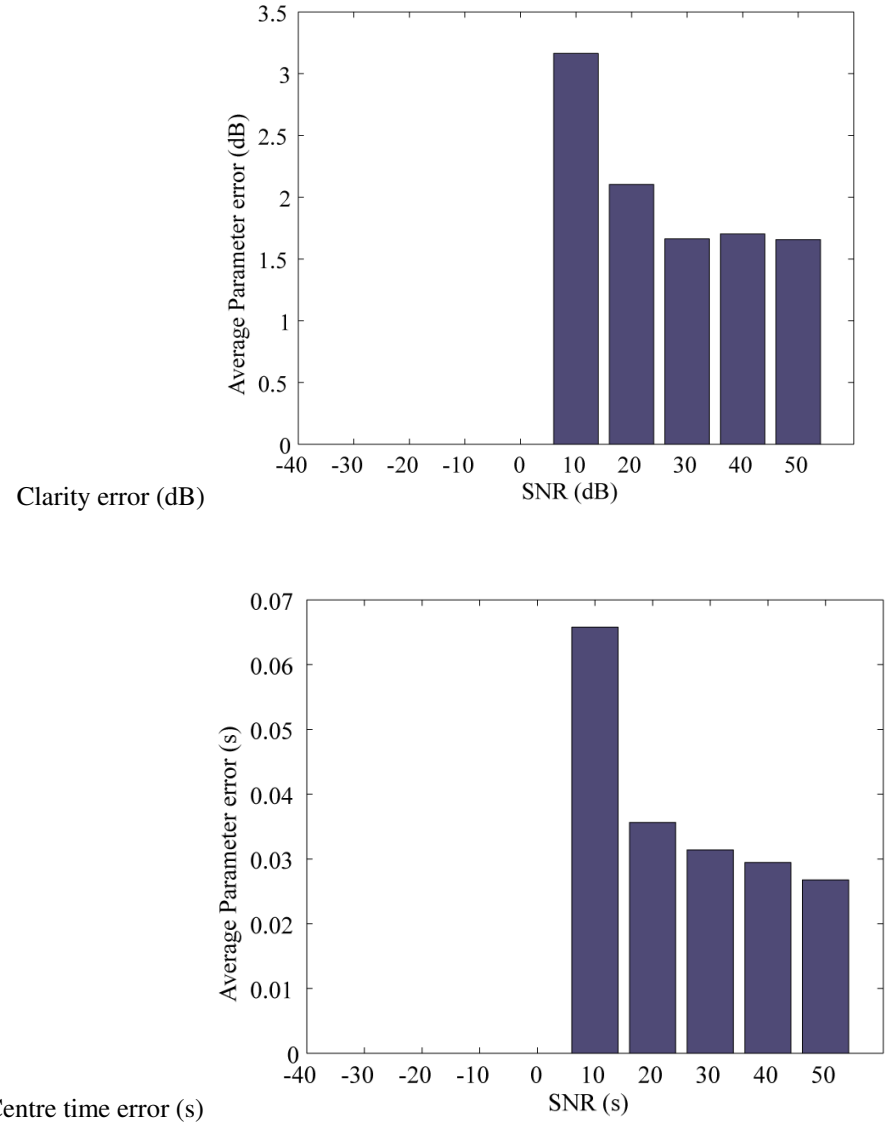


Figure 8-27. Effect of change of signal to noise ratio on parameter estimation accuracy using speech as excitation signal 1) C_{80} 2) t_s

It was found that all these parameters require at least between 10dB and 20dB signal-to-noise ratio to achieve a similar accuracy as the noise-free case. This is not an unrealistic requirement given the nature of the acoustic spaces involved and the probable levels of the signal that will be presented.

8.8 Discussion

This chapter has demonstrated the success of the ML method on both speech and music for blindly estimating a number of acoustic parameters. The method has been shown to give estimation accuracies close to the subjective difference limens for a range of

acoustic parameters with speech and music using both with simulated and real impulse responses and demonstrated excellent estimation performance from in-situ orchestral recordings made in two concert halls.

For speech, estimates of Rt can be gained from the octave bands in the range of 250 – 8000Hz, with an average error of ±0.4s in the 250Hz band, and up to ±0.2s at higher frequencies. The method is blind save for one parameter, the segment length. This parameter must be chosen based on the probability that a suitable decay phase is available within a specified time frame, which is very much dependant on the signal source. It was found that a segment length of 3 minutes was suitable for monologue style speech, though this may differ for other styles of speech as the speed of talking and the lengths of gaps between words may differ. These properties may indeed differ between talkers, but it is not thought that this will impact significantly on the accuracy. Nine minutes of speech was available to conduct these trails. This is the minimum requirement, as only three segments were available and therefore the median estimate was computed from just three points. A better length would be 18 minutes as this would enable the median decay to be estimated from six decays.

The estimation accuracy for the parameters C_{80}, D, t_s and EDT using speech is lower than for Rt, but useful estimates were still gained over the octave bands 500 to 8000Hz. It is postulated that increasing the overall speech length to the previously suggested 18 minutes will significantly increase the accuracy of these parameter estimations. This is because much of the error is due to the stochastic nature of the signal itself. Biasing factors due to the persisting tails of speech utterances in the decay phases have been shown to be reduced by increasing the segment length to 3 minutes.

For music, the accuracy of estimation for all parameters is less than that for speech, but good accuracy for Rt was demonstrated (±0.3) in the octave bands between 1000Hz and 4000kHz. Less accurate, but still useful values were achieved for EDT, C_{80} and t_s in these bands.

Music is a less suitable signal for estimation purposes for a number of reasons:

- The tails of musical notes generally have a longer decay time than the tails of speech utterances,

- The frequency content of music is focused around the equal temperament scale and the key of the music resulting in uneven excitation.

The negative impact of these two factors is reduced by increasing the segment length and the overall length of the recording. There is a much smaller chance of finding regions in a music signal with a decay profile appropriate for ML than there is in a speech signal of the same length, however, such regions do exist. The algorithm automatically searches for sharp, impulsive like, quickly decaying parts of signal followed by a period of silence. In orchestral music these regions can be at the end of a piece of music, for example Mozart often ends his pieces with a sequence of three notes with the last note brought to a sudden sharp stop. Other regions in the music may exist containing loud orchestral 'stabs' or staccato notes. Musicians have a term known as '*Sforzando'* which describes a strong, sudden accent, which is often abbreviated as ***sf***, ***sfz*** or ***fz***. It is likely that music with these '*Sforzando'* passages will be ideal for the ML algorithm. Different pieces of music will have different ideal segment lengths, but in this work it was found that for a ninety minute recording, a segment length of about ten minutes was generally sufficient.

The segment length and overall length of recording have significant impact on the accuracy of the method, in general the longer the recording and the longer each segment is, the better the algorithm will perform. However this is not realistic as only finite lengths of signal are generally available. As previously mentioned, segment lengths of three and ten minutes for speech and music respectively were found to be a good rule of thumb, however some signals may perform better with different segment lengths; for example very long segments with slow legato music. The bootstrap based confidence limit estimation on the decay curve is one way of blindly choosing the ideal segment length. By analysing the spread of parameter values indicated by the 95% confidence limits on the decay curves, it is possible to look at a range of segment lengths and choose one that provides a minimum variance for Rt thus blindly choosing a suitable segment length for the signal.

The method is limited by the level of noise present on the recording. For a live orchestral recording, stationary noise is not generally a problem as concert halls are designed to have low background noise levels from both external sources and internal stationary sources, such as air conditioning. In other situations where stationary background noise is a problem it was shown that a background signal to noise ratio of at least 25dB is sufficient for measurement purposes. Non-stationary noise sources generated by the audience themselves, such as coughing or even clapping, are thought to be a bigger problem. The last note in a piece is often ideal for the ML algorithm, but an audience will generally clap after a short pause, (provided the orchestra are good enough of course!). If the pause is too short then this decay phase cannot be used for estimation purposes. Despite this limitation, the method will automatically fail to produce an answer, if no acceptable decay phases with sufficient dynamic range be present. This enables a level of confidence to be placed in the noise immunity of the method.

In summary, by utilising a multi-decay model of sound decay and a framework to yield decay curve estimates, this maximum likelihood method provides practitioners with a method for determining in-use acoustic parameters, from passively received speech and music signals, to within a reasonable accuracy. Results from this chapter have been published in the following publications; [7, 9, 73, 84].

9 IMPROVING THE EARLY SOUND-FIELD ACCURACY USING THE CEPSTRUM

The ML model uses an unrealistic model for early reflections patterns. By utilizing an echo detection algorithm the early sound field estimation accuracy can be greatly improved. The motivation for this is not only the goal of improving parameter estimation accuracy, but also to provide a subjectively accurate blind-estimate of the acoustic impulse response for the purposes of auralisation. An aurally-accurate recreation of the room impulse response, blindly estimated from speech and music signals would provide producers of music, television, radio and other media, with the ability to mimic the acoustics of a room from a pre-recorded sound. Sounds could then be seamlessly mixed together and perceived as if produced under the same acoustic conditions. This chapter describes the design of an algorithm to blindly detect the time locations of individual reflections by performing a cepstral analysis of reverberated speech signals. The algorithm identifies the pattern of early reflections and these detected reflection locations are then used in combination with the ML estimate of the decay curve to produce a new decay curve estimate which demonstrates more realistic early reflection patterns. This new estimate, rather than a continuous exponential function as estimated by the ML method in chapter 7, has reflections in discrete locations. The new decay is produced by optimising the amplitude of each discrete reflection, by minimising the mean squared difference between its Schroeder curve and the ML Schroeder curve. The resulting parameters indicate that the estimation accuracy is not increased substantially, but informal listening comparisons indicate the impulse response is subjectively much closer to the original than the ML estimate.

9.1 Limitations of the maximum likelihood method

The ML method approximates the fine structure of the impulse response as a random Gaussian variable. This loss of detail may have an impact on the accuracy that can be achieved for a number of room parameters. Parameters such as C_{80} and D are both calculated as a ratio of energy levels in truncated regions of the impulse response. This truncation means that the parameters can be particularly sensitive to strong reflections occurring close to the truncation points. It is postulated that by detecting the time

locations of the early reflections these problematic reflections can be identified and they may contribute to a more accurate parameter estimate.

While the ML estimated acoustic parameters have been shown to be relatively accurate for many examples, subjective comparisons between the original reverberant signal and a signal convolved with a ML RIR estimate indicates they are subjectively very different. This is despite the acoustic parameters indicating they should sound very similar. This is because the early sound field and its structure is of prime importance in the human subjective response and the ML estimate has a very unrealistic early sound field which includes, amongst other omissions no initial time delay gap before the 1st order reflections and no increase of reflection density with time.

9.2 Introduction to the cepstrum

To blindly estimate the early reflection pattern of the impulse response, the locations of the early order echoes must be estimated from the reverberated speech signals. The cepstral signal processing technique has been used for similar applications in the past and is a signal processing technique that is particularly suited to the detection of echoes. It was first developed by Bogert *et al.* [88]. For a review on the method and its related algorithms please refer to [89]. The complex cepstrum ($\hat{x}(t)$) is defined as the inverse Z-transform of the log Z-transform of a signal;

$$\hat{x}(t) = Z^{-1}\left(\log\left(Z\left(x(t)\right)\right)\right) \qquad (9\text{-}1)$$

The cepstrum is a time domain signal with some very useful properties. In the Z-domain, when signals are convolved with each other, as in the case of a signal exciting an acoustic space, the convolution operator becomes a multiplicative operation. By taking the logarithm of the Z-transform, the convolution operation becomes an additive one, and when performing the inverse Z-transform, the additive relationship between the two components remains. The cepstrum exhibits homomorphic properties, meaning that the operation has the characteristics of algebraic addition. Echoes show up as peaks in the cepstrum.

There are two different 'flavours' of cepstrum; the power cepstrum as used by Bogert *et al.* [88] and the complex cepstrum which was developed by Oppenheim and Schafer [51]. The complex cepstrum uses the complex logarithm of the Z-transform, while the power cepstrum uses the real logarithm of the power spectrum (i.e. the squared magnitude of the Z-transform of the time signal). The complex cepstrum retains phase information and in theory is completely reversible. However, because the complex logarithm can be multi-valued, its calculation requires the phase to be unwrapped, this operation may cause errors or inaccuracies. The power cepstrum on the other hand does not contain phase information and the spectrum is always positive and therefore the complex logarithm and phase unwrapping is not required. Childers [89] comments that the echo detecting ability of the power cepstrum can often be superior to the complex cepstrum, due to phase unwrapping errors and a linear phase contribution often tends to mask the echo delay in the complex cepstrum. Therefore it is proposed that the power cepstrum be used to detect the location of echoes that form the early sound field. This is also the approach adopted in references [41] and [90].

The power cepstrum ($\tilde{x}(t)$) is represented by equation (9-2).

$$\tilde{x}(t) = Z^{-1}\left(\log\left(\left|Z\left(x(t)\right)\right|\right)\right) \qquad (9\text{-}2)$$

When evaluating Z-transforms, it is standard practice to describe the Z-transform in terms of its 'poles' and 'zeros'. A Z-transform is often expressed as a ratio of two polynomials and the solutions to the polynomials are the poles and zeros (the denominator solutions are the poles and the numerator solution are the zeros). The Fourier transform of the signal represents the transfer function of the system and is a special case of the Z-transform, where the Z-transform is evaluated around the unit circle (a circle with radius one). The position of the poles and zeros, relative to the unit circle, determine not only the magnitude and phase response of the system, but also its causality and stability.

The power cepstrum assumes that the system is minimum-phase; a system is minimum-phase if process and its inverse are both stable and causal. For a system to be minimum-phase, all its poles and zeroes must be inside the unit circle. This is problematic for room impulse responses as they are generally mixed-phase (mixed

phase means some zeros are within the unit circle while some are outside). When estimating the power cepstrum, the maximum-phase zeros (zeros outside the unit circle) are reflected from outside to inside the unit circle. These spurious zeros contribute inaccuracies to the estimate of the impulse response. To account for this, exponential windowing can be used. Multiplying the signal by an exponential has the property that the Z-transform of that signal is also multiplied by an exponential. Therefore exponential windowing has the effect of moving poles and zeros inwards, towards the origin. Choosing the exponent carefully can enable maximum phase zeros to be moved within the unit circle making the signal minimum phase. After the cepstral operations, inverse exponential windowing moves the poles and zeros back to the correct location. Multiplying each window by a decaying exponential does not introduce any distortion in the convolution relationship due to the following relationship.

$$e^{-\frac{n}{\tau}}.(s[n]*h[n]) = e^{-\frac{n}{\tau}}.s[n]*e^{-\frac{n}{\tau}}.h[n] \qquad (9\text{-}3)$$

Where s[n] represents a speech signal, h[n] the RIR and $e^{-\frac{n}{\tau}}$ the exponential window. To demonstrate this, a 20^{th} order Butterworth, infinite impulse response (IIR) bandpass filter is designed using the Matlab butter command. The pass-band is set to between 0.5 and 0.8 radians/sample. This yields a frequency/phase response as shown in Figure 9-1. The impulse response is then multiplied by the exponential $e^{-0.5n}$ (where n is sample number from 0 to *N-1*). The resulting poles and zeros are shown in Figure 9-2. By multiplying a signal with a decaying exponential, the poles and zeros of the Z transform are moved inwards. There is trade off between having a sufficiently quickly decaying exponential to move a distant zero inwards and the loss of numerical accuracy due to floating point inaccuracies with very small values, due to the tail of the exponential window which moves the zeros towards the origin.

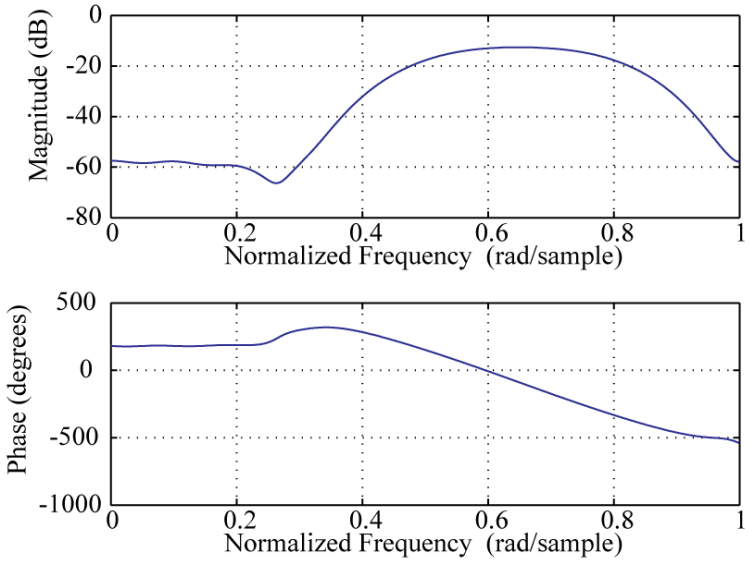


Figure 9-1. Frequency and phase response of the Butterworth filter discussed in the text

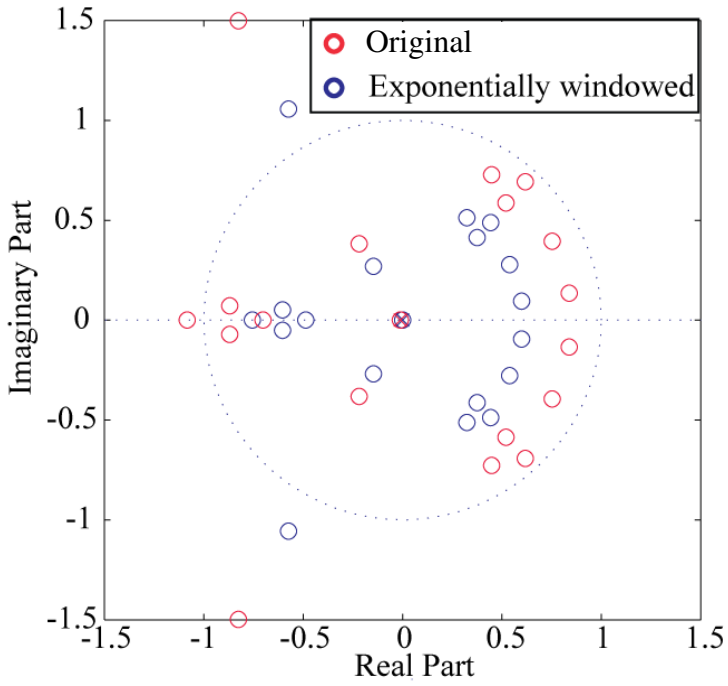


Figure 9-2. Poles (x) and zeros (o) of a band pass Butterworth filter. Blue markers show the poles and zeros after multiplication with exponential window and red markers show them before

9.3 Detecting echoes in speech signals using the cepstrum

Bees *et al*. [90] developed a method of speech dereverberation that utilised cepstral averaging to yield acoustic impulse response estimates from speech signals before using a least-mean-square (LMS) filtering operation to deconvolve the RIR from the speech. Using the homomorphic properties of the cepstrum, Bees windowed the reverberant speech signal and averaged the cepstrum from a sequence of windows. The cepstral speech component can be considered to be random with zero mean and therefore, when averaging the cepstrum, the speech component will be reduced towards zero. Cepstral components due to the room impulse response are stationary and therefore persist in the averaging. After gaining an estimate of the cepstrum of the RIR, an inverse cepstral transformation then yields an estimate of the room impulse response.

Bees encountered a number of sources of error.

1) Estimating the complex cepstrum is difficult and involves phase unwrapping. To counteract this, he multiplied each window by an exponential weighting function (as described in the previous section). This enabled Bees to avoid phase unwrapping and calculate the cepstrum using the log magnitude spectrum of the exponential weighted windows.

2) By windowing the speech signal, the end of the RIR is truncated. This sudden discontinuity produces unwanted effects in the cepstrum. Exponential windowing, mentioned in point 1), reduces the error due to truncation.

3) For each selected window there is some residual signal that is the result of the reverberant tails due to excitation prior to the start of the window. Bees reduced this error by locating the start of each window at the end of a period of silence and the start of a new utterance.

The next section describes some initial results from experimenting with this procedure.

9.3.1 Detecting echoes using the cepstrum

As a proof of concept, 90s of speech was convolved with a simple impulse response which contained a Kronecker delta function of magnitude 1 at time 0 and another of magnitude 0.5 at 50ms. This was then passed through a 1kHz octave band filter. This impulse response is shown as the blue line in **Error! Reference source not found.**. The resulting speech signal was windowed using rectangular windows with 90% overlap. Each window was multiplied by the following exponential function, w, in equation (9-4).

$$w(t) = e^{\frac{-t}{\tau}} \tag{9-4}$$

where τ was set to 0.1 and t is time. The cepstrum can then be computed from the log of the magnitude of the spectrum as the impulse response has (hopefully) been forced to be minimum phase. The cepstrum is given by:

$$\hat{x}(t) = Z^{-1}\left(\log\left(\left|Z\left(w(t).x(t)\right)\right|\right)\right) \tag{9-5}$$

The average cepstrum is then computed for all windows and transformed back to the time domain using the following function. Note that the exponential window has been removed by dividing the function by *w(t)*.

$$x(t) = \frac{Z^{-1}\left(e^{Z(\hat{x}(t))}\right)}{w(t)} \tag{9-6}$$

The result is plotted in **Error! Reference source not found.** as the red line. The Figure shows that the cepstral processing has correctly estimated the time delay of the echo.

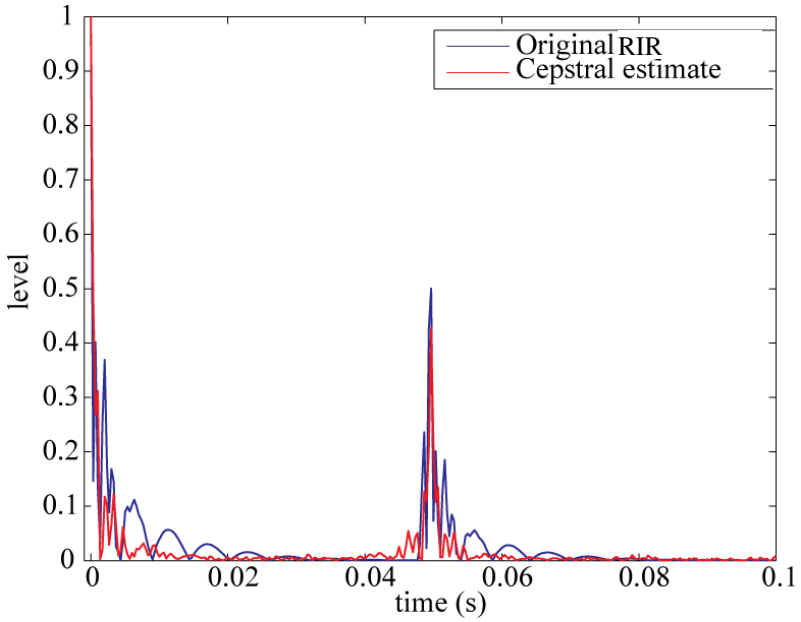


Figure 9-3. Demonstration of echo location using cepstral averaging on speech.

Although the time position is correct, the magnitude of the echo is not. This is due to segmentation errors and the remaining speech component in the cepstrum. Speech produces cepstral peaks at times related to its fundamental period, and this may equate to a cepstral component at or below 8ms [89]. Due to this signal in the cepstrum at low time positions, when converting from the Z-domain to the time domain, this becomes an excess of energy at the beginning of the impulse response estimate, which causes a discrepancy between the direct sound energy level and the echoic energy level that this method estimates.

The method employed by Baskind [41] and Bees [90] used short, exponentially windowed speech segments where the window was located at the start of speech utterances (after a period of silence). An automatic segmentation method has been devised, based on the envelope segmentation algorithm [4] (used to identify decay phases for the ML method in Section 7.3.1). Only the end points of each decay phase from the segmentation procedure are required and these points indicate the end of a period of silence and the start of a speech utterance.

Exponential windows of length T_{win} were applied to the detected start of speech utterances, where T_{win} was chosen to be 500ms to ensure capture of sufficient length of

the early sound field. The exponential decay constant was chosen to be $\tau=T_{win}/5$ as recommended in Baskind [41].

Now that a number of windows are available for computing the cepstrum, it is apparent that a number of these windows exhibit very high segmentation errors due to the persisting presence of reverberant tails from utterances prior to the window in question. In order to improve the quality of the cepstral estimate, a method of locating the 'best' windows is required. The level of segmentation error present can be estimated by looking at the level of the signal that is present immediately prior to the selected window. It is logical to assume that windows, whose preceding signal levels are low, have lower segmentation error. Therefore, the ratio of the average energy in the selected 500ms window, to the average energy in the 500ms prior to the speech utterance is calculated. The windows with the highest ratio are used to calculate the cepstral average (the upper 25% of the distribution of ratios was empirically found to be appropriate). The utterance onset detection algorithm is demonstrated in Figure 9-4. This highlights how the algorithm reduces segmentation error.

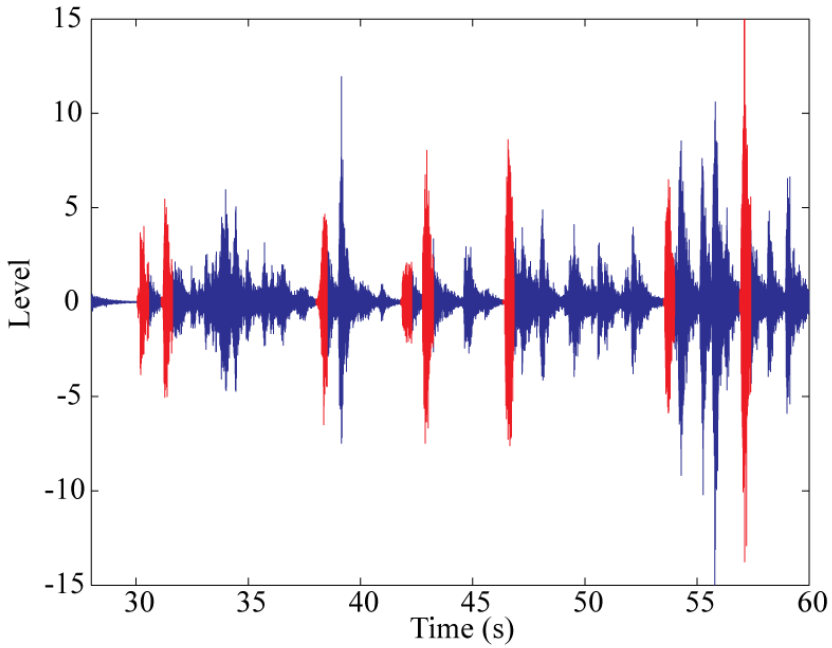


Figure 9-4. Selection of speech utterance onsets, the 500ms sections of speech onsets are highlighted in red

Figure 9-5 shows the result of the cepstral estimation of a room impulse response from the reverberated speech signal. The plot compares results generated from cepstra calculated using only speech onsets (Figure 9-5 a), and results where the cepstrum was estimated from overlapped, averaged windows taken over the whole speech signal (Figure 9-5 b). A marked improvement in the accuracy of the estimation is found by using only the speech onsets. Not shown on the plot, is the large discrepancy between estimated direct sound level (very high) and the subsequent reflection level (very low). This is because part of the speech cepstrum remains in the estimate and because of the loss of non-minimum phase components of the impulse response, both of which contribute to extra components being estimated at very early time locations. This discrepancy is discussed further in [41]. The exponential window will most likely be insufficient to capture all maximum-phase zeros and move them within the unit circle. This loss of non-minimum phase information causes the energy levels of the early reflections to be miscalculated. However as discovered by Baskind, although the energy levels of reflections are incorrect, the time locations of the echoes are correct.

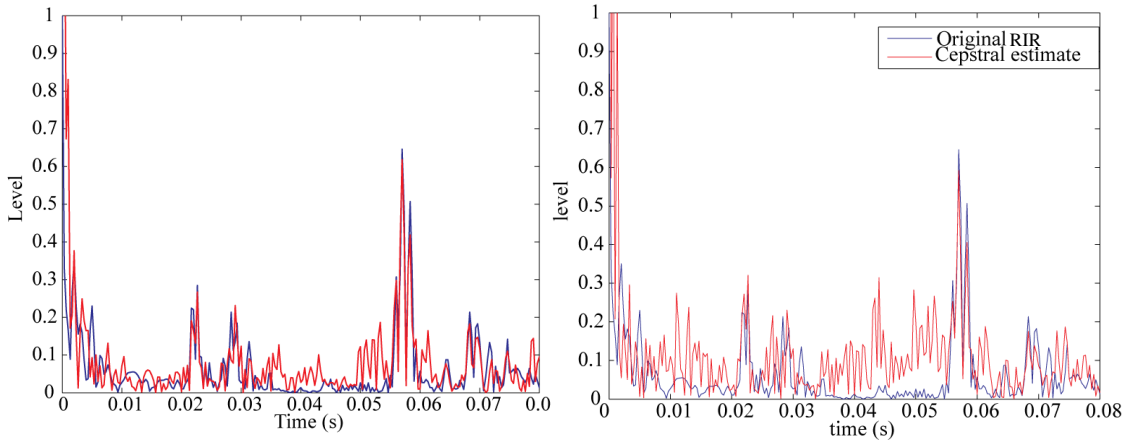


Figure 9-5. Estimated early sound field using averaged cepstra from 500ms windows whose starting point is automatically detected as the initial onsets of speech utterances (a) and averaged cepstra from all 90s of speech using 95% overlapping 3sec windows (b). Plots have been scaled for comparison purpose.

9.4 Combining the cepstral and ML methods

The maximum likelihood method (Chapters 7 & 8) has been shown to be reasonably accurate for Rt and EDT (both estimated from the Schroeder decay curve). When

combining the cepstral and ML methods it is desirable that the new combined ML and cepstral decay curve retains the same overall decay rate and features. Combining the ML and cepstral methods requires a number of steps. These are first summarised in the following list and then the steps will be explained in more detail in subsequent sections.

1. Perform ML estimation of the decay from a reverberated speech signal as described in Chapter 7 (method c).

2. Estimate the early portion of the room impulse response using windowed cepstral averaging as described previously in Section 9.3

3. Process the resulting estimate of the early part of the impulse response to identify the locations, in time, of each echo (Section 9.4.1).

4. Using the detected echo locations, calculate a new decay estimate that exhibits the pattern of discrete reflections at time locations calculated in step 3. The amplitude of these reflections is optimised so that mean square error between the Schroeder curve of this new estimate and the ML estimate is minimised (section **Error! Reference source not found.**).

9.4.1 Identifying echo locations in the cepstrum

A decision-making frame-work is developed that uses the cepstral estimate of the early portion of the RIR from Section 9.3 to identify the locations of individual echoes. To achieve this, a two step approach is adopted; (1) the long term decaying trend of the cepstral RIR estimate is removed, and (2) the strongest reflections are located. The first step makes the second step easier.

To remove the decaying trend of the RIR estimate, a maximum likelihood fit is applied to the RIR estimate. The resulting decay envelope is used to remove the long term decaying trend of the estimate, by dividing the cepstral RIR estimate by the ML decaying envelope.

The objective of this task is to remove, as best as possible, the overall decaying nature of the RIR so that the echo location task will be more successful.

The resulting function with the overall decaying trend removed and normalised to the level of the strongest reflection is shown in Figure 9-6. The ML fit is used as opposed to other methods such as a least-squares line fit, because the ML model has within it, a model of the physical system being measured i.e. the decay of room reflections. It is therefore reasonable to assume that this statistical model enables the ML decay estimate to provide a better estimate of the overall decay of energy than other methods which don't include a model of sound decay. A more reliable estimate of the decay trend means that after application to the RIR estimate, the detection and location identification of individual reflections is easier.

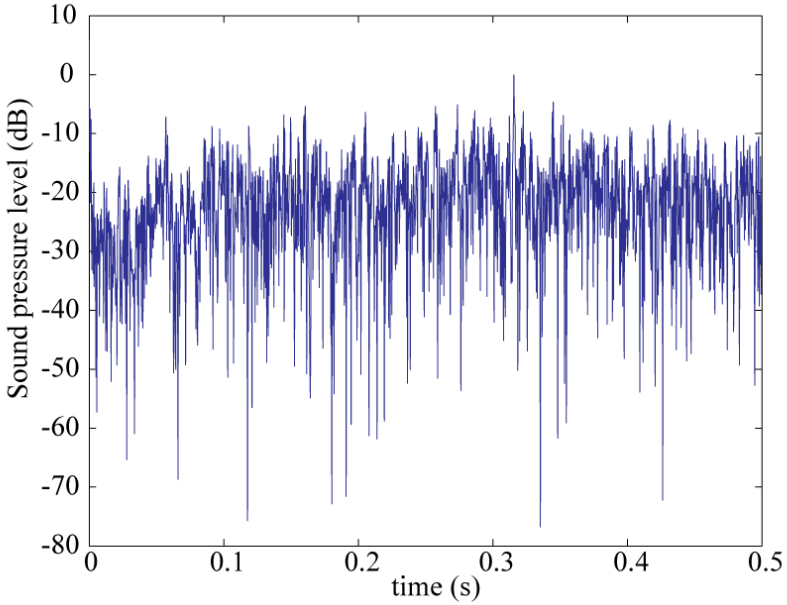


Figure 9-6. Estimated impulse response using the cepstral method with overall decaying trend removed, normalised to strongest reflection

The next task is to locate the strongest reflections. Now that the long term decaying trend has been removed it is a simple matter to locate the strongest reflections. A threshold is set and all reflections that exceed this threshold are identified as individual reflections. For this work, the threshold is set an empirically determined -15dB with all reflections above this level being identified as echo locations. Using these echo locations, a masking function is created, which is 0 for all time except in the sample locations where an echo has been positively identified. This masking function is shown in Figure 9-7.

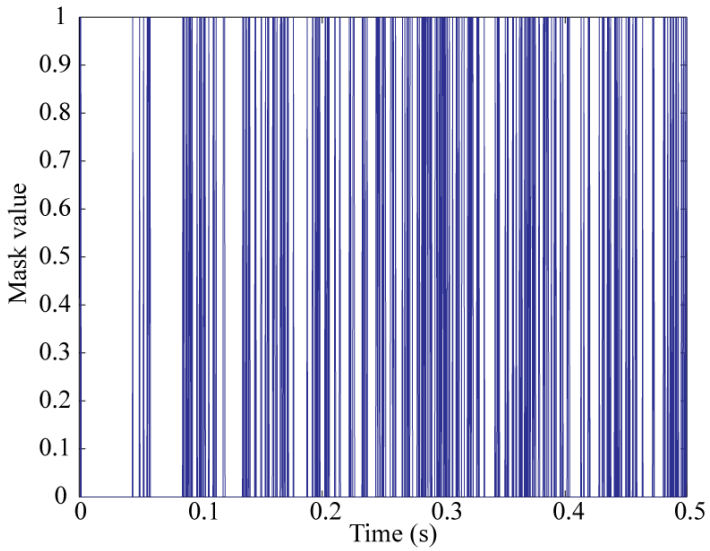


Figure 9-7. Masking function showing the temporal pattern of the early reflections

9.4.2 Assigning levels to detected reflections

The cepstrum method provides estimations of the temporal distribution of reflections in the early part of the impulse response. It does not, however, provide useful information regarding the energy level of these reflections. The ML method provides no information regarding the temporal distribution of the reflections (besides the inaccurate assumption of a Gaussian distribution), however, it does provides an accurate estimate of the smoothed energy decay of the impulse response with time. This energy level is a function of the decay curve model and the estimated variance of the Gaussian variable in the sound decay model. Therefore by combining these two methods it should be possible to assign the correct energy level to each cepstrally detected reflection.

The random Gaussian variable in the ML method is a stationary random process whose distribution does not change with time, whilst, in-fact the RIR fine structure is non-stationary because the distribution type changes with time. Initially the fine structure has a very non-Gaussian distribution with a high kurtosis, (similar to a Laplace distribution) but gradually the distribution changes to become more Gaussian-like as the reflection density increases. A more appropriate ML function incorporates this change in distribution with time so it could be taken into account when estimating the energy decay. However, it is assumed sufficient to utilise the ML estimation of energy as, while unlikely, it is still possible for a random Gaussian variable to produce a sequence

similar to the fine structure of the RIR. Using this assumption it is proposed that the new ML/cepstrum decay model will utilise a Gaussian model of fine structure for the late decay and an alternative distribution with the same variance (or average energy) as estimated by the ML method, for the early region fine structure. The chosen distribution for the early reflection structure is taken from the cepstral estimated masking function in Figure 9-7.

When combining the two methods a transition point needs to be chosen between the two distributions, i.e. between the early sound field and the late sound field. The ideal point for the change over from the early to the late decay would be the mixing time. However no reliable estimate of this is available. Using the masking function to calculate the point where 10 reflections are present in a 24ms window was one option but it was though that this value would be too sensitive to error. A transition point of 100ms was chosen so that the important first 80ms of decay is encompassed, plus an extra region in case there are any strong reflections occurring just outside the first 80ms, which may heavily influence the fixed integration limit parameters (C_{80}, LG). The combined method produces the response pictured in Figure 9-8.

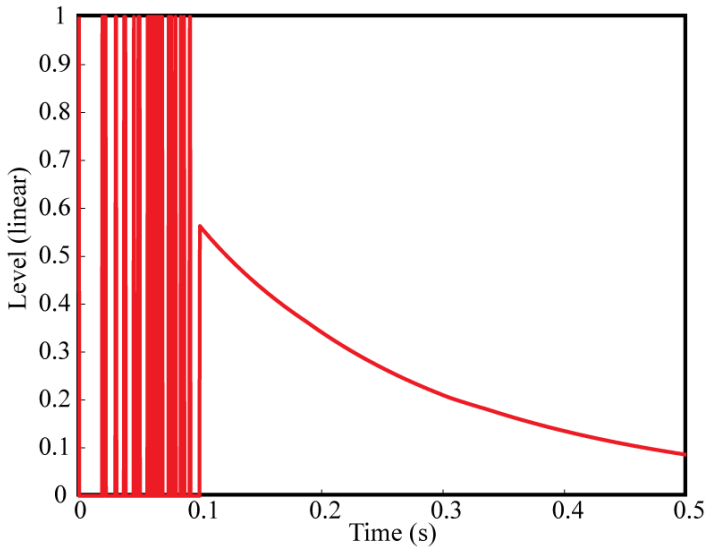


Figure 9-8. Combined ML and cepstral decay estimate prior to individual reflection level optimisation.

In order to adjust the energy levels of the early reflections so that they match the energy decay of the ML decay curve, the energy levels of each of the early reflections are optimised. The energy level of each reflection is adjusted so that total squared

difference between the Schroeder curve of the combined ML/cepstrum estimate and the ML Schroeder curve is minimised. This is expressed as the minimisation of the cost function $E(\theta)$,

$$E(\theta) = \int_t^\infty {h_{ml}}^2(t)dt - \int_t^\infty {h_{ml+ceps}}^2(t,\theta)dt \qquad (9\text{-}7)$$

where h_{ml} is the ML decay curve estimate, $h_{ml+ceps}$ is the combined estimate and θ is a vector containing the level of the reflection at each of the estimated early reflection locations detected by the cepstrum method.

The parameter θ is 1 for all reflections in Figure 9-8. This is chosen as the starting point in the parameter search space. An unconstrained gradient based optimisation method was chosen as there are no absolute limits on the level a reflection could have. The optimisation algorithm used was from the Matlab optimisation toolbox function 'fminunc' which uses an 'interior-reflective Newton method'. Figure 9-10 shows the decay estimate after optimisation.

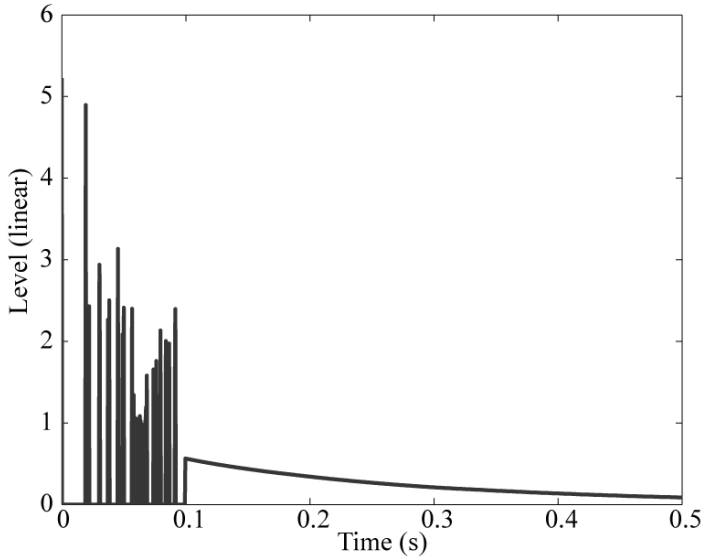


Figure 9-9. Combined ML and cepstral decay estimate.

Figure 9-10 compares the Schroeder curves of the ML estimate and the new ML/cepstrum estimate. It shows how the overall energy decay rate of the new ML/cepstrum estimate follows the same trend as the original ML estimate, the

difference being that in the early sound field the energy occurs in discrete locations, as in the true impulse response.

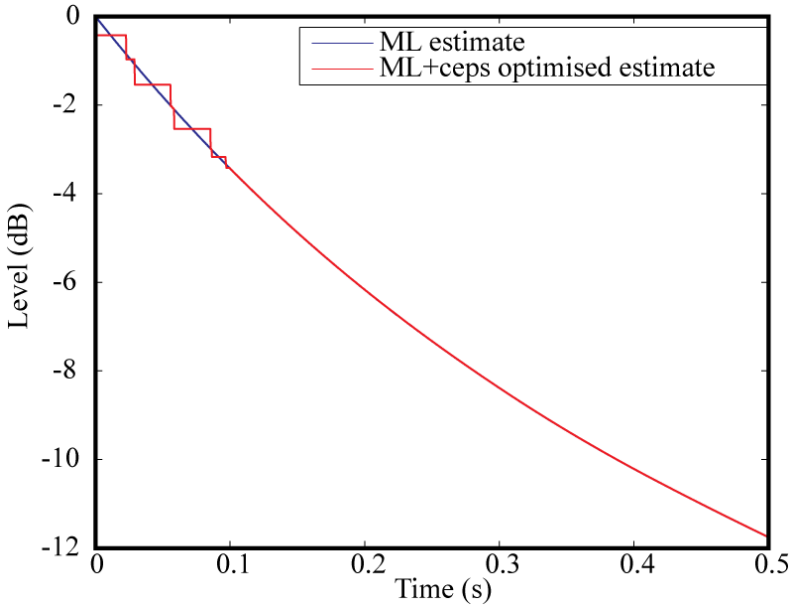


Figure 9-10. Comparison of ML estimated Schroeder curve and ML-cepstrum optimised curve.

To complete the RIR for auralisation purposes, after the transition point, the estimate is used to modulate a Gaussian noise source. This blindly estimated RIR, filtered using a 1kHz octave-band filter, is compared with the original RIR in Figure 9-11. It can be seen that the method has located some of the strongest reflections. While the impulse responses are not identical, the fact that the two strongest reflections and the direct sound (as indicated on the plots) are close to the correct time locations and energy levels, will greatly enhance their subjective similarity. (the magnitude of the RIRs is plotted in Figure 9-11)

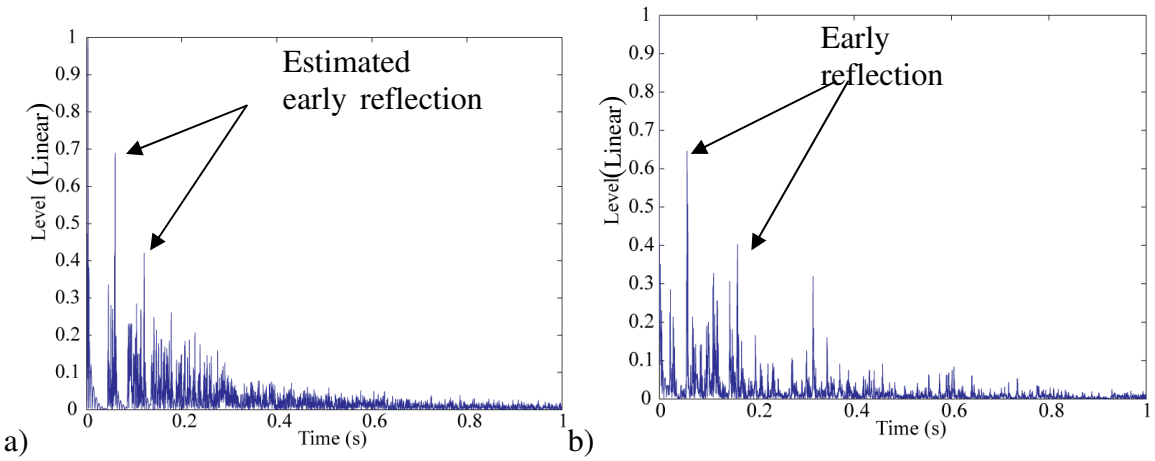


Figure 9-11. Magnitude of the impulse response estimated using combined cepstrum/ML method (a) and true impulse response (b)

A comparison of the decay curve estimated using the enhanced cepstral/ML method and the true decay curve is shown in Figure 9-12. The new estimated decay curve follows the decay pattern of the original impulse response more closely, especially in the first 6 or 7dB of decay. The differences between the two plots in Figure 9-12 are due to the fact that the cepstrum echo detection algorithm has failed to locate a number of low level reflections. When low level reflections are either missed or misidentified this has an impact on the energy estimation accuracy of the higher level reflections. This is because the optimisation uses the Schroeder backwards integrated decay curve; if a reflection is missing, the energy contained in that reflection is added to the energy level of subsequent reflections. The location of all early echoes, including low level ones is important to the success of this method.

The cepstrum algorithm was quite successful at correctly locating the strongest echoes. However, if the ML estimated decay curve differed substantially from the true curve, the estimated level of the reflections were incorrect. This was a particular problem when residual musical notes or utterance caused an overestimate of the energy level along the decay curve. Additionally the ML decay curve is a smoothed estimate of the true decay and as such, some of the features are lost in the estimation. These lost features can have a significant impact on the accuracy of the level of each estimated reflection. Overall the combined ML/cepstrum method is found to be more successful

when the early sound field is populated with a small number of strong reflections and when a reliable estimate of the decay curve has been made.

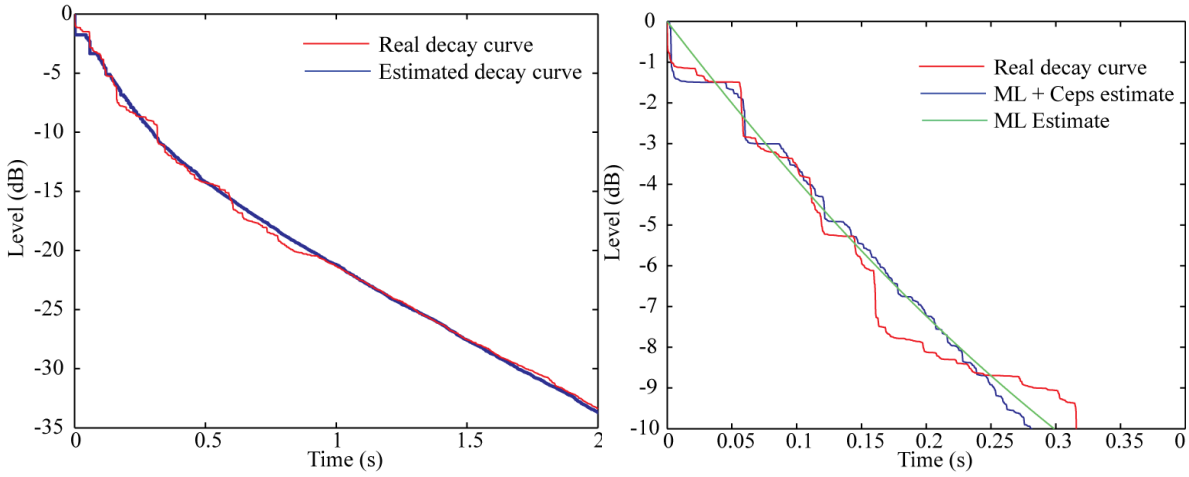


Figure 9-12. Impulse response estimated from reverberant speech using combined cepstrum and ML method.

The strength of the ML/cepstrum method lies in the combination of the two methods. The ML method provides an estimate for the smoothed decay energy of the RIR while the cepstrum method ensures the energy is located in the correct places. The cepstrum/ML method ensures that overall the rate of energy decay is the same as the ML estimate, but forces the energy to be located at discrete points in time. One way of viewing this process is in terms of random variables (where the random variable in question is the fine structure of the RIR). The ML method provides an estimate for the variance of the random variable while the cepstral method provides an estimation of the probability distribution of the random variable.

One weakness in the approach is that the method relies on an accurate ML estimation of, in particular, the early sound field energy decay. As discussed in early chapters, the early sound field accuracy is known to be limited for the ML estimation, partly because of the unsuitability of the ML model for the early sound field but also because of the persisting presence of sounds in the decay phases and the variance in the excitation between impulsive and sustained excitation. Inaccurate early sound field estimation by the ML method will cause the levels of the reflections to be incorrectly assigned in the ML/cepstrum estimate. An improvement to the ML method may be achieved by

utilising the cepstral estimated temporal distribution of the echoes within the likelihood optimisation. It is postulated that by improving the model appropriateness the fit will improve and hence parameter estimates should also improve, however further investigation would be required to confirm this.

Another weakness of the approach lies with the echo detection algorithm. The correct identification of the location of the echoes is key to gaining any improvement in the estimated decay curve. The cepstrum method is sensitive to non-minimum phase RIRs and although some care has been taken to treat any non-minimum phase zeroes by performing exponential windowing, any remaining zeros outside the unit circle will cause the algorithm to identify spurious echoes. As previously mentioned, the cepstrum method works best when presented with strong individual echoes. It is likely that in real rooms, particularly concert halls, the early sound field will not contain as many individual strong reflections as demonstrated by some of the simulated responses but many lower level ones (because of strategic positioning of diffusers). This will deteriorate the echo algorithm's detection performance because, as discussed previously, the algorithm functions best when presented with a small number of strong reflections rather than many lower level reflections. Additionally it is known that as the reverberation time increases the performance of the echo location algorithm will deteriorate, therefore the combined ML/cepstrum method will perform poorly in very reverberant spaces.

9.4.3 Parameter estimation accuracy using the enhanced ML /cepstrum impulse response estimate

This section investigates the impact that the hybrid method has on the accuracy of the resulting acoustic parameters. Parameters which perform an integration over a short rectangular window such as D (50ms) and C_{80} (80ms) can be influenced by the presence of singular discrete reflections close to the 50 or 80ms integration limit. The hybrid method has been used to estimate the parameters from 100 simulated room impulse responses convolved with 90s of anechoic speech and then the parameter estimation accuracy is compared with that of the Maximum likelihood method. The threshold for the detection of a reflection was pragmatically changed to -12dB, as -15dB yielded many spurious reflections. The results for the ML method and the ML/cepstrum

method are presented in Table 9-1. These figures show that while the accuracy for the Rt decreases slightly, the accuracy of the C_{80}, D and t_s all increase marginally. A paired t-test on the same data indicates that while the difference between the two methods give different parameters at a significant level of 95%, the magnitude of the difference is tiny compared to the difference limens for each parameter. Table 9-2 details the performance for real room impulse responses and highlighting a slight improvement in parameter estimation accuracy. The differing performance between the real and simulated rooms is because the real room responses were generally much less reverberant and both the ML and the cepstrum methods are more accurate for less reverberant spaces.

Average parameter error	Rt (s)	EDT(s)	C_{80} (dB)	D (%)	t_s (s)
Hybrid ML+cepstrum method	0.13	0.25	1.54	6.9	0.012
ML Method	0.12	0.25	1.59	7.4	0.012

Table 9-1. Comparison of averaged absolute parameter error for 100 simulated room impulses responses comparing the ML method with the hybrid ML+cepstrum method.

Average parameter error	Rt (s)	EDT(s)	C_{80} (dB)	D (%)	t_s (s)
Hybrid ML+cepstrum method	0.02	0.10	1.37	5.06	0.007
ML Method	0.03	0.08	1.45	5.87	0.008

Table 9-2. Comparison of averaged absolute parameter error for 18 real room impulses responses comparing the ML method with the hybrid ML+cepstrum method.

9.4.4 Comparison of the ML and the ML/cepstrum methods

While the lack of meaningful improvement in parameter accuracy is disappointing for measurement applications, for auralisation, there is a distinct audible difference between the estimated impulse responses. In actual fact the lack of improvement in parameter accuracy is unsurprising becuase the ML/cepstrum decay curve was optimised to be as close to the ML decay curve as possible (in a least means square sense). Therefore the new impulse response estimate exhibits all of the ML likelihood estimated subjective parameters, but with a more realistic early reflection structure. In order to confirm the hypothesis that the new ML/cepstrum method produces RIRs that are more subjectively similar to the original RIRs (than RIRs estimated using the ML method), a set of time consuming, extensive subjective tests would be required. These subjective tests are

beyond the scope of the thesis and therefore a simple, objective comparison is carried out. A mean square difference between the RIR estimates and the original RIR is deemed to be an appropriate choice.

Table 9-3 shows the expected mean square error, of the ML and the hybrid ML/cepstrum impulse resposne estimates compared with the true impulse response. The values were calculated as the mean square difference between the magnitude of the true impulse response and the magnitude of the two impulse response estimates (prior to modulating a Gaussian source and hence an expected value of the mean square difference is computed). It is observed that the hybrid method produces a lower mean square error; this indicates that the hybrid estimate is closer to the original impulse response than the ML method alone.

Method	Mean square error
Hybrid ML+cepstrum method	0.0022
ML Method	0.0285

Table 9-3. Comparison of the mean square difference between the magnitude of the RIRimpulse and the magnitude of the two RIR estimates (ML and ML+cepstrum) calculated from 100 artificially simulated RIR estimates.

9.5 Discussion

This chapter has presented a way of combining the ML estimation of the decay curve, where the subjectively important strong early reflections are not identified, with a cepstral based estimate of the pattern of early reflections. The ML estimate provides the estimation of the level of reflections while the cepstral algorithm identifies the time location of the reflections. The method is most successful where the early sound field consists of a relatively small number of strong reflections and a reliable ML estimate of the decay curve is available. Further accuracy will be gained by improving the echo detection algorithm. Sophisticated blind source separation techniques could be utilised for this aim. Over-estimation of the ML decay curve caused by residual speech or musical notes can cause the estimated reflection level to be inaccurate. Informal subjective listening comparisons suggest that the cepstrum enhanced impulse response estimates are aurally much more comparable to the original impulse response than the

ML estimate. However, in order to compare fully, extensive subjective tests are required.

In order to extend this hybrid method to the use of music signals, further work is required. Music signals, in general, are less random than speech and therefore it is anticipated that a modified or new approach will be required to compensate for the residual cepstral components in the estimated impulse response. Additional further investigation is also required to investigate the performance of the methodology when using real impulse responses, where it is expected the echo detection performance of the cepstrum may be reduced, more sophisticated echo detection algorithms may be required to achieve satisfactory performance.

10 COMPARISON OF THE ENVELOPE SPECTRUM AND MLE METHODS

In this chapter the performance of the two methods of blind acoustic parameter estimation developed in this thesis are compared (envelope spectrum / ANN method, Chapter 6 and MLE, Chapter 7). Two datasets are used to examine the success of the methods, one based on speech or music convolved with simulated impulse responses generated by the geometric room acoustic model, and the second a set based on speech or music convolved with real measured room impulse responses, the datasets were made up of the same RIRs as used in previous chapters. The performance of the parameter estimation methods varies between these two datasets, and comparing the two sheds light on the robustness of the methodologies.

10.1 Speech

About ninety seconds of anechoic, male, narrated running speech [69] is used. By narrated speech, it is meant that the speaker is given a passage of text to read out aloud. This technique tends to slow down the rate at which the narrator speaks, and so gives more gaps between utterances where decays can be seen.

10.1.1 Simulated RIRs

Comparing the results from the ML method and the envelope spectrum using speech and simulated RIRs, the Rt accuracy is very similar, this is shown in Figure 10-1. To improve the accuracy of the methods further, especially at middling reverberation times, a longer section of speech can be used, or further averaging of the estimated parameters across many lengths of recorded speech can be done.

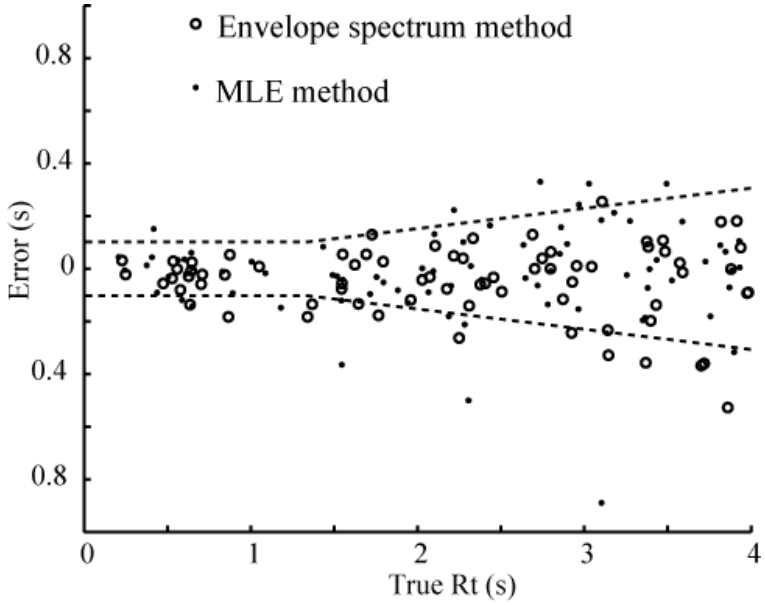


Figure 10-1. Error in parameter estimation versus the true value for reverberation time. The dotted lines indicate the difference limens. • *MLE and* **o** *envelope spectrum method. Validation set using simulated impulse responses and speech excitation. After Kendrick et al. [9].*

Figure 10-2 shows the results for clarity, as an example of one of the other parameters where estimation is more difficult. Clarity requires the accurate estimation of the energy arriving in the first 80ms after the direct sound, and the MLE method finds it difficult to precisely obtain this value because of the natural reverberance of many speech utterances. In contrast, a method based on machine learning such as the envelope spectrum method can learn to compensate for errors in the estimation. Hence for mid to large clarity values, the envelope spectrum method is more accurate than the MLE method. There are no results for the MLE method at very low clarity values, because the segmentation method failed to find any decay phases with sufficient dynamic range. These are very reverberant rooms where the start of the current utterance significantly masks the end of the decay of the previous utterance. To obtain an MLE estimation at such low clarity values requires larger time gaps between the utterances. As the clarity increases, the accuracy of the MLE estimations increases. While the envelope spectrum method provides estimations at these low clarity values, the accuracy of the estimation suffers because there is insufficient information about the late part of the decay, which is masked by subsequent utterances.

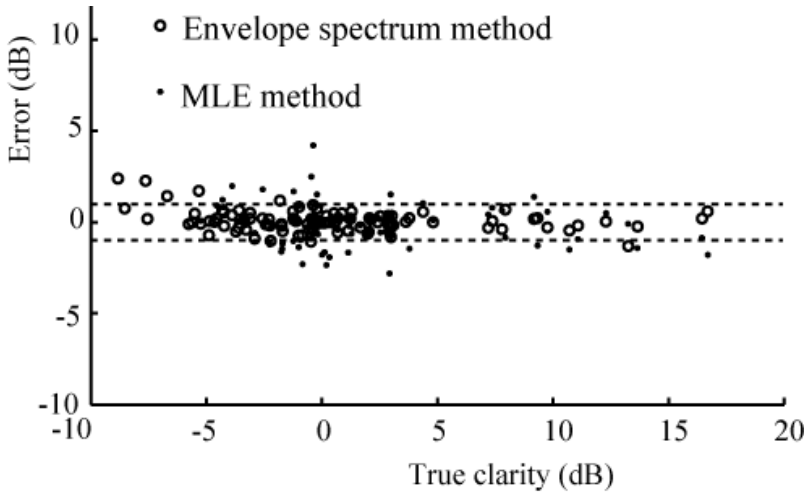


Figure 10-2. Error in parameter estimation versus the true value for clarity • MLE and o envelope spectrum method. Validation set using simulated impulse responses. Speech excitation. After Kendrick et al. [9].

Because the envelope spectrum method is based on experiential learning, the ANN learns to compensate for the over-estimation of the early reverberation that occurs due to the inherent reverberance of the speech utterances. Consequently, the envelope spectrum method is more accurate than the MLE method for parameters such as clarity, centre time and early decay time, but for reverberation time, there is little difference in the accuracy of the two methods.

10.1.2 Real, measured RIRs

Tests on real room measurements, however, yield a slightly different story to that found with the simulated impulse responses. This test set uses impulse responses measured in real rooms convolved with anechoic speech. Both the envelope spectrum and MLE methods provide comparable accuracy for all the parameters. Figure 10-3 and Figure 10-4 show the results for reverberation time and clarity for the real room measurements, which can be compared with those shown in Figure 10-1 and Figure 10-2 for the validation set using simulated impulse responses. For reverberation time, the MLE method becomes more accurate for the real room measurements in comparison to the validation set using simulated impulse responses. (Note, although the graphs may appear to imply this is also true for the envelope spectrum method, the difference is not statistically significant). Figure 10-3 shows that both methods are accurate to within one DL when using real room RIRs.

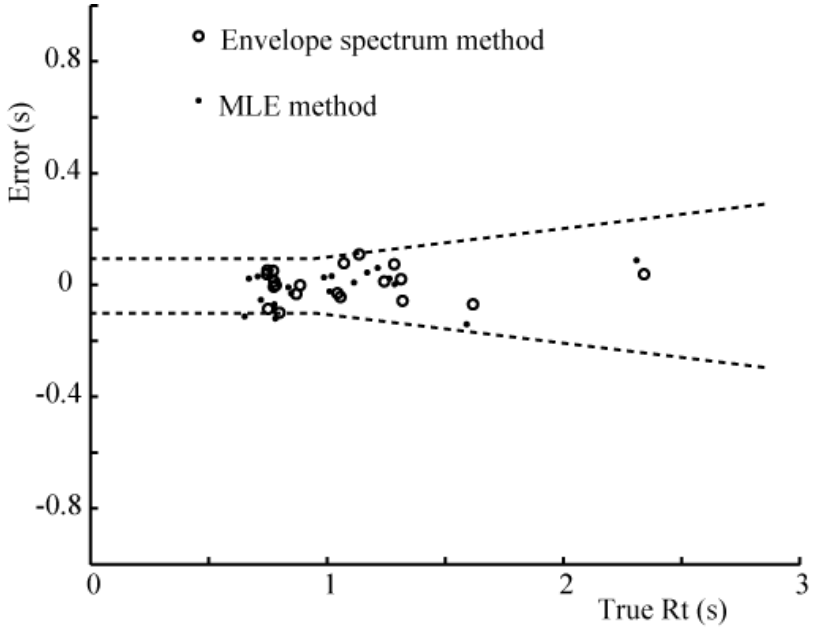


Figure 10-3. Error in parameter estimation versus the true value for Rt. • MLE and ***o*** *envelope spectrum method. Validation set using real RIR. After Kendrick et al. [9].*

For early decay time, the envelope spectrum method is slightly less accurate and the MLE method slightly more accurate when using the real measurements in comparison to simulated ones. For clarity and centre time, the MLE method gives similar accuracy for the real and simulated validation datasets, but the envelope spectrum becomes less accurate, and, furthermore, a bias error is introduced.

The loss in accuracy with the envelope spectrum method when estimating some parameters probably occurs because the simulated room impulse responses used to generate reverberated speech for training the ANN, are not completely representative of real room impulse responses, consequently, the data used for training and validation have some significant statistical differences. The introduction of a bias as shown with some parameters and illustrated in Figure 10-4, is good evidence for this. As an ANN works to minimise the mean square error, a well trained ANN should not generate a bias error, unless something is wrong, such as the test and training sets being different. It might be anticipated that, as the accuracy of geometric room acoustic models improve, this problem should disappear because the training set will better match reality.

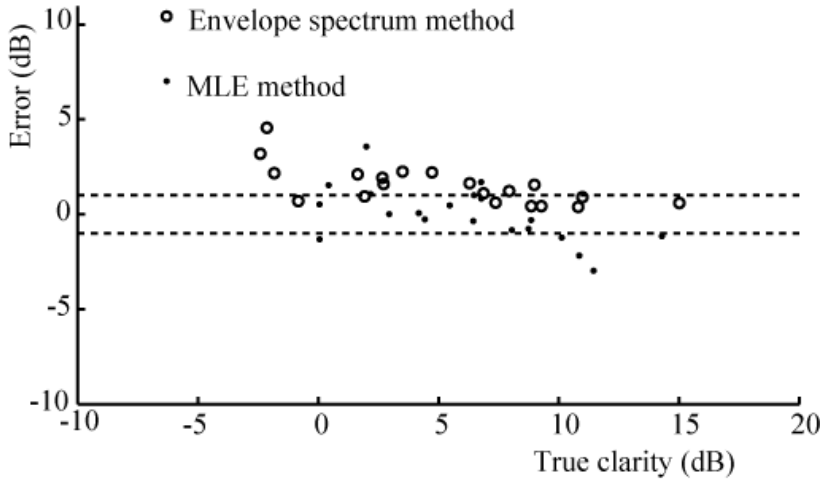


Figure 10-4. Error in parameter estimation versus the true value for clarity. •MLE and ***o*** *envelope spectrum method. Validation set using real RIRs. After Kendrick et al. [9].*

The MLE performs better with real room measurements when compared to the results for the reverberated speech using simulated impulse responses. The real impulse responses have a greater reflection density than the simulated ones, and consequently there is a smoother transition from the early to the reverberant sound field. It is suggested that this might improve the fitting of the simple model of envelope-shaped Gaussian noise, to the real room data, and this could therefore explain the improved accuracy in the parameter estimation.

10.2 Music

In this section the performance of the envelope spectrum and MLE methods are compared using music. Six different pieces of music [72] are used, 2-4 minutes in length. As noted above, the accuracy of the estimation changes from piece-to-piece, for instance the spectral variance affects the accuracy of the envelope spectrum method estimation, consequently, the results from the parameter estimations are averaged across all six pieces of music to improve accuracy and to reveal underlying trends.

10.3 Simulated RIRs

Considering first the validation set generated using simulated impulse responses from the geometric room acoustic model, for reverberation time, the MLE and envelope

spectrum methods produce similar accuracy, as shown in Figure 10-5, although the MLE appears to be marginally more accurate.

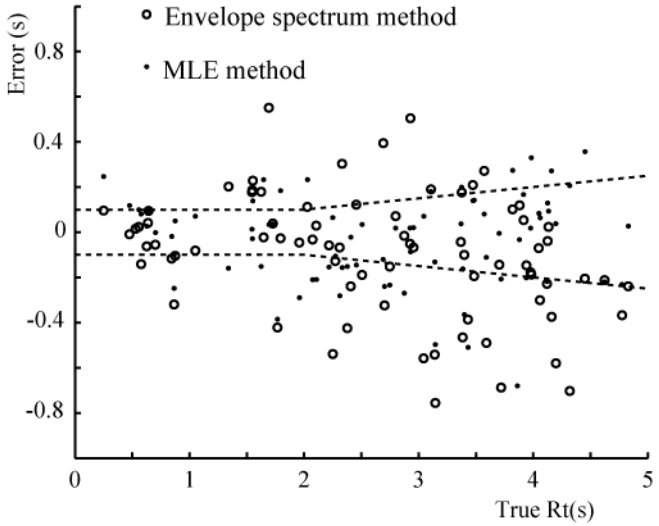


Figure 10-5. Error in reverberation time estimation versus the true value. The •MLE and **o** *envelope spectrum method. Validation set using simulated impulse responses. Music excitation. After Kendrick et al. [9].*

As shown in Figure 10-6, for other parameters (EDT, clarity and centre time) the envelope spectrum method is more accurate than the MLE method. Again, the MLE method finds it difficult to accurately estimate the early parts of the impulse response, due to the presence of the reverberance of the musical notes themselves. For the envelope spectrum method, most of the parameter estimations are within the difference limen. For the EDT and centre time, there is a tendency for the MLE method to overestimate the value by an amount larger than the difference limen. For clarity, there is a tendency for the MLE to overestimate low values of clarity, and underestimate large values, in other words, the range of clarity estimated by the MLE is smaller than the true range.

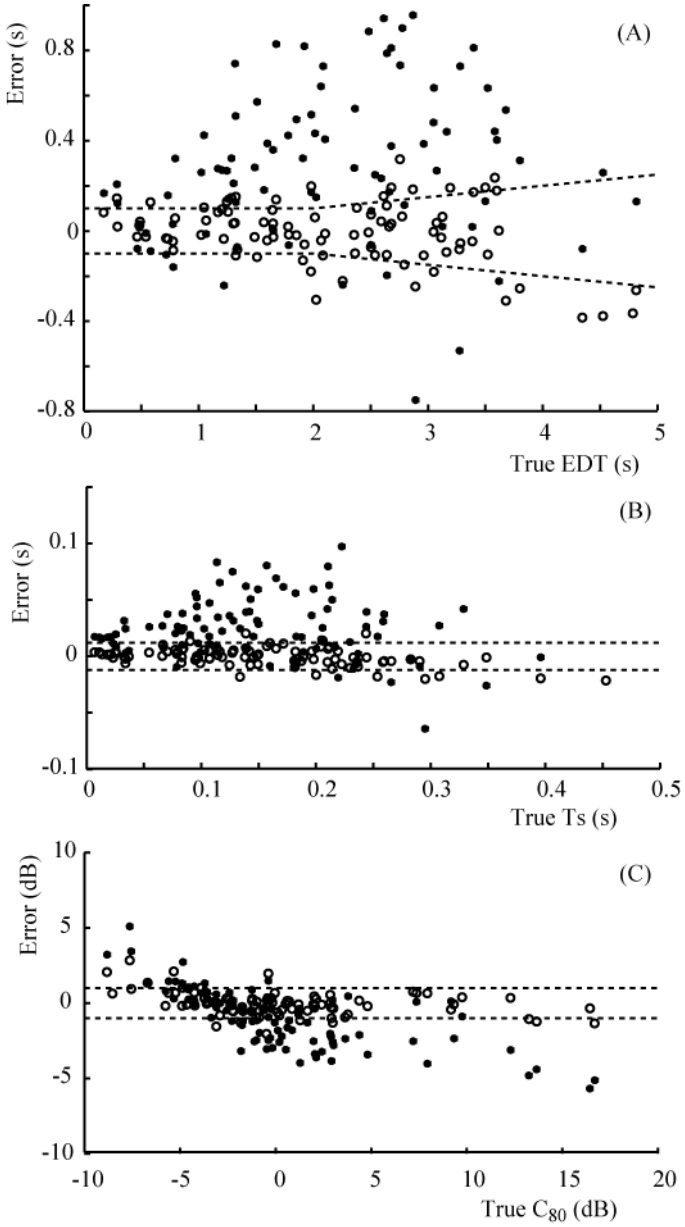


*Figure 10-6. Error in parameter estimation versus the true value: (A) EDT, (B) centre time and (C) clarity • MLE and **o** envelope spectrum method. Validation set using simulated impulse responses. Music excitation. After Kendrick et al. [9].*

10.4 Real, measured RIRs

With real room measurements, the results are again, somewhat different. Figure 10-7 and Figure 10-8 show the error in the four parameters as a function of their true value. In comparison to using the simulated validation set, the estimation of reverberation time and EDT has become more accurate with real room measurements for the MLE, while the envelope spectrum method gives similar accuracy. The estimation of clarity is still

problematic with the MLE method as there is a tendency for overestimation of low clarity values in the envelope spectrum method. For centre time, the MLE method becomes somewhat more accurate with real room measurements, but for the envelope spectrum method a bias error is introduced, with the parameter values being underestimated.

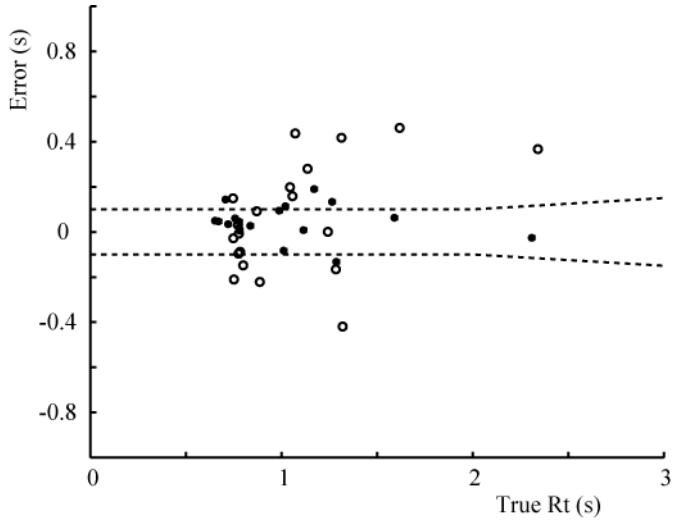


*Figure 10-7. Error in reverberation time estimation versus the true value •MLE and **o** envelope spectrum method. Validation set using real room measurements. Music excitation. After Kendrick et al. [9].*

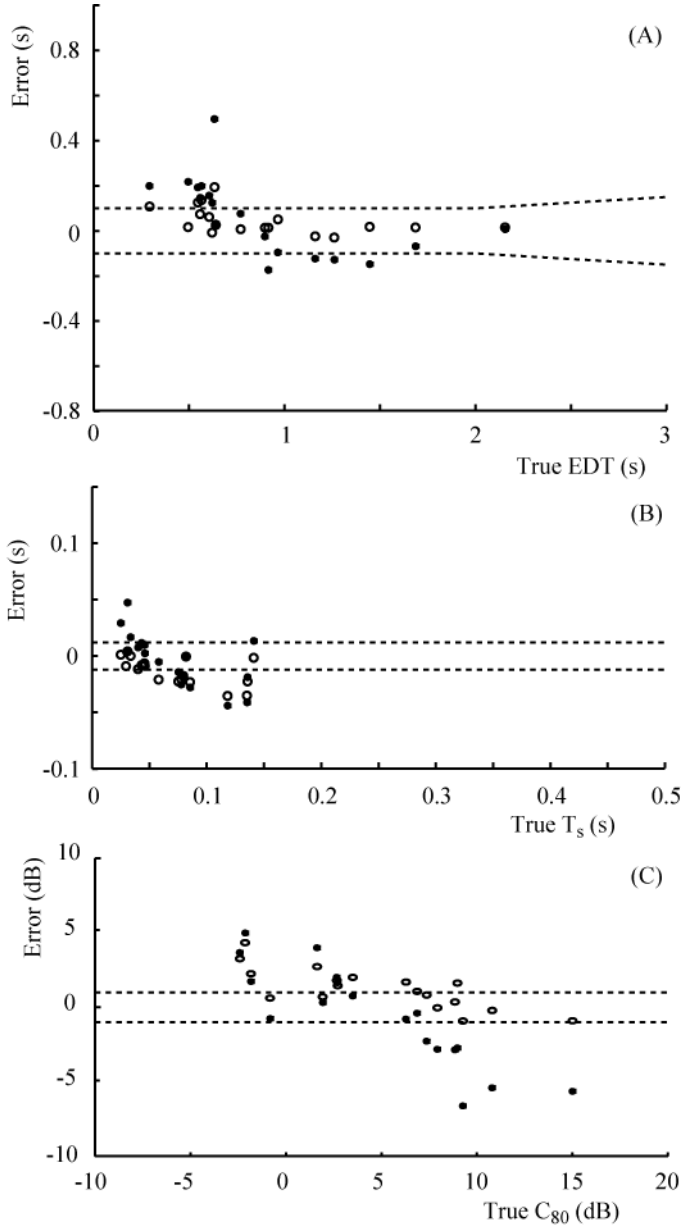


*Figure 10-8. Error in parameter estimation versus the true value: (A) EDT, (B) centre time and (C) clarity. • MLE and **o** envelope spectrum method. Validation set using real room measurements. Music excitation. After Kendrick et al. [9].*

As discussed previously, the introduction of a bias with the estimation results from the envelope spectrum method is probably indicative of differences between the training dataset which were simulated room impulses responses and the real room measurements – see discussions on speech above. The envelope spectrum method has problems estimating reverberation times. It is suggested that this arises because of the masking of the later parts of the impulse response by subsequent notes. The envelope spectrum method carries out an evaluation on the whole music passage, and in a piece of

reverberated music, the early decay of notes is going to be more prominent than later decay portions. Consequently, the envelope spectrum method struggles to accurately estimate reverberation times as the late decay information is lost in the vast amount of data from the whole music passage. To explore this further, companding was carried out on the signal (Chapter 6) before it was fed to the envelope spectrum pre-processor. Companding biases the signal towards late decay and, as expected, this improved the reverberation time estimation but at the expense of accurate EDT estimation. This was not such a problem with speech, because this has more periods of quiet. The MLE deliberately pulls out free decay sections and selects those with sufficient decay, therefore it does not suffer from the same problem. It is suggested, that by applying a similar signal segmentation process to the envelope spectrum method, the accuracy of reverberation time estimation might be improved.

For the MLE method, the improvement in estimation of reverberation time parameters when using real measurements is thought to be due to the real room impulse responses having greater reflection density than the simulated ones and so it is easier to fit the simplified room model to the data. For high clarity values, where underestimation occurs, the reverberance of the notes is causing an overestimation of both the early and late sound energies, hence resulting in a lower clarity than expected. For low clarity values, the MLE method struggles to fit the detailed effects of strong reflections in the early sound field, and so produces a consistent overestimation.

10.5 Discussion

Music and speech are obvious choices for measurement using naturalistic signals in performance spaces, and each has their own particular advantages and disadvantages. Narrative style speech has the advantage of containing plenty of gaps between the signal excitations, and so is richer in clean decay phases. This offers more opportunities for averaging estimates when using the MLE method and more accurate estimation using the envelope spectrum method, as there will be more decay phases where late reflections are not masked by subsequent utterances. However, speech excitation has limited bandwidth and lacks energy in low frequencies, say for the 250Hz octave band and below, and consequently methods that can also provide estimates using music signals are desirable.

Table 10-1 compares the standard error of each parameter estimate for speech and music using both the MLE and envelope spectrum methods. Table 10-1 indicates that the MLE method predicts the late decay properties more accurately than the envelope spectrum method, while the envelope spectrum method predicts the early decay behaviour more accurately than the MLE method.

		Standard error			
Method	Source Signal	Rt (s)	EDT (s)	C_{80} (dB)	Ts (s)
MLE (simulated, RT<5s)	Music	0.18	0.36	1.98	0.023
	Speech	0.17	0.27	1.34	0.017
Envelope spectrum (simulated, RT<5s)	Music	0.35	0.14	0.80	0.008
	Speech	0.19	0.06	0.50	0.004
MLE (Real RT<2.3s)	Music	0.08	0.14	2.52	0.015
	Speech	0.06	0.12	1.47	0.010
Envelope spectrum (Real RT<2.3s)	Music	0.18	0.10	1.51	0.012
	Speech	0.06	0.08	1.05	0.012

Table 10-1. Standard errors for parameters when estimated using the Maximum Likelihood and envelope spectrum method. Standard errors calculated using simulated impulse responses where Rt<5s.

The ideal music signal is one that has plenty of short transient sounds, and whose excitation is even across the octave band, in other words a piece of music which breaks traditional rules of western harmony and gives the same weight to all notes in the chromatic scale. Alternatively, averaging over different pieces of music can be effective, especially if the different pieces are in different musical keys. Another approach is to use pieces with lots of untuned percussion or music based on quarter-tone scales to provide a more broadband response. The most useful music signals also have large gaps between the notes so that the decay phases are prominent.

Neither the envelope spectrum method nor the MLE technique offers a single foolproof method for estimating room acoustic parameters from naturalistic signals. The MLE technique is appealing because it is not empirical and so is potentially more robust. Furthermore, it is truly blind and does not require explicit knowledge of the anechoic excitation signal. However, the MLE method has problems accurately estimating the early sound field because it cannot completely compensate for the inherent reverberance of music notes and speech utterances. As the calculation of reverberation time by

definition avoids this problematic region of decay, the MLE method is most successful in estimating reverberation times.

Currently, the inaccuracies of geometric room acoustic models used to provide training data for the envelope spectrum method limit the accuracy that can be obtained. As geometric room acoustic models improve, this problem will reduce. The method is not blind, but requires knowledge of the test signal during training. When using the envelope spectrum method with music signals, source independent measurement is not possible; this limits the applicability of the envelope spectrum method in that, for example, with a live orchestra the accuracy will be compromised.

11 OVERVIEW, CONCLUSIONS AND FURTHER WORK

This chapter summarises the research presented within this thesis, discusses the merits of the methodologies developed and identifies areas where further investigation could yield either academically interesting or practically useful results.

11.1 Overview

Two methods for extracting room acoustic parameters from music and speech signals have been researched, developed and their advantages and disadvantages explored in detail. The motivation is to enable the in-use measurement of acoustic parameters without using artificial test signals. The first method uses a machine learning approach utilising an envelope spectrum pre-processor. Previously, this method has been used with speech signals. This thesis details the adaptations necessary to enable music signals to be used. The second method applies a maximum likelihood estimation to decay phases at the end of speech utterances or musical notes where a multi-decay rate model of sound decay in a room is utilised within the framework.

11.1.1 Envelope spectrum method

The original envelope spectrum methodology that was developed for speech has been replicated using a new database of geometrically generated RIRs. Attempts to replicate the results with the database showed predictions of Rt with greater errors than demonstrated in previous work. Upon further investigation the decreased accuracy was determined to be due to the increased number of non-uniform decays; this makes the ANN's estimation task more difficult. When presented with many non-uniform decays, the ANN must perform a more complex mapping task as the feature space has become more complex. In addition, in order to accurately represent the global population of envelope spectra (i.e. all possible room responses) the number of RIRs in the training set needs to be very large. In this case the training set used was in excess of 8000 examples; however increasing this number would further improve the representation. By using a companding stage prior to the envelope spectrum detection, the Rt estimation accuracy was increased significantly as the late response is emphasise.

When extending the method to music signals, simulations showed that the accuracy of the ANN estimations was significantly reduced when compared with speech. Investigations found that one important reason for this was the difference between music spectra and speech spectra. Music signals have very uneven spectra in comparison with speech as most signal power in music sits in narrow frequency bands on the equal temperament scale. The simulated room responses showed different reverberant decays when excited only in these narrow $1/12^{th}$ octave bands, compared to the broadband response. Therefore, as the acoustic parameters are always determined from the broadband response, parameters calculated from a decay curve excited only in these narrow equal-tempered spaced bands will always have limited accuracy.

To combat this, a new pre-processor was developed which employed a 12^{th} octave spaced bank of band-pass filters. The envelope in each narrow-band is extracted, then normalised to its average energy before recombining the envelopes. Training the ANN on this weighted envelope spectrum resulted in significant improvements in parameter accuracy for all parameters, particularly for EDT, C_{80} and t_s. On average (over all parameters and for all pieces of music) the modified envelope pre-processor provided and increase in the percentage of estimates within the DL of approximately 23%.

A second phenomenon encountered is related to the inherent temporal patterns within music. These patterns, which are dependent on the tempo and the temporal structure of the musical score, cause the envelope to exhibit periodic behaviour manifested by peaks in the envelope spectrum. Modulation level estimates in-between the peaks in the envelope spectra are less reliable and so the ANN is effectively only trained using information at a small number of modulation frequencies. This limits the possible accuracy that can be achieved with music signals , so to gain a good estimation of features of the decay curve, information over a wider range of modulation frequencies is required. This is especially true for rooms with non-uniform decay curves.

The modified envelope spectrum method, when trained and tested using artificial room responses, yields good accuracy with one music signal for EDT, t_s and C_{80} with up to 85%, 94% and 98% within the difference limens respectively (1kHz octave band). Rt accuracy was less successful with up to 49% being within the difference limens. Different pieces of music yield worse results. The ideal music signal is one that has

plenty of short transient sounds (the best results are from a piece involving ***pizzicato*** strings), and whose excitation is even across the octave band; in other words a piece of music which breaks traditional rules of western harmony and gives the same weight to all notes in the chromatic scale. Alternatively, averaging over different pieces of music can be effective, especially if the different pieces are in different musical keys. Another approach is to use pieces with lots of untuned percussion. The most useful music also has large gaps between the notes so that the decay phases are prominent.

When evaluating the performance on real RIRs it was found that a bias error was introduced. A well trained ANN should not introduce bias errors and this was indicative of the difference between the artificial and the real RIRs. In particular, the simulated impulse responses demonstrated significant discontinuities, in the decay curve properties at the edges of octave bands. The presence or lack of excitation at the edges of octave bands can cause significant differences in the decay curve; this is due to, in combination, the sudden discontinuities and the finite roll-off of the octave band filters. This is a particular problem for music signals as the presence of (or lack of) certain notes at the edge of octave bands will have a major influence on the resulting reverberation decay curve. The bias errors were greatest for parameters dependent on the early sound-field (EDT, C_{80} and t_s) and it is known that the discontinuities have greatest effect on the early reflections indicating that the bias errors are most likely a direct result of these discontinuities.

It is expected that future developments in acoustic simulation techniques will yield further improvements in the accuracy of the envelope spectrum method. While accurate estimates may be reliably yielded for most speech signals, only specific music signals are appropriate for use with the envelope spectrum. It is this, together with the desire for a truly blind method requiring no extensive period of training which motivated the development of a second, more general method.

11.1.2 Maximum likelihood method

A maximum likelihood estimation of reverberation time was initially proposed by Ratnam *et al*. [3]. This method performed a model-based estimation of the decay in very short sections in between speech utterances. However, the exponential decay used

was an idealised model based on the assumption of a diffuse field. In reality non-exponential or weakly exponential decays are quite common and this causes large errors in the parameter estimation. Another feature of the Ratnam method was the use of an order statistics filter to select the Rt from a distribution of estimates; this distribution of estimates is signal dependant and therefore the method is only semi-blind.

The first stage of development of the ML method focused on the reverberant sound model. A multi-decay model of reverberation was developed which had two regions of decay and could effectively model non-diffuse and coupled room responses. The model was inserted into a maximum likelihood framework and an optimisation scheme was defined from which ML model parameters could be estimated. Regions of decay in the music signal were automatically selected using an envelope following algorithm, developed by Yonggang *et al*. [4]. These decay phases were used as inputs to the ML optimisation algorithm.

The accuracy of the method is dependent upon how the method processes the large number of ML decay curve estimates to produce the final result. The final estimate must be selected without using explicit knowledge of the signal. To achieve this, the number of decay phase estimates is first reduced by using only those with sufficient dynamic range. Both speech and music decay phases contain residual amounts of direct sound. This residual signal causes slower than expected rates of decay. To ensure the optimum result, the algorithm searches for the fastest decaying phases by minimising the direct sound in the decay phases. To achieve this, the signal was windowed (>1min windows) so that within each window, a decay curve was estimated by calculating the minimum energy over all decay curves for all points in time.

This minimum energy method is sensitive to stochastic variations in the decay and favours impulsive-like excitation. By favouring impulsive excitation, backwards integration must be carried out on the resulting estimates to calculate the decay curve. A benefit of bias towards impulsive excitation was the preferred spectrum; the algorithm was automatically biased toward selecting decay phases excited by more broadband-like signals. This is a useful result as the non-broadband nature of the music excitation was demonstrated as the major source of error in the envelope spectrum method.

Averaging the resulting decay curve estimates from each window yielded a more robust result. For speech 3 min windows were found to be sufficient, for music 10min windows were required. The accuracy increased with the number of windows used and it is recommended that a total length of between 10 and 20 minutes of speech is used; greater than 90mins is required for music. Different window sizes were optimum for different signals. The size of the window is related to the frequency of occurrence of suitable decay phases. This occurs relatively often in narrated speech but less frequently in music. Different pieces of music contain higher or lower densities of preferred decay phases and the method offers some robustness by only performing estimations using decay phases with sufficient dynamic range. Therefore regions with no suitable decay phases will be omitted from the estimation algorithm.

A method for blindly determining the accuracy of the ML estimation was also described using a bootstrap technique. This method calculated confidence limits on the decay curve which, in turn, indicates the confidence for each parameter. This technique can help to blindly determine the ideal window width for a particular signal and parameter.

The maximum likelihood method using speech yielded estimates of Rt from the octave bands in the range of 1000 – 8000Hz where between 71%- 93% of estimates were within the DL (standard error of about ±0.2s). There is decreased accuracy at 500Hz with 47% being within the DL (±0.4s). The estimated accuracy over the octave bands 500 to 8000Hz using speech for the parameters (C_{80} (37-66%), t_s (47-69%) and EDT (35-67%)) were lower than that for Rt, but useful estimates were still gained. Music achieves good accuracy for Rt estimation with between 46% and 58% of estimates within the DL (±0.3s) in the octave bands 1000 – 4000 Hz. Less accurate but still useful values were achieved for EDT (38-54%), C_{80} (48-59%) and t_s (37-62%) in these bands.

By performing the ML estimation on binaural signals, accurate estimates of the binaural parameters LG_{80} and ELEF (100% and 63% respectively within DL in the 1kHz octave band for speech) are achieved. Estimation of IACC using the ML method is not possible as all information regarding the fine detail of the reflections is lost.

To estimate the fine detail of the reflections, additional complexity in the model is required. The averaged cepstrum was used to blindly estimate the early echo locations but the reflection strength cannot be accurately estimated using this method. By using the echo location information from the cepstrum and combining it with the energy decay information from the ML method, a more accurate decay curve estimate can be determined. While this hybrid method offers no significant increase in parameter estimation accuracy, it enables the blind estimation of RIRs from speech and music for auralisation purposes. Informal subjective tests showed these impulses are more subjectively similar to the originals than those obtained using the simpler ML model.

11.1.3 Comparison of the two methods, implications and accuracy

The two methods have been compared and contrasted for common monaural parameters used in performance space design. The maximum likelihood method is most accurate for estimating reverberation time using both speech and music signals. For centre time, EDT, and C_{80} the envelope spectrum method is slightly more accurate when using simulated RIRs, however, the envelope spectrum method demonstrates a discrepancy between the results from the simulated and real RIRs. As a result, it is recommended that the ML method is most applicable for all parameters. When geometric room acoustic models become more accurate, better quality training data for the machine learning algorithm will be available, and this will ensure the envelope spectrum method will be more successful. ML estimated parameters can be biased, in some cases the early sound-field estimation is inaccurate because of; residual sounds in the decay phases, deviation from impulsive like excitation and the fact that the ML model does not model the fine structure of the early reflections. By utilising a bootstrap-based estimation of confidence limits on the decay curve, the bias can be minimised. The confidence limits can be used to indicate the optimum section length and the number of averages required to produce reliable decay curve estimates.

The maximum likelihood methodology shows particular promise as it is a truly blind method. Measurements carried out in real occupied halls using real orchestras yielded promising results. The accuracy falls a little short of the levels of accuracy yielded from standard measurement techniques, but yields accuracies greater than those achieved by

existing blind methods. When using speech, it is expected that by increasing the length of the signal and therefore the number of averages, the accuracy of the estimations may be increased sufficiently to demonstrate accuracy for most parameters to within the subjective difference limens. The accuracy is slightly compromised with music, but the error can be blindly quantified using the bootstrap method and therefore the user can determine the level of confidence to place in the result.

By facilitating accurate blind estimation of parameters using either speech or music, the ML method enables the estimation of acoustic parameters in occupied spaces that are usually difficult to measure. This research has investigated the accuracy that can be achieved with two methods and provides information for the user to quantify the level of confidence to place in the results. The accuracy of which, although lower than standard methods, is still competitive.

11.2 Further work

The envelope spectrum method requires a vast database of examples. The database used in this research, was comprehensive but may not have been sufficient to represent all possible responses and this caused training to stop early. The number of RIRs generated was limited by the computing power available but as affordable computing power has become more readily available in recent times, it would be straightforward to greatly increase the size of the databases and better represent the feature space. In addition, the acoustic simulation method demonstrated some shortcomings. Particularly, by performing calculations over octave bands the response was fairly uniform over each band which proved to be unrealistic. The discontinuities at the edge of the octave bands can cause problems for the ANN method which would benefit from a more realistic RIR generation method.

The major limitations of the envelope spectrum when applied to music are the uneven excitation exhibited by music and the inherent temporal patterns within music. To overcome these problems the ANN input could be extracted only from sections of the signal which demonstrate likely reverberant decay, rather than from the whole signal. The envelope segmentation used in the ML method could be employed as a pre-processing stage so that the ANN was presented only with reverberant decay. This

could maximise the reverberation to direct sound ratio and is expected to increase the accuracy of the method. This would have the added benefit of making the ANN method less signal dependent and may enable it to move past a one-net one-signal approach and become more versatile as a one-net many-signal approach. The literature also suggests an alternative method could be employed to search for probable reverberant decay. The short-term coherence function calculated from binaural signals as utilised by Vesa *et al*. [42], gives an indication to whether the signal is direct or reverberant. This could be used to detect reverberant sections and apply a dynamic weighting to the envelope spectrum calculation so that reverberant sections are more heavily weighted.

As previously mentioned, a drawback of the envelope spectrum method is the limitation of one network to one piece of music. Due to the large differences in envelope spectra between music signals (unlike speech which demonstrates similarities between talkers) training an ANN to perform source independent estimation by providing the network with many examples of different pieces of music is simply not possible using the existing framework. A fascinating paper by Beran [93] illustrates one aspect of the differences between pieces of music. He shows a positive correlation between composers' dates of birth and note entropy (the randomness of the distribution of notes in one octave). In order to achieve source independent measurement with the envelope spectrum method, a model of the music signal needs to be defined. A very simple model of the power spectrum density of a music envelope is '1/f' noise [91]. 1/f noise is a random signal where the spectral densities vary in proportion to 1/f. By modelling the received envelope as 1/f noise effected by a model of reverberation, as in Couvreur *et al.* [47], a statistical framework not dissimilar to the maximum likelihood method could be developed. Other more advanced stochastic models for music signals could also be utilised, such as Hidden Markov Models. Hopgood [92] demonstrated a model-based method for blind single channel deconvolution that could be used for blind estimation of acoustic parameters. The method incorporates a two part model of the reverberant signal; the source signal is modelled as a time varying autoregressive process and the room response as an IIR filter. By utilising the non-stationary nature of the signal i.e. decoupling the stationary room response from the non-stationary signal, the stationary parameters of the room model can be estimated, and from there the acoustic parameters estimated.

A limitation of the ML approach is that the model does not account for the convolution of the source signal with the room response; reverberant decay is always assumed. An alternative approach could utilise a signal model in addition to the RIR model. Signal models such as those suggested by Hopgood (autoregressive process) [92] and Couvreur (HMM) [47] could be formulated into a probabilistic framework. The optimisation procedure would then yield parameter estimates for both the room model and the signal model. This would make an interesting academic investigation but the focus on accurate acoustic parameter estimation must be maintained. Defining a more complex problem poses the risk that the problem becomes so difficult to solve that the acoustic parameter estimation accuracy may suffer. The success of such a method would depend on the appropriate choice of signal model; one which retains sufficient complexity to yield an improvement in accuracy but does not overcomplicate the problem.

The cepstrum method, when combined with the maximum likelihood estimation, provides a blind estimation of the RIR from speech and music signals. Only cursory comparisons with the true RIR were carried out. To determine the level of similarity with the original RIR, subjective listening tests need to be carried out. In addition, the combined ML/cepstrum method should provide the opportunity to estimate IACC, a parameter that the ML method was previously unable to estimate. Other extensions to the procedure could incorporate the reflection pattern detected by cepstral averaging into the ML estimation. The motivation for this is that by performing the ML estimation with the correct distribution (as opposed to the assumed Gaussian distribution) the energy level estimate may be more accurate and therefore yield more accurate parameter estimations. The cepstral estimation of the early echo location is not foolproof and a number of spurious echo locations do appear. As a result, more accurate methods to detect echo locations are desirable. Blind deconvolution and source separation techniques may provide useful ways to achieve better echo detection.

The noise immunity of the maximum likelihood method could be improved as currently any stationary background noise currently causes the late portion of the decay curve to be overestimated. Utilising methods designed to deal with excessive noise on measured impulse responses, such as proposed by Xiang [44], may help overcome some of these weaknesses.

Baskind [49] demonstrated the use of a pitch tracking algorithm, used to design a comb filter, to isolate single notes and their associated harmonic series. This same principal could be used in the lower frequency bands to isolate individual notes. This would provide much longer decay phases from which to perform the estimations as subsequent notes would be removed while preserving the decay excited by the previous note. The periodic structure of the filtered notes may cause problems in the ML estimation which assumes a noise like structure. This procedure would only be applicable at low frequencies as it only works with single, monophonic notes which are more likely to occur at low frequencies. The overlapping of harmonics from other notes at higher frequencies will most likely negate any advantages gained by this procedure.

In addition to these ideas a number of other interesting research questions were raised.

- Concert hall acoustics are generally measured using a broadband excitation. In concert halls, music signals are predominant and, as demonstrated in this thesis when the RIR is filtered using an averaged music spectrum, the acoustic parameters are significantly altered (especially EDT C_{80} and t_s). It would be interesting to discover what effect filtering of the RIRs has on the correlation with subjective opinion scores of the acoustics.

- Room acoustics are generally quantified by measurements from a single excitation source. The source and receiver are moved around, multiple measures made and the parameter averaged. However, in a concert the excitation is created by a number of different sources simultaneously. Depending on the listener's location, the superimposed reverberant decay response will differ from one measured at a single source position and the distribution of sources causes a distribution in path length and thus propagation time. Does this have a significant subjective effect on the listener and how can the effect be quantified? For example, would a larger orchestra with more varying path lengths cause the perceived clarity to reduce, and can this be taken into account when calculating acoustic parameters?

APPENDIX

A. Image source description of sound decay in a room

The decay of sound in a room can also be described by using image source modelling. Sound energy is lost by the following processes; absorption in a lossy medium (air), boundary absorption (wall materials) and by spherical spreading. Spherical spreading causes losses proportional to $^{1}/_{(ct)^{2}}$, ct being the distance travelled by a ray. The losses due to air absorption can be modelled by e^{-mct} , were m is the attenuation constant. Finally, energy loss due to boundary attenuation (assuming frequency independence) can be modelled by the multiplication by the energy reflection coefficient $(1-\alpha)$ where α is the absorption coefficient of the boundary. Boundary attenuation will happen many times, therefore sound rays will decrease in energy by $(1-\alpha)^{nt}$ where n is the average number of time a ray strikes a wall per second. All of this information combined to represent the energy from all reflections at a point of observation is;

$$E(t) = E_0 e^{(-mc + n\ln(1-\alpha))t} \qquad (A.1)$$

This also describes the decay of sound energy in a room as an exponential decay and as long as the absorption of the surfaces are reasonably similar, the decay is uniform (exponential). This will deviate from a purely exponential decay when the surfaces have very different absorptive properties.

B. Acoustic parameter distribution for database of simulated room responses used in ML estimation.

The following waterfall plots describe the distribution of acoustic parameters for the database of 100 simulated impulse responses used in the ML estimation validation.

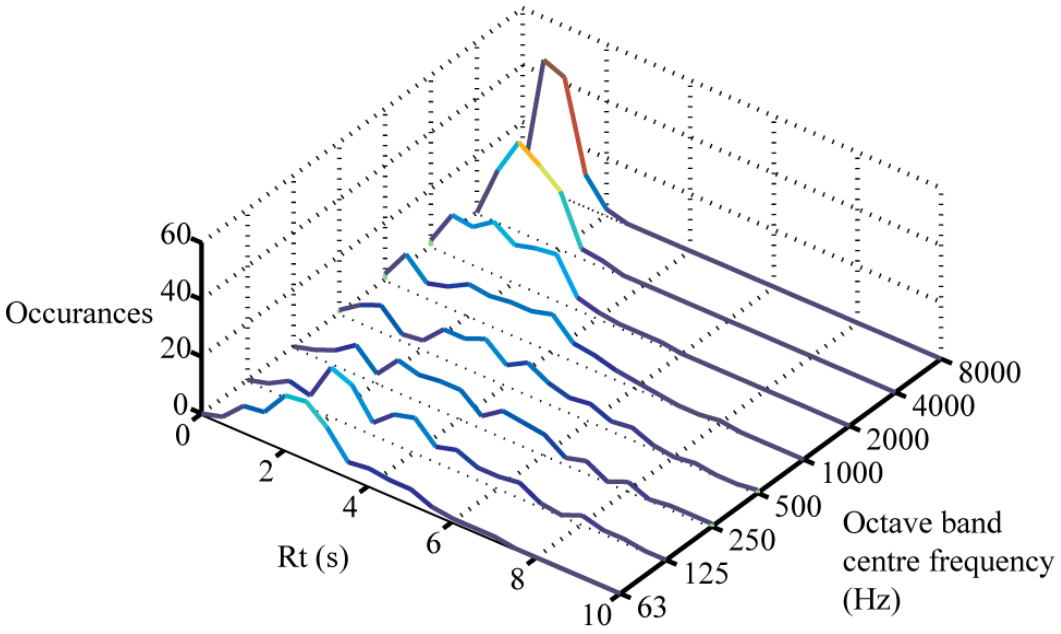


Figure A.1. Distribution of Rt values for database of 100 simulated room impulse responses for the octave bands from 63Hz to 8000Hz.

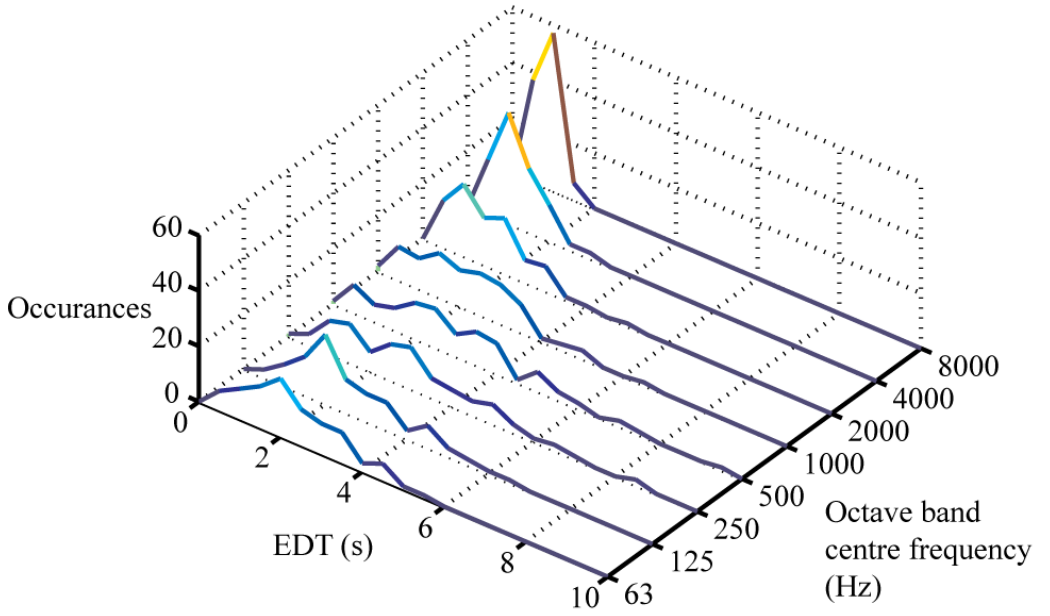


Figure B-2. Distribution of EDT values for database of 100 simulated room impulse responses for the octave bands from 63Hz to 8000Hz.

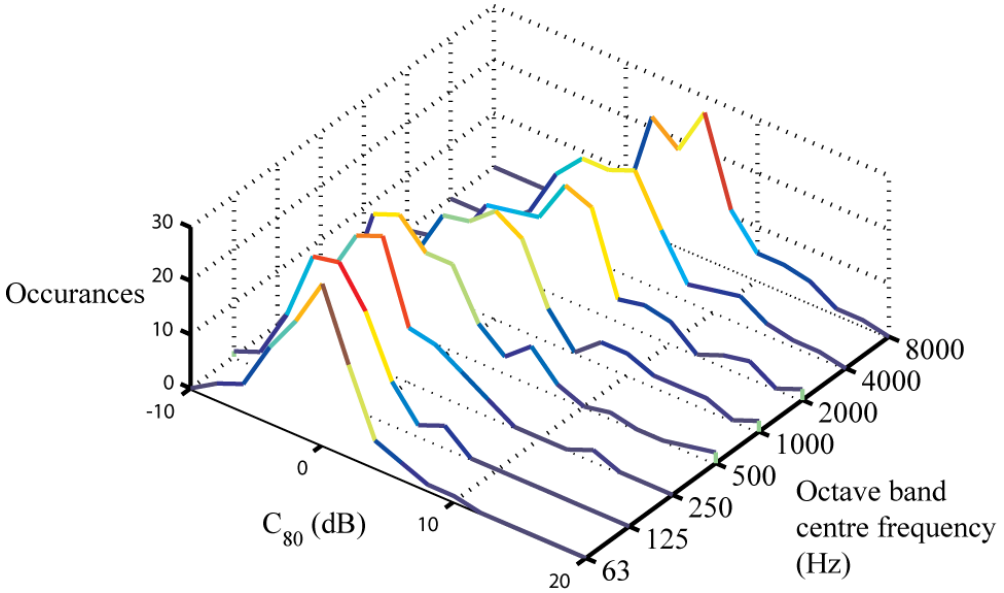


Figure B-3. Distribution of C_{80} values for database of 100 simulated room impulse responses for the octave bands from 63Hz to 8000Hz.

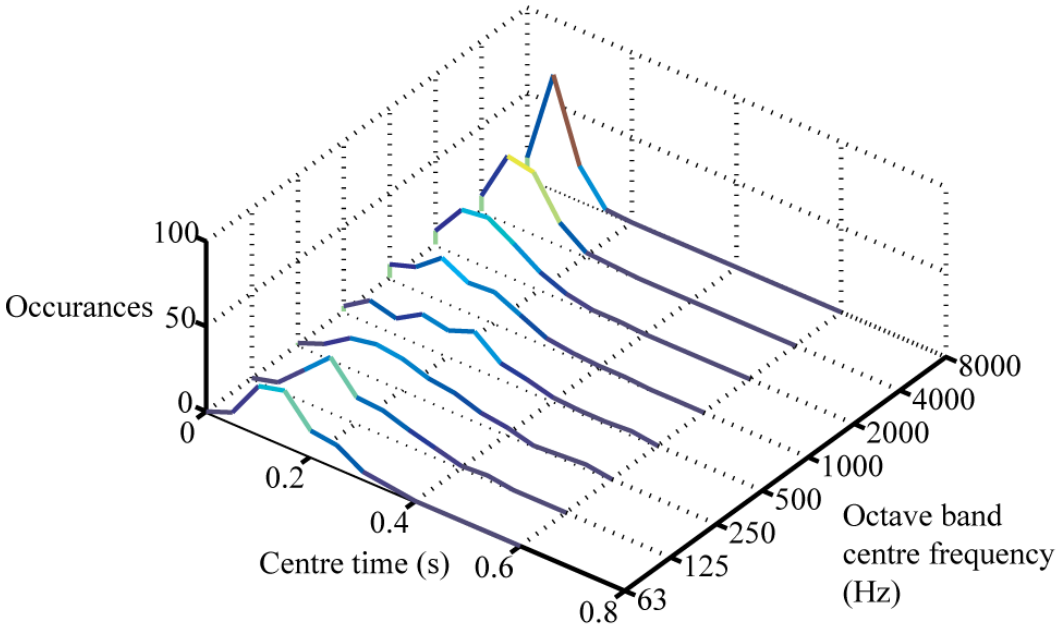


Figure B-4. Distribution of t_s values for database of 100 simulated room impulse responses for the octave bands from 63Hz to 8000Hz.

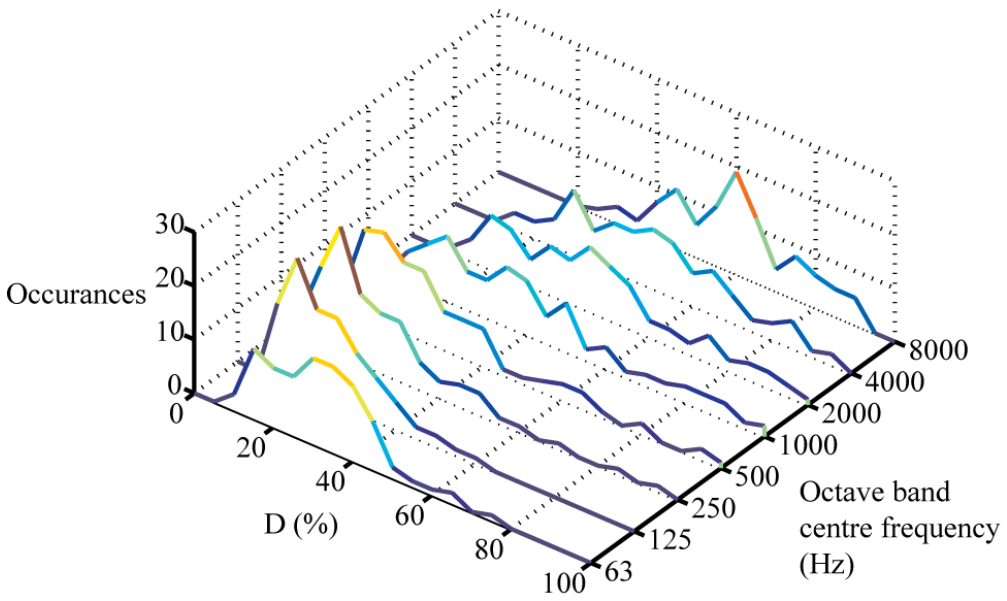


Figure B-5. Distribution of D values for database of 100 simulated room impulse responses for the octave bands from 63Hz to 8000Hz.

C. Acoustic parameter distribution for database of real room responses used in ML estimation

The following waterfall plots describe the distribution of acoustic parameters for the database of 18 real impulse responses used in the ML estimation validation.

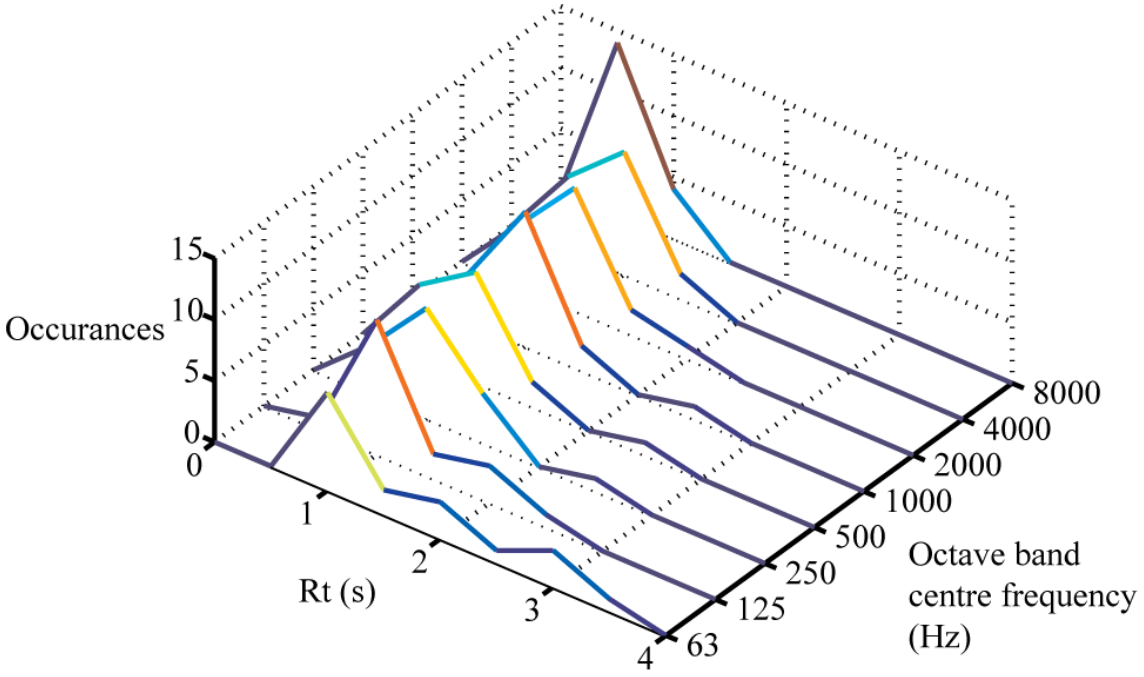


Figure C-6. Distribution of Rt values for database of 18 real room impulse responses for the octave bands from 63Hz to 8000Hz.

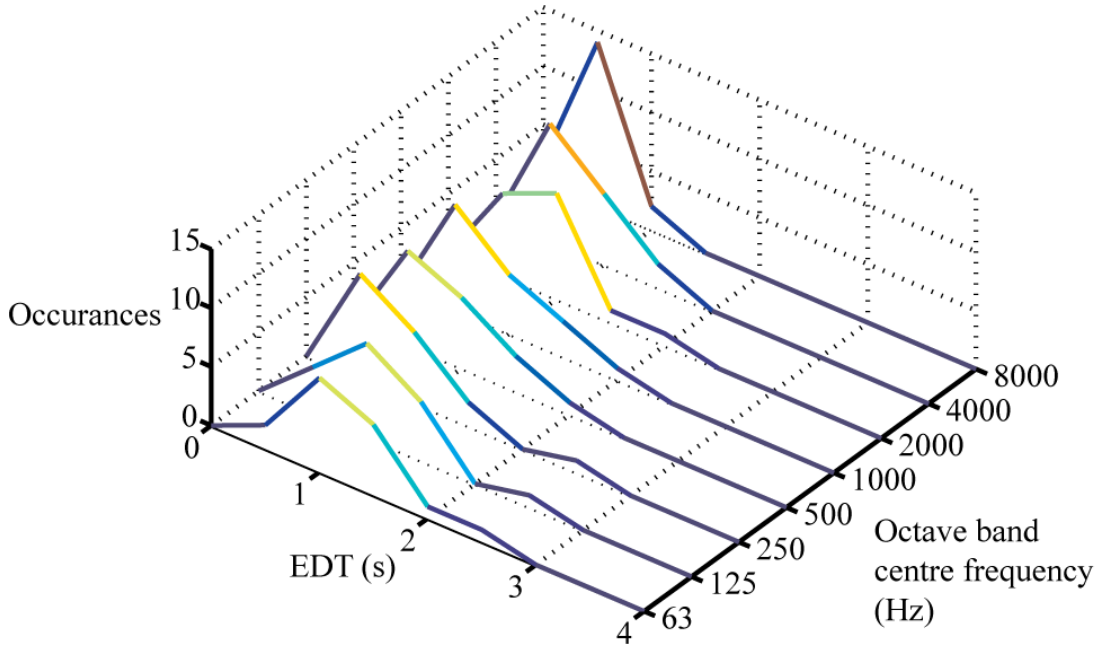


Figure C-7. Distribution of EDT values for database of 18 real room impulse responses for the octave bands from 63Hz to 8000Hz.

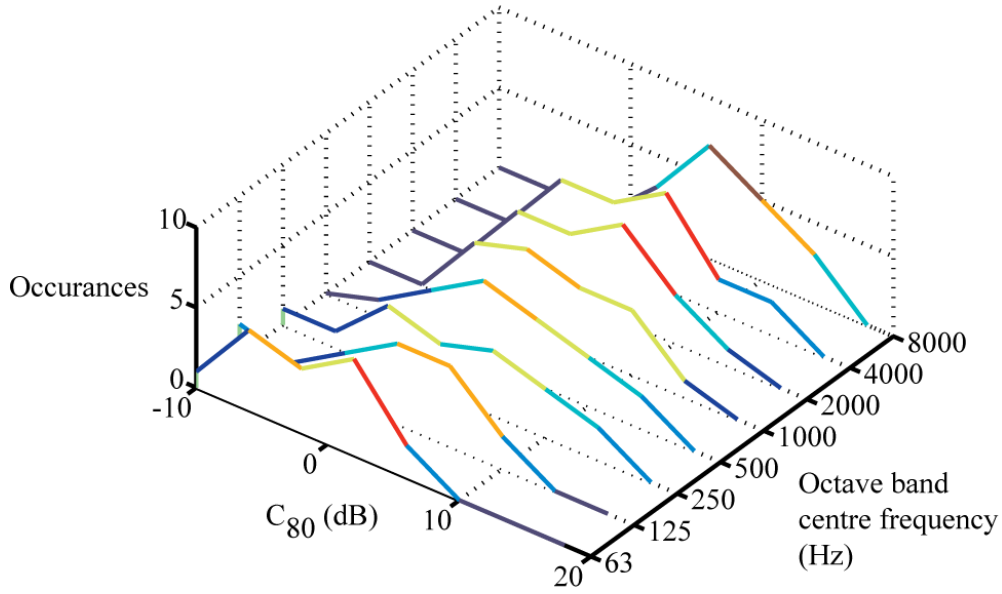


Figure C-8. Distribution of C_{80} values for database of 18 real room impulse responses for the octave bands from 63Hz to 8000Hz.

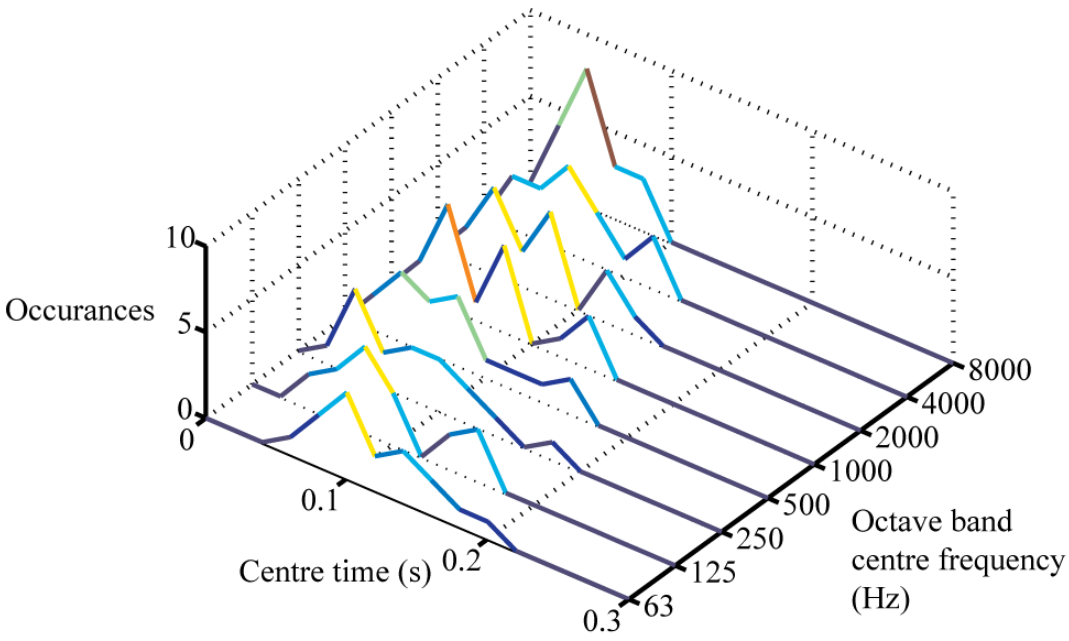


Figure C-9. Distribution of t_s values for database of 18 real room impulse responses for the octave bands from 63Hz to 8000Hz.

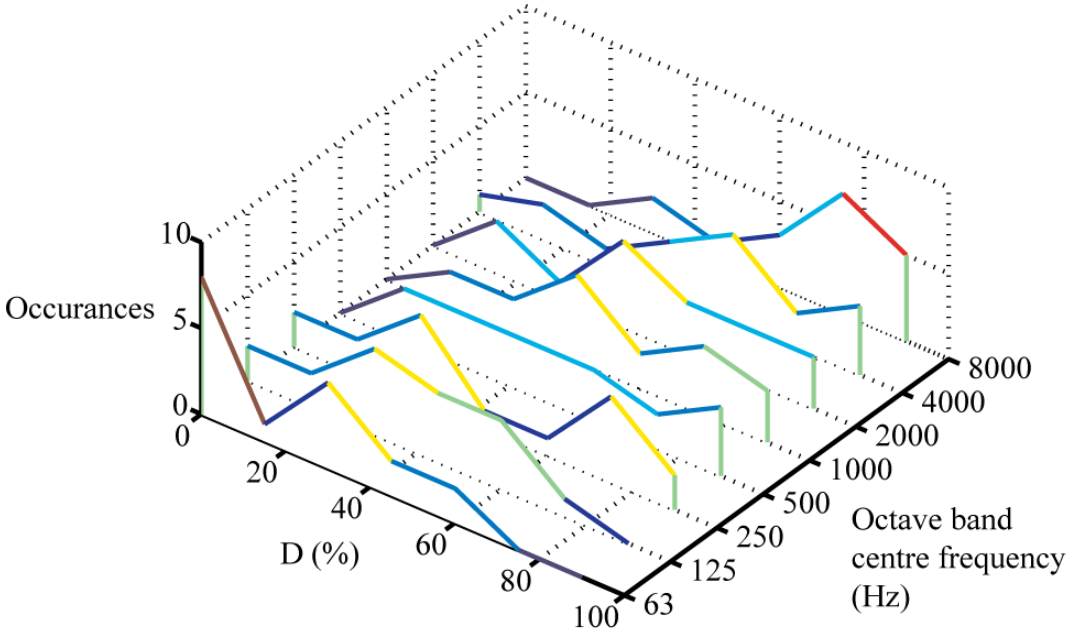


Figure C-10. Distribution of D values for database of 18 real room impulse responses for the octave bands from 63Hz to 8000Hz.

D. Anechoic music pieces

The following tables list the anechoic orchestral music recordings used in this thesis.

Table D-1 lists the pieces from [72]. These were created by recording a full orchestra in an anechoic chamber.

Title	Duration (mm:ss)
1. Mozart: Le Nozze De Figaro: Overture	4:19
2. Mendelssohn: 4th Mov. -SymphonyNo.3 In A Minor, Op.56 'Scottish', Bars 396-490	2:20
3. Bizet: L'Arlesienne, Suite No.2: Menuet, Bars 396-490	4:13
4. J. Strauss: Pizzicate-Polka	2:35
5. Pushkin: Ruslan And Lyudmila	5:22
6. Verdi: La Traviata	3:27
7. Bruckner: SymphonyNo.4 In EFlat 'Romantic',Bars 517-573	1:41
8. Debussy: Prelude AL'Apres-Midi D'un Faune, Bars 1-20	1:55

Table D-1. Denon anechoic orchestral music recording

Additionally the following recordings, available from [86], were used. These pieces were recorded one instrument at a time in an anechoic chamber and later reconstructed to realise the full orchestral recording.

Ttile	Duration (mm:ss)
W. A. Mozart (1756-1791): An aria of Donna Elvira from the opera Don Giovanni.	3:47
L. van Beethoven (1770-1827): Symphony no. 7, I movement, bars 1-53	3:11
A. Bruckner (1824-1896): Symphony no. 8, II movement, bars 1-61.	1:27
G. Mahler's (1860-1911): Symphony no. 1, IV movement, bars 1-85.	2:12

Table D-2. Anechoic recordings of symphonic music; Helsinki University of Technology for more information.

E. Alternative methodology – training the ANN on the MTFs directly

By training the ANN to predict the MTF from an envelope spectrum and then extracting the acoustic parameters from the estimated MTF, the ANN is being trained to perform approximate blind deconvolution. The ANN, rather than performing a non-linear mapping of envelope spectrum to a room parameter, is trying to perform the approximately linear mapping of envelope spectra to MTF. It was hoped that this will have benefits to parameter estimation accuracy. A possible problem with this method is that the MTF is in fact only an approximate description of the effect of reverberation and noise on speech and signals. By using an ANN it is hoped that the ANN can account for this and still yield accurate MTF estimates.

Another advantage to this method is that, by yielding an estimate of the MTF then, via an inverse Fourier transfer, the squared impulse response can be calculated and from this all the parameter estimates may be determined simultaneously. This eliminates the need to train a separate ANN for each parameter.

E.1 Methodology

The proposed scheme is described in Figure E-11. Firstly, the envelope spectra are generated (in this case high resolution linear frequency envelope spectra are used, sampled at 120 Hz using a window length of 4.3s). The MTF is calculated directly from the first 4.2s of the RIR (ensuring that the MTF and envelope spectra have the same spectral resolution). A neural network is defined with the same number of inputs as outputs, that number being the number of frequency bins (257). Initial experiments

used one hidden layer with linear activation functions. This means the ANN maps the process as linear function, MTF is an accurate representation of a transfer function that affects the anechoic envelope to produce the reverberant envelope this is sufficient, these results are presented in E-2 and E-3. Later experiments found this was insufficient and therefore increased the number of hidden layers and used non-linear activation functions these results are present in E-4. The ANN was trained using the magnitude of the MTFs as the training set.

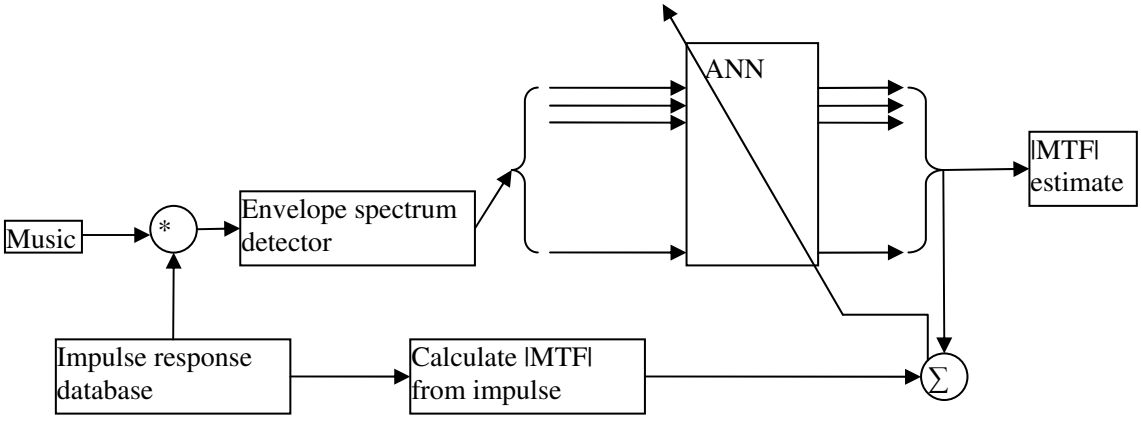


Figure E-11. Neural network method to estimate MTF directly from envelope spectra

E.2 Results – Estimating the MTF from the envelope spectrum

Figure E-12 shows a selection of a few of the resultant MTF estimates after training the ANN using reverberant music envelope spectra. Comparing some of the estimated MTFs with the actual MTFs shows reasonable agreement, although there are some differences. These differences are due to the inability of the ANN to find a function which maps, via a simple linear relationship, the MTF to the envelope spectrum. By using a single hidden layer with linear activation functions, the ANN does not have the capacity to account for the approximate nature of the MTF.

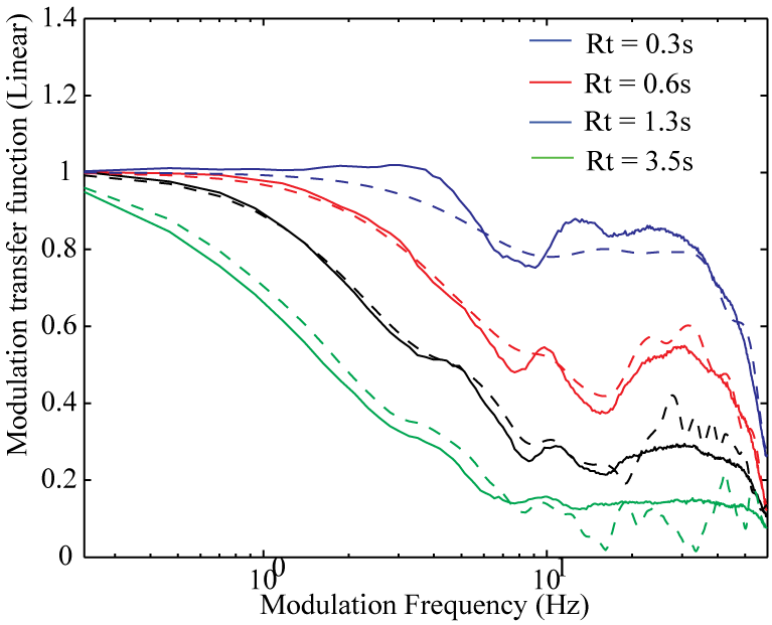


Figure E-12. ANN estimates of MTFs when presented with envelope spectra of reverberant music signal, dotted lines show the true MTF while the bold lines show the estimates

E.3 Results – extracting the decay curve from the MTF estimates

Now that an estimate of the MTF is available, the squared impulse response can be calculated by performing an inverse FFT on the data. However, prior to this procedure further treatment of the estimated MTFs is required as only positive frequencies and magnitude data are available which are insufficient for performing an inverse Fourier transform. The Fourier transform of a real signal always has a symmetrical magnitude spectrum and an anti-symmetrical phase spectrum. Therefore it is straight forward to calculate the negative frequencies from the positive ones. The phase information however has been lost and must be reconstructed using only the magnitude information. One method of reconstructing phase information from only magnitude data is known as minimum-phase reconstruction [51]. This methodology assumes that prior to the magnitude spectrum calculation, the signal was minimum phase. This implies that all poles and zeros of the z-transform are contained within the unit circle and therefore the phase response can be reconstructed using only the magnitude response. Performing this reconstruction requires the use of a signal processing methodology known as cepstral processing; this methodology is further discussed in Chapter 9. The real

cepstrum is the real part of the inverse Fourier transform of the log of the magnitude spectrum of a signal.

$$\hat{x}(t) = \Re\left[F^{-1}\left(\log\left(\left|X(f)\right|\right)\right)\right] \tag{E-1}$$

The cepstrum is a time domain signal and one of its properties is that if the signal *x(t)* has minimum phase, then the real cepstrum, $\hat{x}(t) = 0$ when *t<0.* Therefore to compute the phase response from the magnitude spectrum of a signal while enforcing the minimum phase condition the following procedure is followed. First the real cepstrum is computed, then all values in the real cepstrum for t<0 are set to zero, this is achieved by multiplying the cepstrum with a windowing function *w(t)* where *w(t)=1, t≤0* and *w(t)=0, t<0.* Performing an inverse cepstrum transform on this signal yields the reconstructed signal with an enforced minimum phase condition.

$$X_{\min\ phase}(f) = F\left[e^{\hat{x}(t)w(t)}\right] \tag{E-2}$$

An estimate for the squared impulse response can be gained via an inverse Fourier transform of $X_{\min\ phase}(f)$. Figure E-13 shows a number of decay curve estimates calculated from the ANN estimated MTFs (The same examples as depicted in Figure E-12 are used). What is apparent from Figure E-13 is that the method appears to correctly identify the early decay features, but struggles to indentify the correct late decay features. This is consistent with previous results where training on the EDT was more successful than training using the Rt. A very notable feature of these plots is the presence of a noise floor at relatively high SPL values. This noise floor causes problems when estimating the Rt and is due to the inability of the ANN to map the envelope spectrum to the MTF correctly.

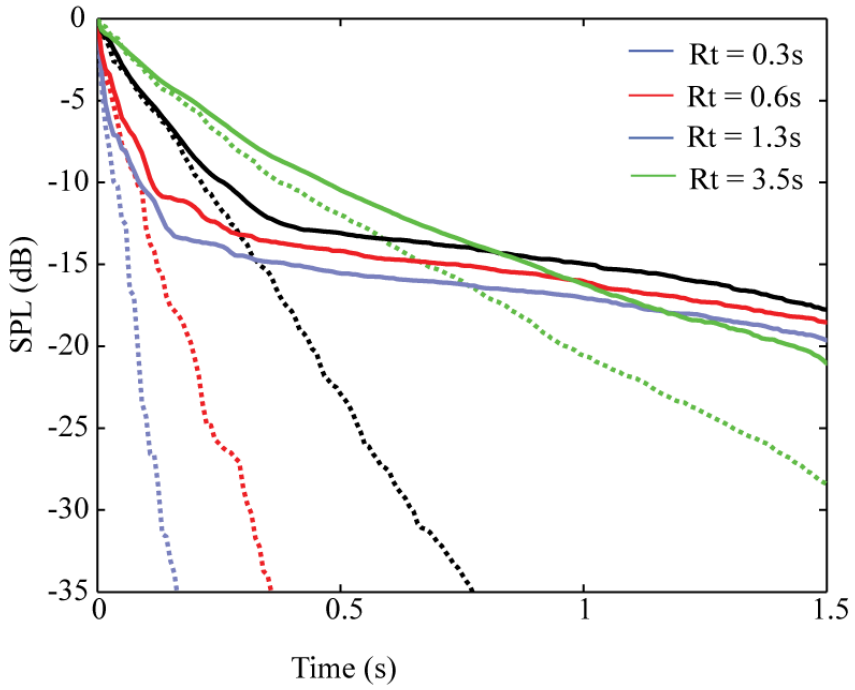


Figure E-13. Backwards integrated decay curve estimates calculated from MTFs estimated by the ANN method, dotted lines shows the estimates while bold lines show the true MTF.

The success of this method depends on the ability of the ANN to perform the deconvolution. However it was found to be less successful than the original envelope spectrum method. One possible reason for this is the large size of the ANN. The network was exceptionally large with 257 inputs and 257 outputs, while there was only one hidden layer with linear activation functions. This is a very large network and large networks are notoriously difficult to train. Another problem lies with the fact that the MTF can only approximately describe the envelope transfer characteristics of speech and music signals. Therefore significant errors occur when the simple network is unable to find the function which maps all the envelope spectra to their respective MTFs. Therefore to improve accuracy, methods are sought to reduce the data dimensionality and enable the ANN to successfully perform the mapping from envelope to MTF.

E.4 Reducing network dimensions by PCA pre-processing the data

In order to reduce the network size the dimensionality of the input and output data needs to be reduced. To achieve this, a dimensionality reduction technique known as principal component analysis (PCA) was used. This is a technique where directions of maximum variance are found within a data set, and these directions, known as components, and are then placed in order of decreasing variance. By discarding the latter components and retaining the first few it is hoped that the important information is retained while reducing the dimensionality. The PCA transformation matrix was calculated from the training MTF data set and the transformation matrix was applied to both the test and validation sets for both input and teacher vectors (i.e. both the envelope spectrum and the MTF). The PCA transformation is calculated using the Matlab command 'princomp'. Before calculating the principal components the data are centred and normalised to a variance of unity.

The dimensionality is reduced using principal component analysis and the 25 components with the largest variance are retained (in PCA analysis, components with larger variance is assumed to contain more information). PCA pre-processing is carried out on both the inputs and targets (the envelope spectra and MTFs). A network with two hidden layers was used (25-25-25-25).

Figure E-14 (1) shows some example MTF estimates, compared with the true MTFs. Speech convolved with the simulated database was used for these examples. Figure E-14 (2) shows the estimated decay curves calculated via an inverse Fourier transform on the estimated MTF. The phase response of the MTF is lost and so is reconstructed using a minimum-phase assumption (see E.3 for details). It is apparent from these examples that the method is achieving some success at estimating the correct decay curve. Problems arise in decay estimates in the later regions and manifest as an increased noise floor. This is for the same reasons that the Rt estimate is inferior to the EDT estimate in the previous methodology i.e. the late decay is masked by subsequent utterances and therefore the early reflections are dominant in the envelope spectrum.

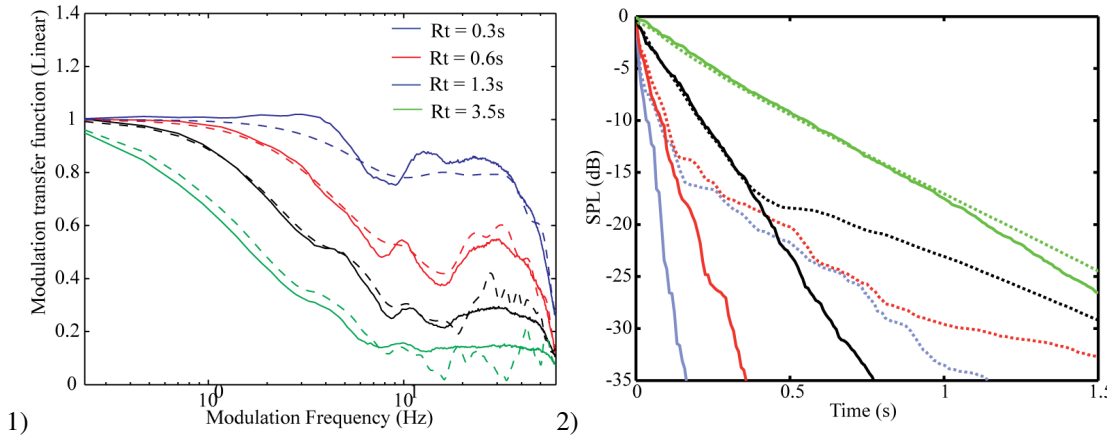


Figure E-14. Backwards integrated decay curve estimates(2) calculated from MTFs(1) estimated by the ANN method using PCA pre-processing, dotted lines shows the estimates while bold lines show the true MTF.

Figure E-15 presents a number of parameter estimates of Rt and EDT, calculated from the MTF estimates for a number of simulated RIRs. Figure E-15 (1) shows the EDT estimation accuracy is quite good especially up to about 2s. There appears to be a small bias towards overestimation of the EDT (<0.4s) at very low decay rates. The Rt estimation presented in Figure E-15 (2) shows generally poor accuracy due to the dominance of the early reflections in the signal. Similar to the overestimation at low EDTs there is also a trend for overestimation at low Rts (<1s). This occurs because the ANN cannot perform the correct mapping from envelope spectrum to MTF. It likely that the approximate relationship between the anechoic and reverberant envelope spectra means the problem is intractable or it may be that PCA pre-processing has removed critical information from the input or target data.

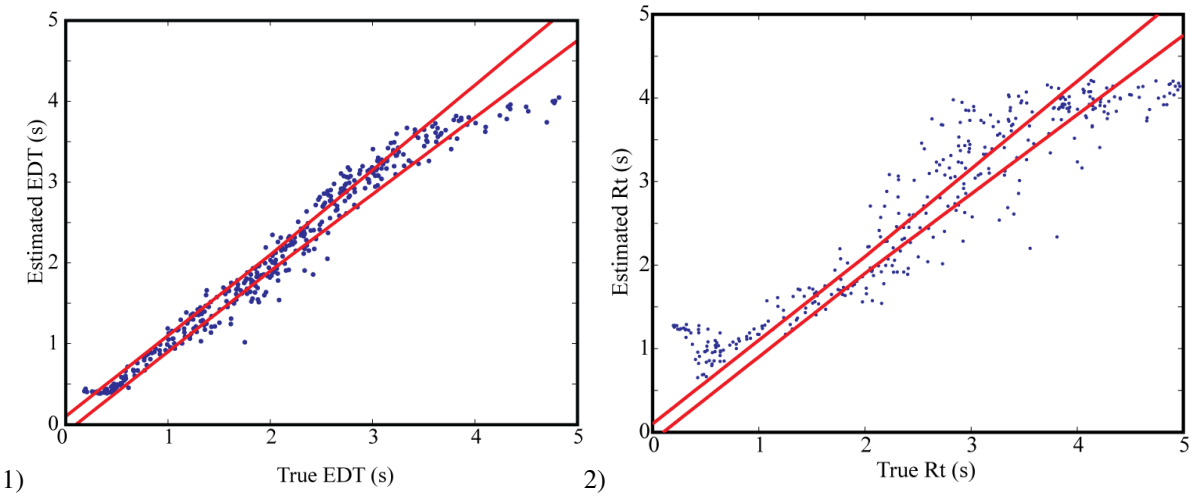


Figure E-15. Estimation of EDT(1) and Rt(2) from PCA pre-processed signals using an ANN with two hidden layers to estimate MTF

Compared with training directly for each room parameter, the results indicate that the error is increased when compared with the results in 6.3.2, the reason for this is thought to be due to the increase in problem complexity. The ANN is trained to produce estimates in a 25 dimensional feature space as opposed to the one dimensional space when trained directly to estimate acoustic parameters. Training an ANN using the parameters as targets yielded Rt and EDT estimates with 62% and 92% of the results with the DL respectively. Estimating the MTF with a single ANN and estimating all room parameters from the MTF predicted only 25% and 55% of results within the DL. Despite this decrease in accuracy, the ability to blindly estimate the decay curve is possibly a useful and valuable tool.

F. Additional results – MLE

F.1 Parameter estimates from speech convolved with real room impulse responses

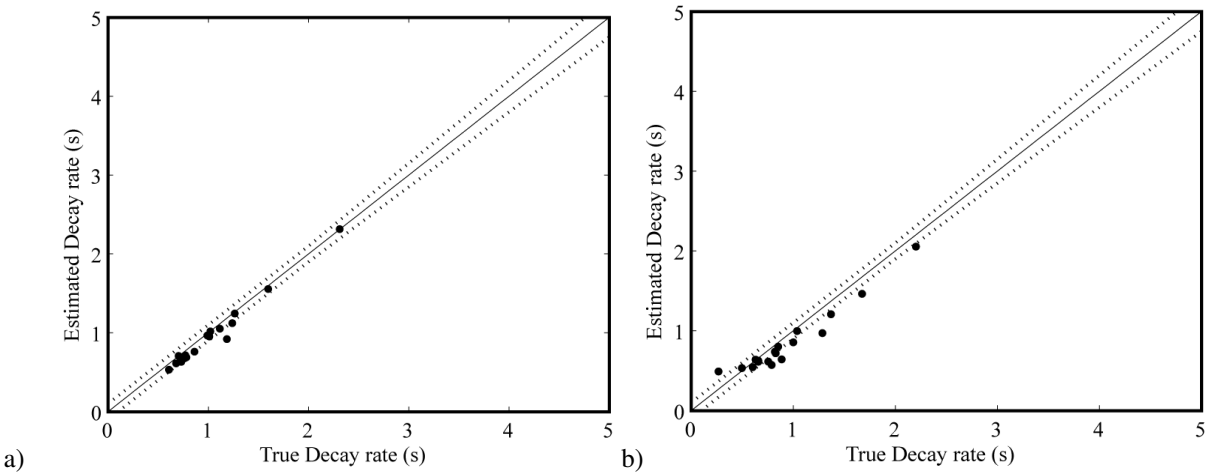


Figure F-16. Real impulse responses 9 mins of speech windowed into 1 ½ minute segments, results presented are for Rt(a) and EDT (b).

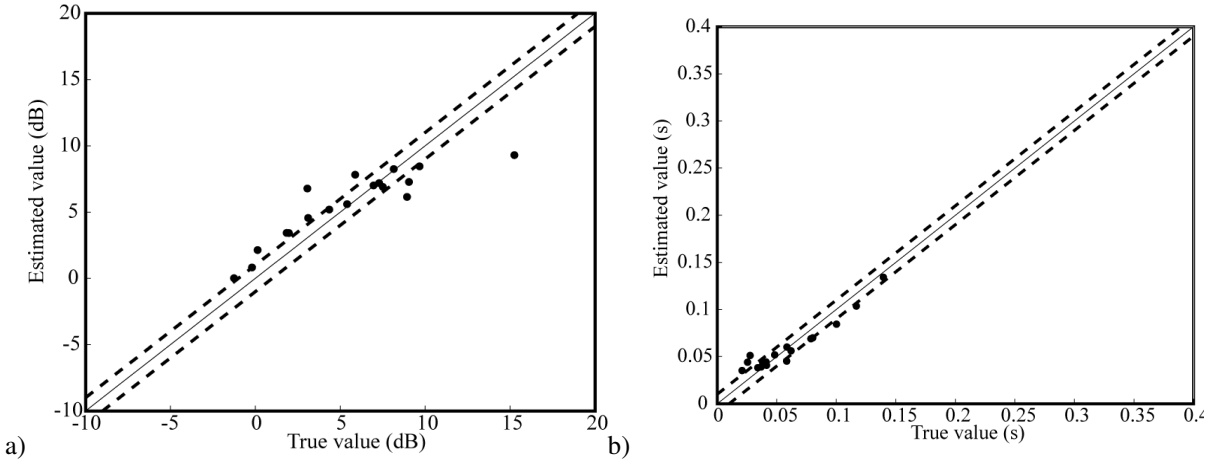


Figure F-17. Real impulse responses, 9 mins of speech windowed into 1 ½ minute segments, results presented are for C_{80} (a) and t_s(b).

F.2 Parameter estimates from speech convolved with real room impulse responses for all octave bands

	Octave band (Hz)							
	63	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(%)	11	39	22	50	83	89	89	100
EDT(%)	17	39	22	50	44	61	89	78
C80(%)	22	44	39	33	33	50	72	56
D(%)	28	33	33	17	28	33	33	44
Ts(%)	28	61	61	56	61	78	72	61

Table F-3. Percentage of estimates within difference limens from 9 mins of anechoic speech split into 7 segments. Results are shown for all octave bands.

	Octave band (Hz)							
	63	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(±s)	1.65	0.39	0.19	0.16	0.12	0.07	0.08	0.06
EDT(±s)	0.63	0.51	0.34	0.35	0.24	0.25	0.15	0.17
C_{80}(±dB)	4.20	3.28	4.43	4.51	4.24	3.54	3.62	4.13
D(±%)	22	19	20	21	17	14	16	13
Ts(±s)	0.04	0.02	0.03	0.03	0.02	0.02	0.02	0.02

Table F-4. Average parameter error from decay curve estimates yielded from 9 mins of anechoic speech split into 7 segments. Results are shown for all octave bands.

F.3 Parameter estimates from music convolved with real room impulse responses

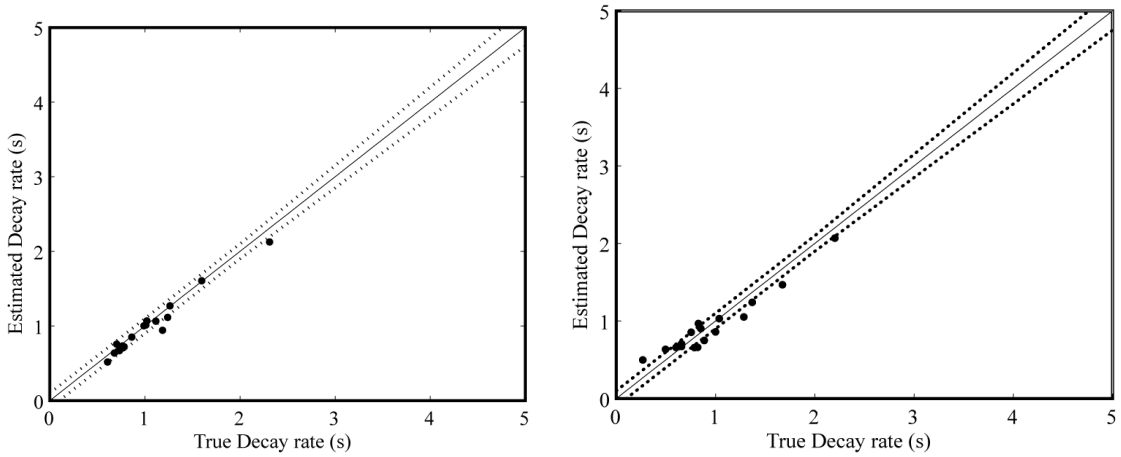


Figure F-18. Real impulse responses 40 mins of speech windowed into 4 minute segments, results presented are for Rt (a) and EDT (b).

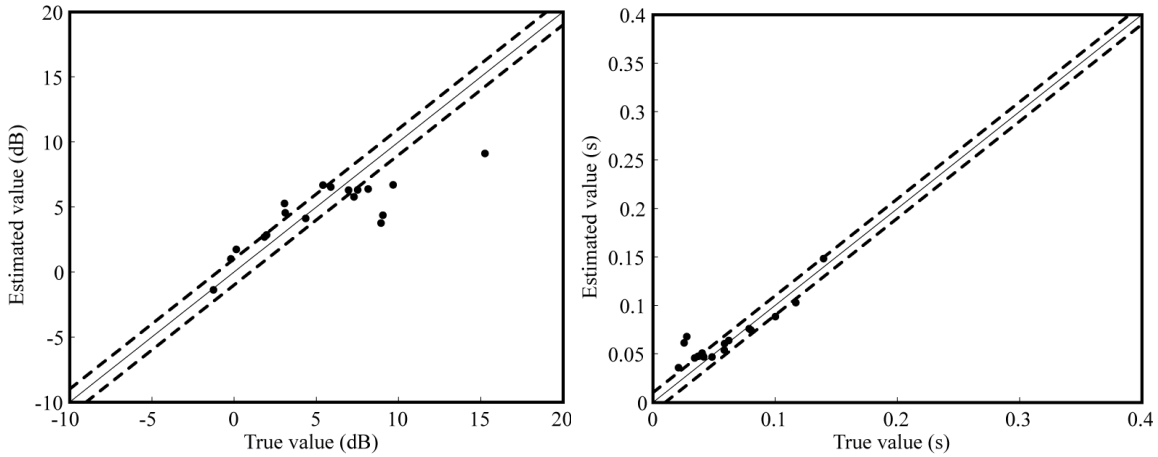


Figure F-19. Real impulse responses, 40 mins of speech windowed into 4 minute segments, results presented are for C_{80} (a) and t_s (b).

F.4 Parameter estimates from music convolved with real room impulse responses for all octave bands

	Octave band (Hz)							
	63	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(%)	6	28	44	44	83	78	89	33
EDT(%)	22	33	33	50	33	56	56	39
C_{80}(%)	22	39	28	33	44	33	39	6
D(%)	17	17	6	17	17	28	17	0
Ts(%)	33	44	28	50	56	50	56	17

Table F-5. Percentage of ML parameter estimate from 40 mins of anechoic music segmented into 8, 5 min segments, the music has been convolved with real room impulse responses.

	Octave band (Hz)							
	63	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Rt(±s)	1.8	0.7	0.4	0.2	0.1	0.2	0.1	0.3
EDT(±s)	0.7	0.7	0.5	0.3	0.3	0.3	0.2	0.2
C_{80}(±dB)	4.8	4.2	6.0	5.8	5.2	4.8	4.2	4.5
D(±%)	25	22	27	26	22	20	18	15
Ts(±s)	0.05	0.03	0.04	0.03	0.03	0.03	0.02	0.03

Table F-6. 95% confidence bounds on ML parameter estimate from 40 mins of anechoic music segmented into 8, 5 min segments, the music has been convolved with real room impulse responses.

12 REFERENCES

[1] T. Hidaka, N. Nishihara, and L. L. Beranek, "Relation of acoustical parameters with and without audiences in concert halls and a simple method for simulating the occupied state," *J. Acoust. Soc. Am.,* vol. 109, pp. 1028-42, 2001.

[2] F. F. Li, "Extracting room acoustic parameters from received speech signals using artificial neural networks," PhD thesis, *Acoustics Research Centre*: University of Salford, 2002.

[3] R. Ratnam, D. L. Jones, B. C. Wheeler, W. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.,* vol. 114, pp. 2877-92, 2003.

[4] Y. Zhang and J. A. Chambers, "Blind Estimation of Reverberation time," PhD thesis, University of Cardiff, 2006.

[5] T. J. Cox, F. Li, and P. Darlington, "Speech transmission index from running speech: A neural network approach," *J. Acoust. Soc. Am.,* vol. 113, pp. 1999-2008, 1999 (2003).

[6] F. F. Li and T. J. Cox, "A neural network for blind identification of speech transmission index," in *IEEE ICASSP2003*, 2003.

[7] P. Kendrick, T. J. Cox, F. L. Francis, Y. Zhang, and J. A. Chambers, "Blind estimation of clarity, centre time and deutlichkeit from speech and music signals," in *19th International Congress on Acoustics* Madrid, 2007.

[8] ISO 3382:1997 "Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters"

[9] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, and J. A. Chambers, "Blind Estimation of Reverberation Parameters for Non-Diffuse Rooms," *Acta Acustica United with Acustica,* vol. 93, pp. 760 -770, 2007.

[10] M. R. Schroeder, "New Method of Measuring Reverberation Time," *J. Acoust. Soc. Am.,* vol. 37, pp. 409, 1965.

[11] W. C. Sabine, *Collected Papers on Acoustics*: Harvard University Press, 1922.

[12] V. L. Jordan, "Acoustical criteria for Auditoriums and Their Relation to Model Techniques," *J. Acoust. Soc. Amer,* vol. 47, pp. 408-412, February 1970.

[13] A. C. Gade, in *Proceedings of the Sabine Centennial Symposium* Cambridge, Mass., 1994.

[14] W. Reichardt, O. A. Abdel, and W. Schmidt, 'Defintion und Meβgrundlage eines objektiven Maβes zur Ermittlung der Grenze zwischen brauchbarer und unbrauchbarer Durchsichtigkeit bei Musikdarbietung'.*Acustica,* vol. 32, pp. 126-137, 1974.

[15] R. Kürer, "Zur gewinnung von eizahlkriterien bei impulsmessungeg in der raumakustik," *Acustica,* vol. 21, 1969.

[16] H. Kuttruff, *Room acoustics*: Spon press, 2000.

[17] H. J. M. Steeneken and T. Houtgast, "A Physical Method for Measuring Speech-Transmission Quality," *J. Acoust. Soc. Am.,* vol. 67, pp. 318-326, Jan 1980.

[18] T. Houtgast and H. J. M. Steeneken, "Envelope spectrum and intelligibility of speech in enclosures," in *IEEE-AFCRL Conference on Speech Communications and Processing*, pp. 392-395, 1972.

[19] T. Houtgast and H. J. M. Steeneken, "Envelope spectrum and intelligibility of speech in enclosures," *IEEE-AFCRL Conference on Speech Communications and Processing,* pp. 392-395, 1972.

[20] T. Houtgast and H. J. M. Steeneken, "A Multi-Lingual evaluation of the RASTI-method for estimating speech intelligibility in auditoria," *Acustica,* vol. 54, pp. 185-199, 1984.

[21] IEC, "Sound system equipment " in *Part 16: Objective rating of speech intelligibility by speech transmission index*. vol. 60268-16, 2003.

[22] M. R. Schroeder, "Modulation transfer functions: Definitions and measurement," *Acustica,* vol. 49, pp. 179-182, 1981.

[23] J. D. Ploack, H. Alrutz, and M. R. Schroeder, "The modulation transfer function of music signals and its application to reverberation time measurement," *Acustica,* vol. 54, pp. 257-265, 1984.

[24] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *Journal of Sound and Vibration,* vol. 77, p. 211, 1981.

[25] J. S. Bradley, G. A. Souladore "The influence of late arriving energy on spatial impression," *Journal of the Acoustic Society in America,* vol. 97, pp. 2263-2271 ,1995.

[26] G. B. Stan, J. J. Embrechts, and D. Archambeau, "Comparison of Different Impulse Response Measurement Techniques," *Journal of the Audio Engineering Society,* vol. 50, pp. 249-262, 2002.

[27] W. D. T. Davis, "Generation and Properties of Maximum length Sequences," *Contro, vol. 10, pp* 302*,* 1966.

[28] M. R. Schroeder, "Integrated Impulse Method Measuring Sound Decay Without Using Impulses," *J. Acoust. Soc. Am.,* vol. 66, pp.497-500, 1979.

[29] H. Alrutz and M. R. Schroeder, in *Proceedings of the 11th International Congress on Acoustics*, Paris, 1983.

[30] C. Dunn and M. O. Hawksford, "Distortion Immunity of MLS-Derived Impulse Response Measurements," *Journal of the Audio Engineering Society,* vol. 41, 1993.

[31] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Audio Engineering Society Pre-print* Paris, 2000.

[32] I. Bork, "A comparison of room simulation software — the 2nd round robin on room acoustical computer software," *Acta Acustica,* vol. 86, pp. 943–956, 2000.

[33] T. J. Cox, W. J. Davies, and Y. W. Lam, "The Sensitivity of Early Sound Field Changes in Auditoria," *Acustica,* vol. 79, pp. 27–41, 1993.

[34] I. B. Witew, "Is the perception of listener envelopment in concert halls affected by clarity?," in *DAGA*, 2006.

[35] J. L. Nielsen, "Maximum-Length Sequence Measurement of Room Impulse Responses with High Level Disturbances," in *100th Audio Engineering Society Conventions* Copenhagen, 1996.

[36] J. Meyer, "Equalization Using Voice and Music as the Source," in *76th Audio Engineering Society Convention*, 1984.

[37] W. T. Chu, "Impulse-response and reverberation-decay measurements made by using a periodic pseudorandom sequence," *Applied Acoustics,* vol. 92, pp. 193-205, 1990.

[38] L. Cremer and H. Müller, *Principles and Applications of Room Acoustics: Geometrical, Statistical and Psychological Room Acoustics Vol 1*: Spon Press, 1982.

[39] L. Beranek and D. W. Martin, "Concert and Opera Halls : how they sound," *J Acoust Soc Am.,* vol. 99, pp.779-780, 1996.

[40] T. J. Cox, F. Li, and P. Darlington, "Extraction of room reverberation time from speech using artificial neural networks," *J. Audio. Eng. Soc.,* vol. 49, pp. 219-230, 2001.

[41] A. Baskind and O. Warusfel, "Methods for blind computational estimation of perceptual attributes of room acoustics," in *AES 22nd international conference on virtual, synthetic and entertainment Audio*, Espoo, Finland, 2002.

[42] S. Vesa and A. Härmä, "Automatic estimation of reverberation time from binaural signals," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* 2005.

[43] J. Vieira, "Estimation of Reverberation Time without Test Signals," in *118th Audio Engineering Society Convention*, Barcelona, 2005.

[44] N. Xiang, "Evaluation of reverberation times using a nonlinear regression approach," *J Acoust Soc Am.,* vol. 94, pp. 2112-2121, 1995.

[45] M. Hansen, "A method for calculating reverberation time from music signals," in *The Acoustics Laboratory* Technical University of Denmark, 1995.

[46] R. Ratnam, D. L. Jones, and J. William D. O'Brien, "Fast Algorithms for Blind Estimation of Reverberation Time," *IEEE Signal Processing Letters,* vol. 11, pp. 537-540, June 2004.

[47] L. Couvreur, C. Ris, and C. Couvreur, " Model-based Blind Estimation of Reverberation Time: Application to Robust ASR in Reverberant Environments " in *7th European Conference on Speech Communication and Technology* Denmark, 2001.

[48] M. Wu and D. Wang, "A Pitch-Based Method for the Estimation of Short Reverberation Time," *Acta Acustica United with Acustica,* vol. 92, pp. 337-339, 2006.

[49] A. Baskind and A. D. Cheveigne, "Pitch-Tracking of Reverberant Sounds, Application to Spatial Description of Sound Scenes," in *24th AES International Conference Multichannel Audio: The New Reality* Canada, 2003.

[50] Y. Zhang, J. A. Chambers, P. Kendrick, T. J. Cox, and F. F. Li, "A combined blind source separation and adaptive noise cancellation scheme with potential application in blind acoustic parameter extraction," *Neurocomputing,* vol. 71, pp. 2127-2139, 2008.

[51] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*: Prentice Hall, 1975.

[52] "CATT-Acoustic v8.0c," p. Room acoustic modelling software.

[53] T. Lahti, A. Ruusuvuori, and H. Moller, "The acoustic conditions in Finnish concert spaces - Preliminary results," in *Audio Eng.Soc. 110th convention*, 2001.

[54] V. L. Jordan, *Acoustical design of Concert halls and Theatres*, 1980.

[55] "DIN 18041:2004 Acoustical quality in small to medium-sized rooms," 2004.

[56] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications* Springer, 2007.

[57] R. Hooke and T. A. Jeeves, ""Direct search" solution of numerical and statistical problems.," *J. Ass. Comput. Mach,* vol. 8, pp. 212-29, 1961.

[58] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, pp. 671-680, 1983.

[59] K. V. Price, R. M. Storn, and J. A. Lampinen, *Differential evolution: a practical approach to global optimization*: Birkhäuser, 2005.

[60] S. Haykin, *Neural Networks A Comprehensive Foundation*: Pearson Prentice Hall, 1999.

[61] W. S. McCulloch and W. H. Pitts, "A logical calculus of the ideas immanent in nervous activity. ," *Bulletin of Mathematical Biophysics,,* vol. 5, pp. 115-133, 1943.

[62] F. Rosenblatt, "The Perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological Review,* vol. 65, pp. 386-408, 1958.

[63] J. Heaton, *Introduction to Neural Networks for Java*: Heaton Research, Inc., 2008.

[64] B. Widrow and M. E. Hoff, "Adaptive Switching Circuits," *IRE WESCON Convention Record,* 1960.

[65] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature,* vol. 323, pp. 533-536, 1986.

[66] Mathworks, "Matlab," 7.1 ed, 2005.

[67] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*: Wiley, 1996.

[68] P. C. Mahalanobis, "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India,* vol. 2, pp. 49-55, 1936.

[69] B&O, "Music For Archimedes," C. 101, Ed., 1992.

[70] M. T. Hagan and M. Menhaj, "Training Feedforward Networks with the Marquardt Algorithm," *IEEE Transaction od Neural Networks,* vol. 5, pp. 989-993, 1994.

[71] B. Laboratories, *Transmission systems for communications*: Western Electric co., 1971.

[72] Denon, "Anechoic Orchestral Music Recording." vol. PG-6006, 1988.

[73] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural room acoustic parameters from music and speech," *J. Acoust. Soc. Am.,* vol. 124, pp. 278-287, July 2008.

[74] P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers, and F. F. Li, "Room acoustic Parameter Extraction from music signals," in *IEEE ICASSP* Toulouse, 2006.

[75] J. Aldrich, "R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922," *Statistical Science,* vol. 12, pp. 162-176, 1997.

[76] L. Cremer and H. Müller, *Principles and Applications of Room Acoustics: Geometrical, Statistical and Psychological Room Acoustics* vol. 1: Spon Press, 1982.

[77] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America,* vol. 65, pp. 943-950, April 1979.

[78] J. D. Polack, "La transmission de l'égie sonore dans les salles," Universite du Maine, 1988.

[79] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model," in *103rd Conv. Audio Eng. Soc*, New York, 1997.

[80] E. Kreyszig, *Advanced engineering mathematics*, 7th ed.: John Wiley & Sons, Inc, 1993.

[81] V. Poor, *An introduction to signal detection and estimation*. New York Springer-Verlag, 1994.

[82] S. J. Apollo, "Maximum likelihood estimation of exponentials contained in signal-dependent noises," in *Department of Electrical Engineering*: The University of Texas at Arlington, 1991.

[83] R. Fletcher, *Practical Methods of Optimization*: John Wiley, 2000.

[84] T. J. Cox, P. Kendrick, F. F. Li, J. Chambers, and Y. Zhang, "Extracting room acoustic parameters from music," in *Proc. IoA. Auditorium Acoustics*, 2006.

[85] IEC 1260:1995 "Electroacoustics - Octave-band filters and fractional-octave band filters."

[86] T. Lokki, J. Pätynen, and V. Pulkki, "EAA Auralization Symposium" 2008: http://auralization.tkk.fi/

[87] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*: Cengage Learning, 2000.

[88] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, New York, pp. 209-243, 1963.

[89] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The Cepstrum: A Guide to Processing," *Proceedings of the IEEE,* vol. 65, pp. 1428-1443, October 1977.

[90] D. Bees, P. Kabal, and M. Blostein, "Application of the complex cepstrum to acoustic dereverberation," in *Proc. Biennial Symp. Commun.* , Kingston, pp. 324-327, 1990.

[91] R. F. Voss and J. Clarke, ""1/f noise" in music: Music from 1/f noise," *J Acoust Soc Am.,* vol. 63, pp. 258-263, 1978.

[92] J. R. Hopgood, "Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments," in *Signal Processing Laboratory ,University of Cambridge*. vol. PhD Cambridge: University of Cambridge, 2000.

[93] J. Beran. "Which aspects of music can be described by quantitative models? Music- Chaos, Fractals and Information", *Chance*, vol. 17 (4), pp. 7-16, 2004.

[94] B. L. Dalenbäck. "Room acoustic prediction based on a unified treatment of diffuse and specular reflection", *J. Acoust. Soc. Am.* Volume 100, Issue 2, pp. 899-909, 1996.

[95] R. Thiele . “Richtungsverteilung und Zeitfolge der Schallrückwürfe in Räumen”, *Acustica* 3, pp. 291, 1953