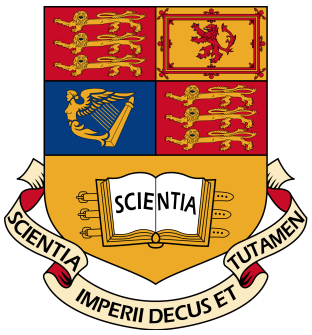
Roomprints:

Identifying a room from its sound characteristics

Final Year Project Report

Department of Electrical and Electronic Engineering
MEng Electrical and Electronic Engineering
Imperial College London

June 21, 2017



Aidan O. T. Hogg
aidan.hogg13@imperial.ac.uk
CID: 00733984

Project Supervisor: Dr Alastair Moore
Second Marker: Mr Mike Brookes

Abstract

Room identification is an important task due to the current trend in the use of location based multimedia applications. Current positioning methods available like GPS (Global Positioning System) are advantageous when it comes to determining which building the user is located in. These techniques, however, lack accuracy indoors, this is why methods to identify rooms inside of buildings are appealing. The benefits of being able to utilise acoustics is that no external hardware, such as Bluetooth beacons, are required. It also reduces concerns surrounding privacy in places where using visual identification is inappropriate.

The aim of the project is to investigate the viability of room identification using acoustics where properties of the room are estimated using a supervised scheme. This will involve implementing algorithms which estimate acoustic parameters of rooms in a supervised scenario. The success of these algorithms will be validated first on simulated synthetic data then on existing data and, finally, on newly collected audio data. The parameter of interest in this project was the reverberation time at different frequency subbands which was used for room classification. The project ultimately involved evaluating how practical it was using a setup of a speaker and microphone to recognise a room using only speech.

Acknowledgements

First and foremost I'd like to thank my project supervisor Dr Alastair Moore for always finding time to give me his very useful advice and guidance. I would also like to thank Dr Patrick Naylor for making this exciting project available to me. I am also very grateful to the Department of Electronic and Electrical Engineering for providing me with the tools needed to carry out my project. Finally, I would like to thank my friends and family for always being there for me through the good times and the bad.

Soli Deo gloria

Table of contents

Abbreviations

ACE acoustic characterisation of environments.

C_{80} clarity index.

DFT discrete Fourier transform.

EDT early decay time.

FD frequency-domain.

FFT Fast Fourier transform.

HPF high pass filter.

ISM image-source model.

LTI linear-time-invariant.

NPM normalized projection misalignment.

RIR room impulse response.

RT reverberation time.

RTF room transfer function.

SNR signal-to-noise ratio.

SSNR segmental signal-to-noise ratio.

1 Introduction

The word 'perception' is used to describe the multiple ways in which people collect information from their surroundings and thus allowing them to know their environment. We as humans mainly rely on visual perception to identify the environment in which we find ourselves. This is why so much research over the last few decades has been invested into computer vision so that computers can be trained to see in a similar way to humans. The eyes, however, are not the only sense that people have at their disposal when it comes to identifying the environment around them. Humans are also well adept at sound perception; the study of which is called, 'psychoacoustics'.

If a person was to enter a room blindfolded and speak they would probably be able to guess what sort of room they were in just by what they heard. Most people have had the joy of shouting loudly in an environment with a lot of reverberation (e.g. a cave) and seeing how long the echo lasts. Most of us probably have an intuitive feeling that just by speaking and listening we could probably tell whether we were standing in a cathedral or a cupboard. People encounter multiple complex audio scenes everyday and are able to make sense of them.

Humans to some degree can identify a room or an environment using only sound, a form of automaticity, then surely it would then be possible for a machine (e.g. a robot or mobile device) to perform the same task? Indeed, it is possible and a lot of research has been carried out in the field of audio signal processing to this end. This project will therefore investigate the viability of room identification using acoustics where properties of the room are estimated using a supervised method and modify existing algorithms to improve performance in this particular situation.

Room identification is an important task for many applications including robotics and hearing aids. This is because factors such as filtering in hearing aids may vary depending on the room the wearer is in. The filtering required in a living room, where most people chat and watch television aims to cut out background noise. This, however, will be very different to the kitchen, where background noise may be important and, therefore, should not be filtered out. There is also a current trend in the use of location based multimedia applications. These applications currently depend on GPS (The Global Positioning System) to determine their position. This is, however, inadequate when location identification is required indoors. One solution that is currently being explored to increase positioning accuracy is the use of WiFi strength. This solution nevertheless requires the device to support this technology and if it does not then it is not possible to estimate the location. It would be very useful, therefore, to be able to identify a room solely on acoustic properties and without the necessity of extra hardware (i.e. Bluetooth beacons).

In 1997, Sawhney and Maes[1] were the first to focus solely on auditory room classification; until this point it was mostly conducted as background. Since then many more algorithms have been developed for the function of environment identification[2][3][4][5][6]. This report focuses on this concept in the context of room identification. Can a room be identified by its audio characteristics? The current methods that exist, the latest being the estimation of the direct-path relative transfer[7] (2016), lack robustness. This is a major shortcoming when it comes to unsupervised estimation methods. This project aims to explore a supervised approach to classifying the rooms. The setup that was envisaged used speech signals along with a microphone and a loudspeaker to identify rooms. A big emphasis was placed on the robustness of the implementation.

1.1 Supervised estimation of room acoustic characteristics

Supervised estimation of room audio characteristics corresponds to a situation where the estimations are made using the known transmitted signal and the recording of the reverberated signal from the room. This would be analogous to a robot entering a room speaking and then listening to the output. The audio characteristics of a room are usually determined by inspecting the room impulse response (RIR). The RIR is comparable to a 'fingerprint' or, in this case, a 'roomprint' of a particular room. The name roomprint, also known as a 'quantifiable description of an acoustic environment'[8], comes from an analogy to fingerprints where a database of reference roomprints can be collected which is the equivalent to a database of rolled fingerprints. Then these reference roomprints can be compared with a roomprint derived from an environment under test, just as a latent fingerprint would be compared to a database of reference fingerprints. The RIR is a recording of what it would sound like if a loud and short click were played. This is why sometimes balloon pops[9] or starter pistol shots are used to calculate the RIR. If loudspeakers are being used, then an exponential sine-sweep method is generally more appropriate due to the fact that it yields more accurate and more reliable results. This exponential sine-sweep method was used in this project by a way of evaluating the accuracy of an RIR estimation from a speech signal. The RIR depends both on the position of the audio source and the audio receiver which means no RIR within a room is completely similar to any other. However, certain audio characteristics contained in the RIR are constant for a specific room, for example, the reverberation time (RT).

There is, however, one major practical disadvantage to this approach. In order to gather useful results, the room in question would need to be occupied to its normal capacity. This is because human bodies absorb sound and therefore the level of occupancy of a room can severely alter the room's acoustics.

Hidaka and Nishihara published a paper[10] (2000) which looked at five different audio characteristics which include: reverberation time (RT), early decay time (EDT) and clarity index (C_{80}). Hidaka and Nishihara showed that RT can vary as a result of the occupancy level. They found that RT differs by 250ms on short RIRs and 2000ms on long RIRs. A significant impact on the EDT and C_{80} measures was also observed, which for the EDT measure is to be expected as it is correlated with the RT. The C_{80} measure also showed that a high occupancy level can reduce the clarity by 1.4dB. These results show that although measurements are usually taken in an unoccupied room, the occupancy level needs to be taken into account when investigating room classification. To limit the scope of this project, occupancy will not be taken into account and, therefore, the assumption that the occupancy level is negligible when calculating the RTs at different octave frequency bands will be accepted.

The main reason this is still an active area of research is the complexity of sound propagation in natural acoustic environments. Acoustical signals have a range of frequencies from 20Hz to 20kHz and a room's impulse response can have thousands of coefficients. This makes the estimation of a room impulse response (RIR) very difficult. This problem is exacerbated by the presence of noise.

This project aims to tackle these problems by finding a robust solution to room identification in a supervised situation. It focuses on how well a room can be classified using a supervised approach that first estimates the RIR and uses this result to obtain audio characteristics of a room which can then be used for identification. The main focus will be on achieving this task using speech as

the test signal and exploring the minimum conditions required.

An additional outcome of this report would be to prove that a robot could store a database of such roomprints. These reference roomprints would then be consulted when room identification was required. This process would ultimately lead to correct classification using machine learning techniques.

1.2 Aim

The aim of this project is to explore the hypothesis:

Can a room be identified by its observable acoustic characteristics in a supervised scenario?

This is to build on research already carried out in this field to look at the viability of different methods and limitations involved when performing room identifications.

1.3 Objectives

The objectives for this project can be broken down into a number of specific goals.

Implement algorithms

This project will use a supervised method to estimate the RIR of a room. This means that the output speech signal is known and the RIR is calculated by using one microphone. This approach can be undertaken in many different ways. The algorithms that will be implemented in MATLAB exploit the time and frequency domain. The advantage to this supervised approach is that it is a very robust method.

Evaluate algorithms using simulated signals for simulated RIR data

Synthetic room impulse responses (RIRs) will be used for verifying the implemented algorithms in the context of a supervised scenario. This evaluation will be performed first due to the possibility of controlling most aspects of the simulated RIR data. This allows parts of the problem to be examined in isolation, such as the effects of: noise contamination; length of the filter being estimated and length of speech used.

Evaluate algorithms using simulated signals for measured RIR data

The supervised method will then be tested on real room impulse responses acquired from the acoustic characterisation of environments (ACE) corpus[1]. This process will verify the validity of the selected algorithm on real measured data which is likely to be more complex and realistic than simulated RIR data.

Evaluate algorithms using real recordings

Data will be collected using the microphone and speaker setup that reflects a real-world scenario. This data will then be used for offline processing in MATLAB to evaluate how well the selected algorithms perform in a practical situation. The data will be gathered from 5 different rooms from 4 different positions.

Evaluate effectiveness of algorithms in room identification task

The supervised approach will then be evaluated by its ability to perform classification using the estimated audio characteristics of the room. This will involve looking at the performance and

[1] http://www.ee.ic.ac.uk/naylor/ACEweb/

results observed throughout the project and verifying the robustness of the approach that was pursued.

1.4 Report Structure

The rest of the project report is structured as follows:

Chapter 2: Background - This section will cover the basic background required for this project including concepts, algorithms and evaluation metrics to be used.

Chapter 3: Implementation and Results - This section will detail the implementation of each algorithm. It will also look at the performance of the supervised approach when evaluated using: simulated synthetic data; existing data and newly collected audio data.

Chapter 4: Conclusion - This section will assess how well the aim (Can a room be identified by its observable acoustic characteristics in a supervised scenario?) has been answered in the course of this project by outlining the validity of using a supervised approach for room identification. It will also give an overview of the final thoughts in regard to this project and how it could be taken further in the future.

2 Background

This chapter will cover the background required to define a room's acoustic characteristics. This will mainly focus on the monaural reverberation time (RT). It will also introduce several methods that are used to calculate the room impulse response (RIR) from which multiple parameters can be determined. Finally, it will look at different evaluation metrics, one being the concept of the normalized projection misalignment (NPM) metric which can be used to measure how accurately the room impulse response (RIR) has been estimated. The second being machine learning classifiers where room identification can be evaluated.

2.1 Roomprint

First, a roomprint is defined[8] as 'a quantifiable description of an acoustic environment' which can be measured under controlled conditions and estimated from a monophonic recording made in that space. A roomprint must therefore utilise room features that allow the room to be distinguished from other similar rooms.

In [8] the feature chosen is the reverberation time (T_{60}) as this does not require a special microphone arrangement and does not depend on the source directivity and orientation. In fact, the frequency-dependent reverberation time was explored due to the the bass ratio (ratio of low and high frequency reverberation times) being a source of dissimilarity between different rooms.

In this project, the reverberation time (T_{60}) (see section 2.3) will be used for room identification. However, in this case it will be calculated from the estimated room impulse response (see section 2.2).

2.2 Room impulse response

Rooms are acoustic spaces which can be modelled as passive linear-time-invariant (LTI) systems.

Figure 2.1: Room response from a sound source

LTI systems have two properties:

1. **Linear**: $\mathscr{H}(\alpha u[n] + \beta v[n]) = \alpha\mathscr{H}(u[n]) + \beta\mathscr{H}(v[n])$

   A linear system maps an input to an output using only linear operations

2. **Time invariant**: $y[n] = \mathscr{H}(x[n]) \Rightarrow y[n-r] = \mathscr{H}(x[n-r])\forall r$

   A time invariant system means that the system's response is not changed but only delayed if the input to the system is delayed

The response of a room to a sound source can be seen in figure 2.1 The behaviour of an LTI system is completely defined by its impulse response: $h[n] = \mathscr{H}(\delta[n])$. This implies that the room

response can be quantified by the RIR which is the room's response to a Dirac delta function ($\delta[n]$).

In the discrete domain Dirac's delta function (impulse) is defined by the following property:

$$\delta[n] = \begin{cases} 0, & n \neq 0 \\ 1, & n = 0 \end{cases} \tag{2.1}$$

The response of a room to a sound source is the convolution RIR $h[n]$ with the source signal $x[n]$.

$$y[n] = x[n] * h[n] \tag{2.2}$$

This property can be viewed in the frequency domain by using a feature of the circular convolution. This feature is that the circular convolution is the multiplication of the DFT of the RIR and the DFT of the input signal in the frequency domain.

$$\begin{aligned} y[n] &= x[n] \circledast h[n] \\ Y(\omega) &= X(\omega)H(\omega) \end{aligned} \tag{2.3}$$

This is important due to the fact it is possible to make the circular convolution of x[n] and h[n] a linear convolution if we append zeros onto x[n].

The room impulse response (RIR) completely defines the room's response to a single audio source with respect to a single sound receiver where both are in particular positions. The room transfer function (RTF) is the frequency-domain (FD) equivalent. This is based on the assumption that a room is equivalent to an LTI system. In reality, time variance is caused mainly by temperature but the changes are sufficiently slow that the model is widely used.

Figure 2.2 shows an RIR. The direct-path is shown in red, which is the short initial period of non-zero amplitude. The amplitude of this direct-path peak maybe larger or smaller than the amplitude of later reflections, depending on the positions of the source and the receiver. It also depends on the absorption coefficients of the room walls. In figure 2.2 the direct-path component is very large which implies that the distance between the microphone and the speaker was short when this RIR was measured; this was indeed the case.

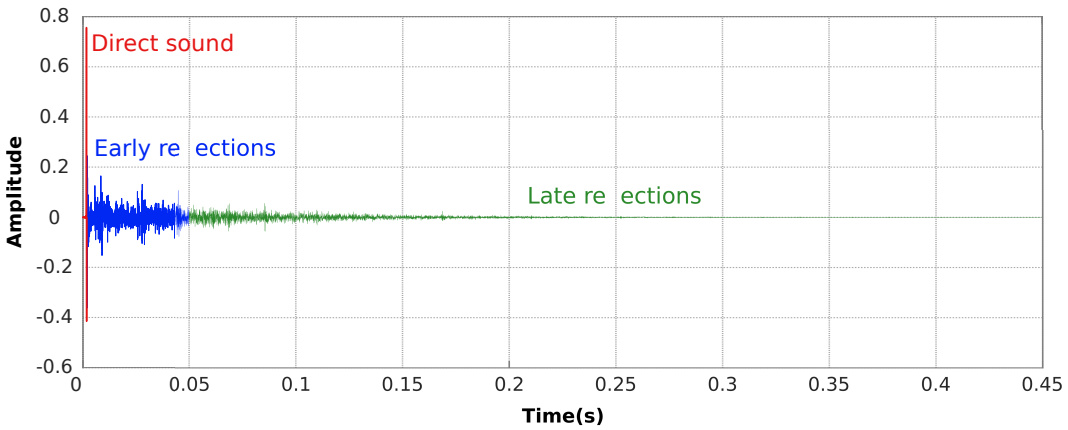


Figure 2.2: Room impulse response example

Figure 2.2 also shows the early and late reflections in blue and red respectively. The early reflections are taken to be the reflections that occurred in the first 50ms[11] of the RIR. These are normally well defined impulses in comparison with the late reflections, sometimes called the reverberation tail, which are more diffused in nature. The reverberation tail is the cause of the 'echoey' sound which one comes across in one's everyday experience of room acoustics.

2.2.1 Synthetic room impulse responses

This project heavily utilises the RIR to calculate audio characteristics of the room under test. It is, therefore, desirable to be able to generate realistic synthetic room impulse responses. These synthetic RIRs can then be used extensively to evaluate different methods which achieve room classification; the advantage being that all the characteristics of a synthetically constructed room are known and thus can be compared with their estimates.

The image-source model (ISM)[12] is a commonly used method for generating a synthetic room impulse response (RIR). When viewed in the frequency-domain, this is the RTF between a sound source and an acoustic sensor within a given room. This RIR can then be used to calculate the reverberation time (RT) of the specified room.

This method was first proposed by Allen & Berkley in a revolutionary paper[12] in 1979. Their method, however, had a few disadvantages due to the way that the image-source model (ISM) was implemented. An improvement to this original ISM implementation was put forward by Peterson[13] in 1986. Then in 2008 Lehmann & Johansson[14] improved the ISM algorithm which addressed the drawbacks for the earlier ISM implementations. This improved ISM method[1] was used in this project to create synthetic RIR's.

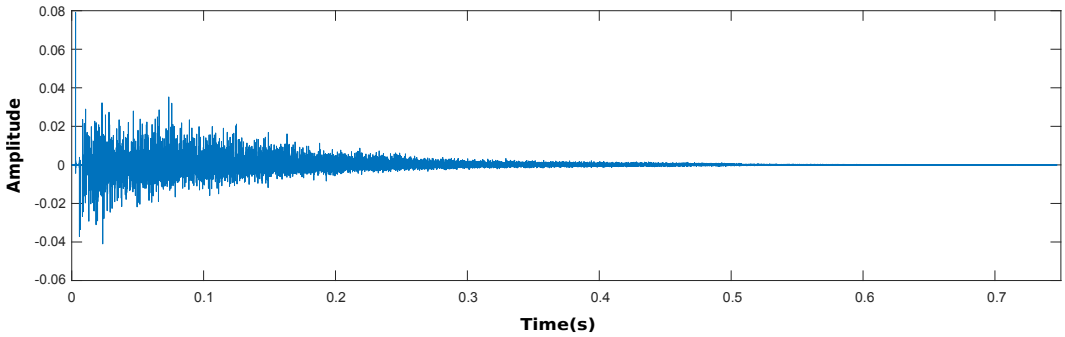


Figure 2.3: Typical RIR obtained using the ISM method

This technique makes it possible to simulate a realistic speech signal using the ISM generated RIR. This is accomplished by artificially adding noise to the clean speech signal after it has been convolved with a RIR created using the ISM method. This is later shown in 2.6 where x is the realistic speech signal, convolution is denoted by *, y is the additive noise and the RIR generated using the ISM is denoted by h.

2.3 Reverberation time (RT)

It is possible from the room impulse response (RIR) to calculate the reverberation time (RT) which is the parameter of focus in this report.

[1] http://www.eric-lehmann.com/ism_code.html

Firstly, it is also important to note that the properties of rooms have a major impact on how a sound source (i.e voice) will be heard by a sensor (i.e human ear). This is due to the fact that reflections of the sound waves are created at the room's boundaries and by objects within the room. This means that when a speech signal is produced in a room by a sound source then the listener will hear a signal consisting of the superposition of many delayed and attenuated copies of the original speech signal produced in the room. Figure 2.4 shows an example of these reflections from a given sound source.

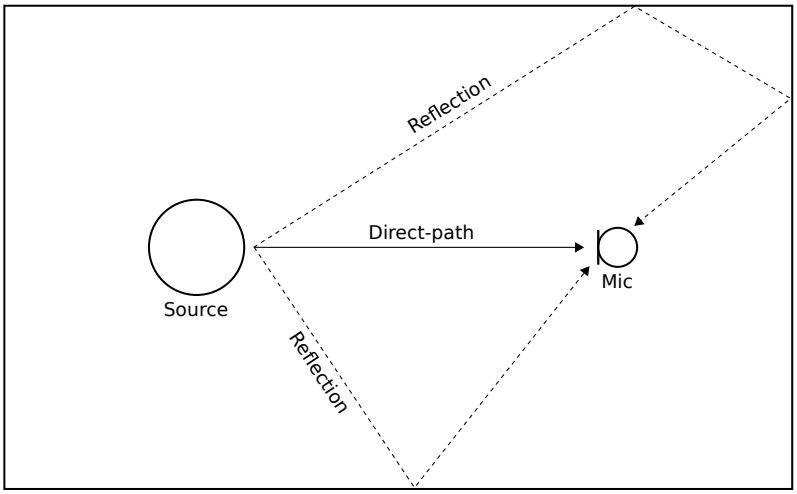


Figure 2.4: Illustration of room reverberation

In acoustics, 'free field' is a situation where these reflections do not exist and the transmission of sound occurs only via the direct path. This is where the direct-path is defined as the acoustic propagation path from the talker to the listener without any reflections.

The reverberation time (RT) is an important quantification of a room. It was first developed in the late 19th century by Sabine[15]. By using organ pipes as a sound source, Sabine discovered that the time for the sound to become inaudible after the pipe was turned off was related to the size of the room and the amount of absorption inside it. This work resulted in the RT measure which is defined as the time taken for the reverberant energy to decay by 60dB once the sound source has been abruptly turned off.

The reason that this is such a useful measure is that within any given room, independent of the amplitude and location of the sound source, the RT is constant. Therefore, although the RIR changes depending on the positions of the sound source (speaker) and sound sensor (microphone) used to estimate the RIR; the RT remains the same for each given room.

2.4 Room impulse response measurement methods

There are multiple methods that can be deployed to estimate the room impulse response. This section will introduce some of these techniques with their advantages and drawbacks.

2.4.1 Exponential sine-sweep method

This project requires a method to accurately measure the RIR in order to evaluate how good the RIR estimation is when using speech signals. This exponential sine-sweep method[16][17] was used for this purpose. The advantage of this method being that it overcomes the problems of non-linearities in loudspeakers which will used to measure the RIR.

This technique works by using a sine wave, whose frequencies exponentially increase over time to give greater emphasis to the low frequencies. This is due to the fact that there is often a lower signal-to-noise ratio (SNR) at lower frequencies.

The signal-to-noise ratio (SNR) is defined as the ratio of the power of a signal to the power of background noise. Acoustic signals have a wide, dynamic range and, therefore, often expressed using the logarithmic decibel scale.

$$\begin{aligned} \text{SNR} &= \frac{P_{\text{signal}}}{P_{\text{noise}}} \\ \text{SNR}_{dB} &= 10\log_{10}(\text{SNR}) \end{aligned} \tag{2.4}$$

In equation 2.5 $x(t)$ is a band-limited sinusoidal sweep signal where the frequency varies exponentially from f_1 to f_2.

$$x(t) = \sin\left[\frac{2\pi f_1 T}{\ln\left(\frac{f_2}{f_1}\right)}\left(e^{\frac{t}{T}\ln\left(\frac{f_2}{f_1}\right)} - 1\right)\right] \tag{2.5}$$

The output signal used in this project to obtain the RIR from the exponential sine-sweep method is shown in figure 2.5

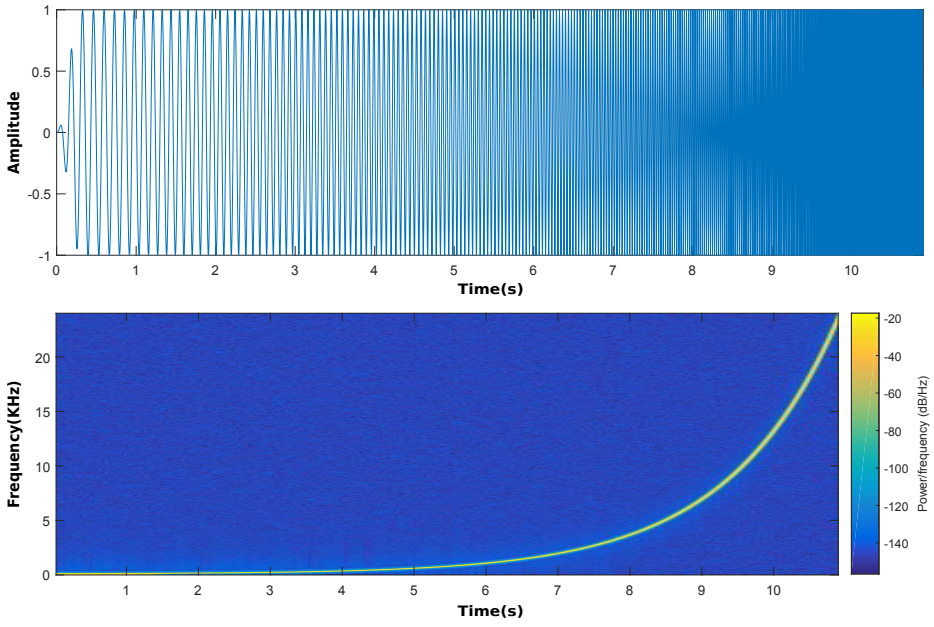


Figure 2.5: Output signal - Exponential sine-sweep

The inverse filter used in this project, which is the equalised time-reversed output signal, is shown in figure 2.6.

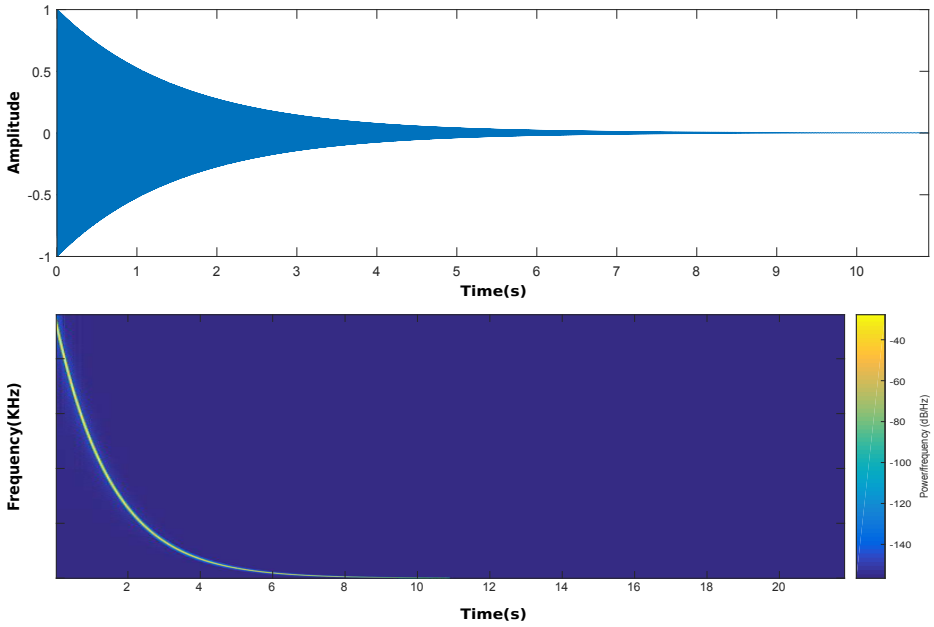


Figure 2.6: Result of the deconvolution

In figure 2.7 the non-linear behaviour of the loudspeaker causes a few harmonics to appear. The advantage to this method is that the loudspeaker distortions result in harmonic distortions at higher multiples of the excitation frequencies. Therefore, these harmonic distortions can be totally removed and thus the non-linearities caused by the loudspeaker when making the measurement do not effect the subsequently calculated RIR.

The RIR is then obtained from the recording of the exponential sine-sweep by convolving the measured signal (the recording of the exponential sine-sweep, see figure 2.7) with the inverse filter which is the equalised, time-reversed output signal (see figure 2.6). The output of this process, which is named the deconvolution, can be seen in figure 2.8.

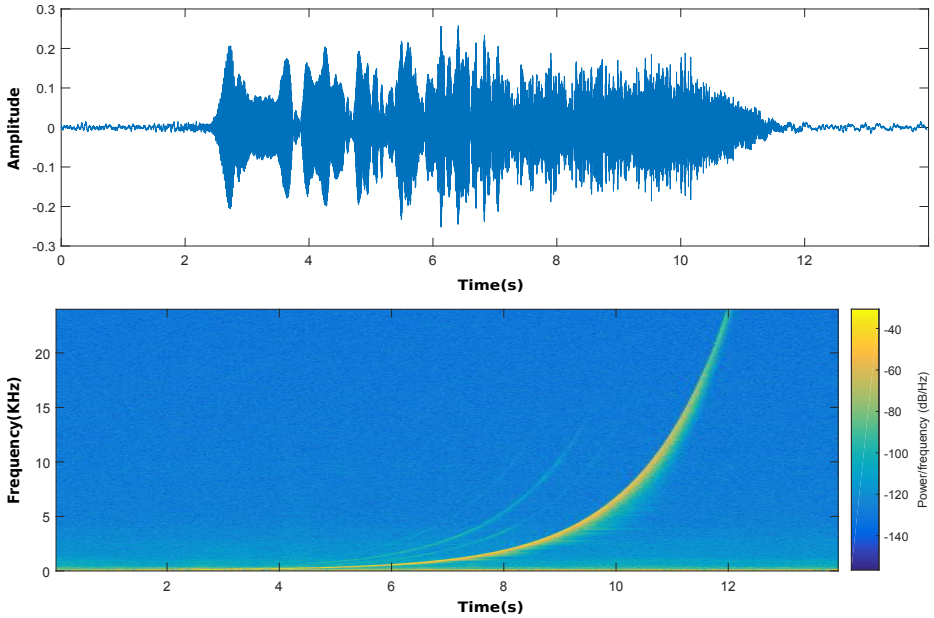


Figure 2.7: Measured signal - Recorded exponential sine-sweep

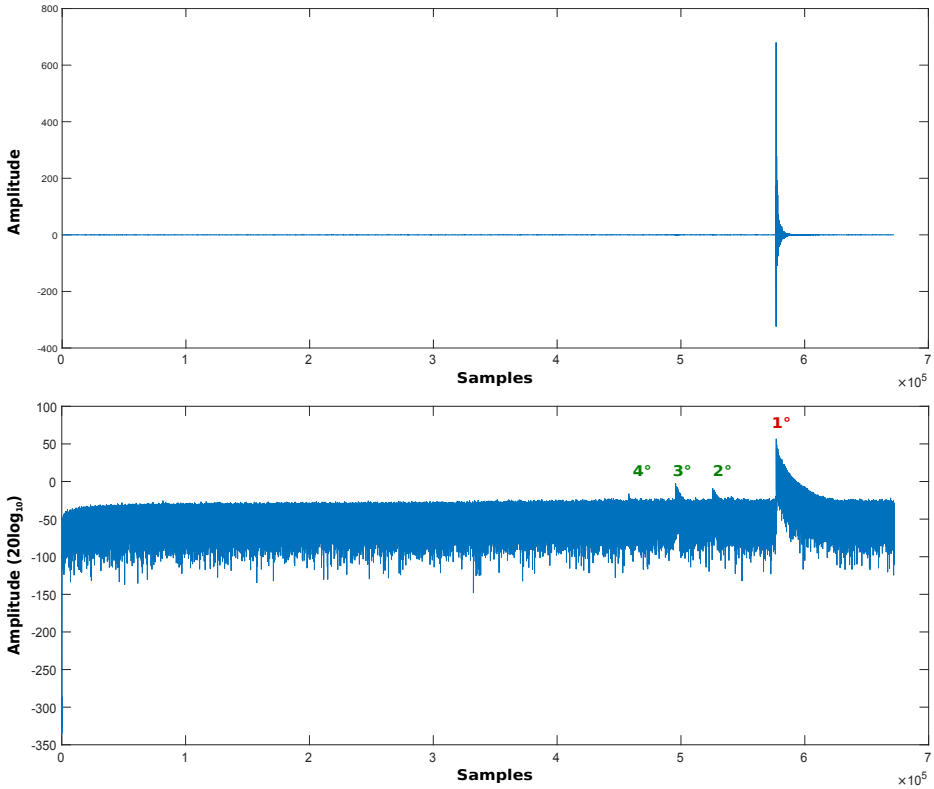


Figure 2.8: Result of recorded sine-sweep convolved with the inverse filter (Top) $20log_{10}$ of recorded sine-sweep convolved with the inverse filter (bottom)

In figure 2.8 the last impulse response is the linear one (red); any preceding ones (green) are the harmonics distortion products of various orders.

2.4.2 Estimation of RIRs using speech signals

In this project a supervised RIR estimation approach is pursued, where the speech signal at the loudspeaker is known, such that robustness to noise at the microphone is the main focus. This can be achieved through the sine-sweep method, however, this is an intrusive procedure on any external listener in the room, whereas using speech signals would be far more pleasant to any listener.

This project looks at a supervised approach where only one microphone is used and the target source is known. This is an ideal scenario where only noise contamination will prevent a near perfect estimation of the room impulse response.

Problem description

Therefore to analyse the problem we will assume that there is one microphone available and the target source (output from speaker) is known.

A target source x(n) is described by:

$$x(n) = \{h(n) * s(n)\} + y(n) \tag{2.6}$$

Where the time index is denoted by n taking values from 1 to N, x is the input signal from the microphone, convolution is denoted by *, y is the contaminating noise and the microphone-target acoustical impulse responses are then denoted by h.

The short-term frequency domain equivalent description of (2.6) is:

$$X(\theta, \ell) = H(\theta)S(\theta, \ell) + Y(\theta, \ell) \tag{2.7}$$

where frequency is denoted by θ and the frame index is denoted by ℓ.

Time-domain estimation using least squares

If we assume noise-free conditions then we have to solve equation:

$$x(n) = h(n) * s(n) \tag{2.8}$$

This could be represented in a matrix form:

$$\mathbf{x} = \mathbf{Sh} \tag{2.9}$$

Where $\mathbf{h}$ is a vector of L coefficients of h, $\mathbf{x}$ is a vector of $N + L - 1$ coefficients of x and:

$$\mathbf{S} = \begin{bmatrix} s(1) & 0 & \cdots & 0 \\ s(2) & s(1) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ s(N) & s(N-1) & \cdots & s(N-L+1) \\ 0 & s(N) & \cdots & s(N-L+2) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & s(N) \end{bmatrix}$$

Note: This is due to the fact that circular convolution can be performed by a circulant matrix multiplied by a vector representation of the other signal. It is also possible to make the circular convolution of s(n) and h(n) a linear convolution if we append $L-1$ zeros onto s(n).

Therefore we can solve this equation by $\mathbf{h} = \mathbf{S}^{-1}\mathbf{x}$. However, we have an overdetermined set of equations ($\mathbf{S}$ is not of full rank) then the inverse $\mathbf{S}$ does not exist.

This is why least squares method[2] maybe be used to estimate the first L coefficients of h as follows:

Therefore we want to minimize $\|\mathbf{e}\|^2$ where $\mathbf{e} = \mathbf{Sh} - \mathbf{x}$ which is the case if:

$$\mathbf{h}_{LS} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{x} = \mathbf{R}^{-1}\mathbf{p} \tag{2.10}$$

where

$$\mathbf{R} = \mathbf{S}^T\mathbf{S}, \mathbf{p} = \mathbf{S}^T\mathbf{x} \tag{2.11}$$

Frequency-domain estimation

If noise-free conditions are assumed then in the short-term frequency-domain the equation 2.12 has to be solved.

$$X(\theta, \ell) = H(\theta)S(\theta, \ell) \tag{2.12}$$

where frequency is denoted by θ and the frame index is denoted by ℓ.

A simple conventional frequency-domain method[3] to estimate the RTF (H) is calculated by:

$$\hat{H}(\theta) = \frac{\sum_\ell \overline{S(\theta,\ell)} X(\theta,\ell)}{\sum_\ell |S(\theta,\ell)|^2} \tag{2.13}$$

Frequency-domain estimator using nonstationarity

Another frequency-domain estimator was proposed by S. Gannot *et al*[4] that is able to perform with noisy signals.

If we now allow the presence of noise, the short-term frequency domain equivalent description of (2.6) is:

$$X(\theta,\ell) = H(\theta)S(\theta,\ell) + Y(\theta,\ell) \tag{2.14}$$

where frequency is denoted by θ and the frame index is denoted by ℓ.

Assume this model is valid for a particular interval where H is approximately constant. Then if the interval is split into k frames by considering the k^{th} frame we have from equation 2.14:

$$\begin{gathered} \Phi_{XS}^{(k)}(\theta) = H(\theta)\Phi_{SS}^{(k)}(\theta) + \Phi_{YS}^{(k)} \\ k = 1, ..., K \end{gathered} \tag{2.15}$$

where K is the number of frames used, $\Phi_{AB}^{(k)}$ denotes the cross power spectral density between A and B during the k^{th} frame.

$$\begin{bmatrix} \Phi_{XS}^{(1)}(\theta) \\ \vdots \\ \Phi_{XS}^{(K)}(\theta) \end{bmatrix} = \begin{bmatrix} \Phi_{SS}^{(1)}(\theta) & 1 \\ \vdots & \vdots \\ \Phi_{SS}^{(K)}(\theta) & 1 \end{bmatrix} \begin{bmatrix} H(\theta) \\ \Phi_{YS}(\theta) \end{bmatrix} \tag{2.16}$$

The estimate of $H(\theta)$ can now be calculated by replacing the cross-PSDs in (2.16) by their sample-based estimates. Thus by using least squares the overdetermined system of equations can be solved.

The solution to (2.16) is given by:

$$H(\theta) = \frac{\langle \Phi_{SS}(\theta)\Phi_{XS}(\theta) \rangle - \langle \Phi_{SS}(\theta) \rangle \langle \Phi_{XS}(\theta) \rangle}{\langle \Phi_{SS}^2(\theta) \rangle - \langle \Phi_{SS}(\theta) \rangle^2} \tag{2.17}$$

2.5 Reverberation time measurement methods

A paper written by Alastair H. Moore *et al* in 2013[8] showed that a a room can be identified by its reverberation time. This report focuses on a different way of calculating the RT measure by first estimating the room impulse response (RIR). This section will look at how the RT metric can be estimated from the RIR and the different approaches available.

2.5.1 Third-octave frequency band processing

The reverberation time (RT) is obtained from the room impulse response and can be stated as a single value if it is measured as a wide band signal. This approach will, however, make classification very difficult as many rooms have similar RTs. A solution to this problem can be found in the fact the RT is frequency dependant. It, therefore, makes sense to calculate the RT in third-octave frequency bands. This will give a more precise result that is suitable for classification as the narrow frequency bands will differ more between rooms depending on the frequency band being measured.

In this project Christophe Couvreur's matlab code[2] was consulted to calculate the RT at third-octave frequency bands.

Schroeder integration

The RT is calculated using the reverse time integration of the squared RIR which has been filtered into octave frequency bands (section 2.5.1). The reverse time integration is also referred to as the Schroeder integration due to the fact that it was developed by Schroeder in 1967[18]. It is based on a simple method but can be challenging to perform accurately.

It works by starting at the end of RIR and moving backwards to the beginning while adding up the squares of each sample in the RIR in the process. The obstacle to this method is the noise floor. The integration flattens when the reverberant decay slope reaches the noise floor which can cause an overestimation of the RT. This is exacerbated when the RIR has a small dynamic range or a long reverberation tail.

One clear solution to this problem is to identify the intersection between the reverberant decay slope and the noise floor, commonly called the saddle point. Then it is possible to start Schroeder integration from this point instead of at the end of the RIR. This can be seen in green on figure 2.9. This saddle point, however, is extremely difficult to estimate which is why this method can be difficult to perform accurately.

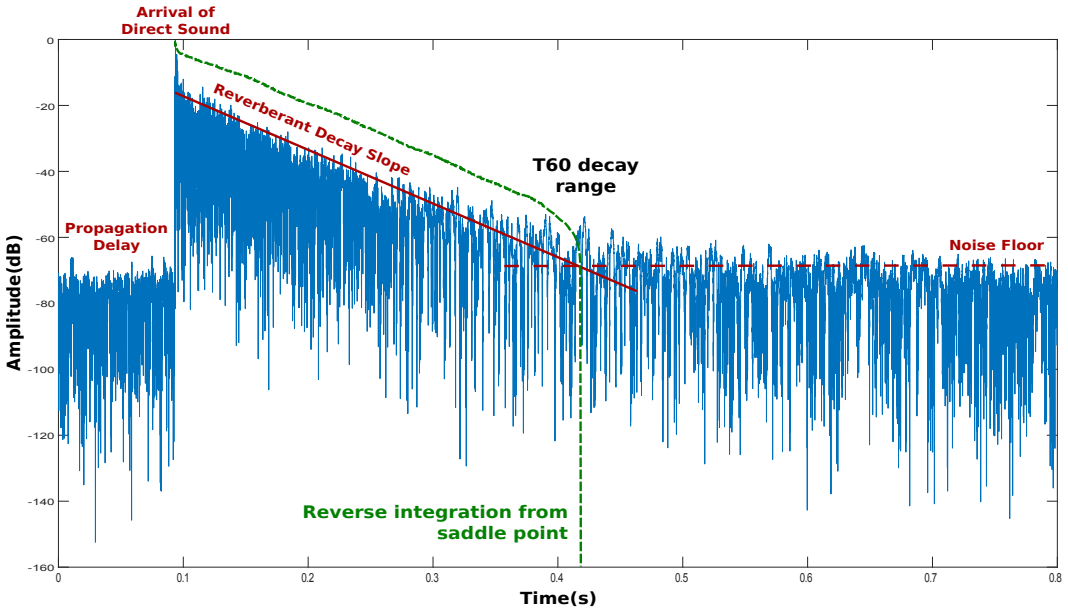


Figure 2.9: Estimating reverberation time by reverse integration of the squared RIR

[2] https://uk.mathworks.com/matlabcentral/fileexchange/69-octave

The reverse time integration can be used to estimate the RT by interpolation of the reverberant decay slope. This can be achieved in a linear or non-linear manner.

Linear approach

If a linear approach is taken, then the interpolation of the reverberant decay slope is performed with linear function: $L = At + B$. Then the RT (T60) depends on the gradient of this linear line L and is estimated by equation 2.18.

$$RT = \frac{-60}{A} \tag{2.18}$$

Non-linear approach

It is also possible to calculate the interpolation of the reverberant decay slope using a non-linear function[19]. This approach tends to give more consistent results which is why this approach was used for the majority of this project. It was also the approach that was used in the ACE challenge[20]

2.6 Evaluation metrics

This project requires that the methods taken are evaluated to compare their performance. This section will explore these evaluation metrics. It will start by looking at the normalized projection misalignment measure that is used to evaluate the similarity between two different room impulse responses. It will then explain the different machine learning approaches that will be used for room identification. Room identification is the main goal of the project and, therefore, the percentage of rooms that can be correctly classified is of absolute importance.

2.6.1 Normalized projection misalignment

This project requires a method of evaluating performance of RIR estimators. If we have an estimated impulse response $\hat{\mathbf{h}} = [\hat{h}_0, \hat{h}_1, ..., \hat{h}_{L-1}]^T$ given by an estimator and a true impulse response $\mathbf{h} = [h_0, h_1, ..., h_{L-1}]^T$ then a suitable error measure can be given by:

$$\begin{aligned} \xi(\mathbf{h}, \hat{\mathbf{h}}) &= \min_{\beta} \frac{\|\mathbf{h} - \beta\hat{\mathbf{h}}\|^2}{\|\mathbf{h}\|^2} \\ &= 1 - \left(\frac{\mathbf{h}^T\hat{\mathbf{h}}}{\|\mathbf{h}\|\|\hat{\mathbf{h}}\|} \right)^2 \end{aligned} \tag{2.19}$$

This error is obtained by minimizing overall possible gain values β. The error can also be viewed from a geometric perspective as the normalized minimum squared distance from $\mathbf{h}$ to the linear manifold of $\hat{\mathbf{h}}$, which is obtained by projecting $\mathbf{h}$ onto $\hat{\mathbf{h}}$. Thus this important measure is normally given the name normalized projection misalignment (NPM)[21].

2.6.2 Machine learning classifiers

This project will use the reverberation time (RT) to classify each given room. To achieve this machine learning techniques will be exploited. The RT will be calculated at different frequency bands which will constitute the different features that will be used for classification. This section will focus on the different classifiers that were used to ultimately perform room identification.

Naive Bayes classifier

The naive Bayes classifier[22] is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. The naive Bayes classifier has been studied

extensively since the 1950s.

Bayes' theorem was named after Thomas Bayes (1701–1761) and states:

$$\mathsf{P}\big(c_j \mid d\big) = \frac{\mathsf{P}\big(d \mid c_j\big)\mathsf{P}\big(c_j\big)}{\mathsf{P}\big(d\big)} \tag{2.20}$$

The Bayes classifier estimates $\mathsf{P}\big(d|c_j\big)$ (likelihood) and $\mathsf{P}\big(c_j\big)$ (prior) in equation 2.20 while learning from training data. Then during classification it uses equation 2.20 to calculate $\mathsf{P}\big(c_j \mid d\big)$.

The (naive) independence assumption of the naive Bayes classifier is that features are independent given class:

$$\begin{aligned}\mathsf{P}\big(d_1, d_2 \mid c_j\big) &= \mathsf{P}\big(d_1 \mid d_2, c_j\big)\mathsf{P}\big(d_2 \mid c_j\big) \\ &= \mathsf{P}\big(d_1 \mid c_j\big)\mathsf{P}\big(d_2 \mid c_j\big)\end{aligned} \tag{2.21}$$

More generally:

$$\mathsf{P}\big(d_1 ... d_n \mid c_j\big) = \prod_{i=1}^{n} \mathsf{P}\big(d_i \mid c_j\big) \tag{2.22}$$

The decision rule for the naive Bayes classifier is:

$$h_{NB}(d) = \arg\max_{c_j} \mathsf{P}\big(c_j\big) \prod_{i=1}^{n} \mathsf{P}\big(d_i \mid c_j\big) \tag{2.23}$$

If the assumption holds then the naive Bayes classifier is an optimal classifier.

Training and testing the naive Bayes classifier

This section presents key details regarding the training and testing the naive Bayes classifier[23].

K-fold Cross Validation

K-fold cross validation is where the data is divided into k subsets where one of the k subsets is used as the test set and the other k-1 subsets are used to form the training set. This process is repeated k times using a different subset for the test set each time. The higher the value of k, the lower the variance of the resulting estimate. Then the average error across all k trials is calculated. The advantage to using this method is that the data divisions matter less due to the fact that each data point gets tested exactly once and is used in the training set k-1 times. The downside being that the training algorithm has to be rerun k times which is high in computational cost.

Leave-one-out cross-validation

Leave-one-out cross-validation is the same as K-fold cross validation where k, the number of subsets, is equal to n, the number of data points in the set. This means the classifier is trained n times on all the data except for one point. This is an advantageous method when a limited amount of data is available, however, it is very computationally intensive, which is not a concern for this project.

3 Implementation and Results

This section focuses on the implementation of the supervised method to identify a room by its audio characteristics. This involves estimating a room impulse response (RIR) from speech signals and then using this estimate to determine the reverberation time (RT) of the room. The section starts by verifying these estimation techniques on synthetic rooms to prove that the theory works in practice. It then looks at real data that was taken from the ACE corpus to prove that the approach still performs well when applied to real RIRs. To conclude, this section experimental data that was collected for the purpose of this project, is explored to demonstrate that the algorithms implemented work in a realistic scenario.

3.1 Synthetic data

The synthetic rooms, used in this section, were generated by means of the image-source model (ISM) approach. This way all the parameters of the room were known so a comprehensive evaluation could be performed. Figure 3.1 shows a synthetic room that will be adopted for assessment purposes. The parameters of the room are as follows: room dimensions = 3m, 4m, 2.5m; position of microphone = 2.6m, 1m, 2.3m; position of source = 1.7m, 3m, 1m and the RT was set to T60 = 0.4. The resulting RIR can be seen in figure 3.2 with the results of the estimated RTs ($T60$ measure) at multiple third octave frequency subbands. The non linear approach was used to calculate the RTs due to it being more consistent when compared with the linear approach.

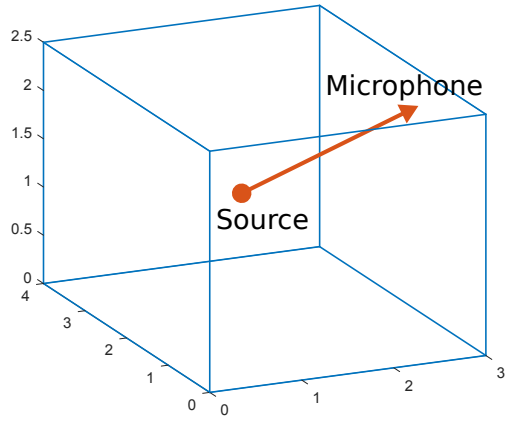


Figure 3.1: Synthetic room dimensions in metres (direct-path in orange)

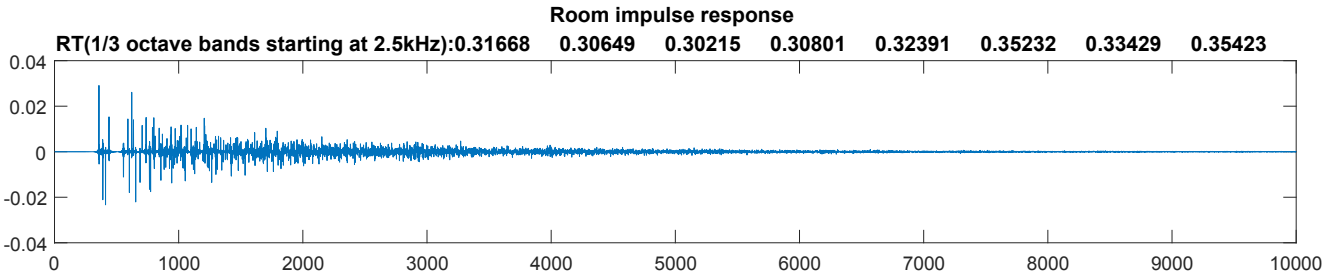


Figure 3.2: Synthetic room impulse response

Table 3.1 shows RT estimates when the RT value, in the ISM algorithm, is set to different lengths. These estimates, calculated by non linear approach, are taken to be the base case as the RT value set in the ISM is only used as a guideline.

	RT at centre octave band frequencies(s)							
ISM RT(s)	2500Hz	3150Hz	4000Hz	5000Hz	6300Hz	8000Hz	10000Hz	12500Hz
0.1	0.08	0.08	0.07	0.08	0.07	0.08	0.08	0.08
0.4	0.32	0.31	0.30	0.31	0.32	0.35	0.33	0.35
0.9	0.75	0.73	0.75	0.72	0.76	0.77	0.77	0.77

Table 3.1: RT estimates for different RT values set in ISM algorithm

3.2 Room impulse response estimation

The synthetic RIR can be used to investigate estimation methods based on speech signals using equation 2.6 shown ealier. In this case, the time index is denoted by n taking values from 1 to N, x is a realistic synthetic recording which is calculated by convolving h, the synthetic RIR, with s, the clean speech signal, with contaminating white Gaussian noise y being added to this result to make it more authentic. Table 3.2 shows the effects of noise on the three known estimators that have been implemented when the speech length used was 36.3s and the filter length was set to 1s. The time-domain estimator performs poorly in noisy conditions which is to be expected and, therefore, this method will not be examined further in this report. In the next section the two frequency-domain approaches described in section 2.4.2 are explored with $x(n)$ and $s(n)$ from equation 2.6 given as inputs. The FFT shift was set to 512 for the remainder of this report.

SSNR	TD NPM	FD NPM	NS NPM
-2.254	-0.663	-1.271	-2.411
2.747	-1.916	-2.984	-4.533
7.748	-4.400	-6.308	-7.455
12.750	-8.196	-10.236	-9.903
17.748	-12.746	-14.000	-11.299

Table 3.2: Effects of noise on the time-domain(TD), frequency-domain(FD) and nonstationarity(NS) estimators

3.2.1 Effects of background noise

The first metric that needs to be considered is the amount of noise that is added to contaminate the synthetic record signal x. Tables 3.3 & 3.4 and figures 3.3 & 3.4 show the effect of noise on the estimates given by the two frequency approaches when the speech length used was 36.3s and the filter length was set to 1s.

Conventional frequency-domain estimate									
SSNR	NPM	RT Error (1/3 octave bands starting at 2.5kHz)							
-32.246	-0.003	cannot be calculated							
-27.250	-0.003	cannot be calculated							
-22.245	-0.014	cannot be calculated							
-17.250	-0.060	cannot be calculated							
-12.250	-0.105	0.020	0.027	0.044	0.030	0.034	0.005	0.054	0.121
-7.254	-0.464	0.011	0.012	0.032	0.004	0.016	0.012	0.014	0.089
-2.256	-1.271	0.002	0.015	0.022	0.012	0.013	0.008	0.003	0.006
2.748	-2.984	0.010	0.017	0.028	0.020	0.016	0.040	0.017	0.009
7.742	-6.308	0.019	0.013	0.020	0.013	0.014	0.022	0.018	0.010
12.749	-10.236	0.014	0.013	0.020	0.016	0.016	0.032	0.018	0.041
17.754	-14.000	0.017	0.016	0.017	0.018	0.018	0.036	0.018	0.022
22.756	-16.785	0.016	0.013	0.013	0.016	0.018	0.032	0.018	0.016
27.759	-18.486	0.016	0.013	0.015	0.016	0.018	0.033	0.016	0.020
32.741	-19.086	0.015	0.012	0.015	0.016	0.019	0.031	0.016	0.021
37.739	-19.316	0.015	0.012	0.015	0.016	0.019	0.031	0.016	0.022
42.753	-19.437	0.015	0.012	0.015	0.016	0.019	0.031	0.016	0.021
47.738	-19.449	0.015	0.012	0.015	0.016	0.019	0.031	0.016	0.021

Table 3.3: Effects of background noise on conventional FD estimate

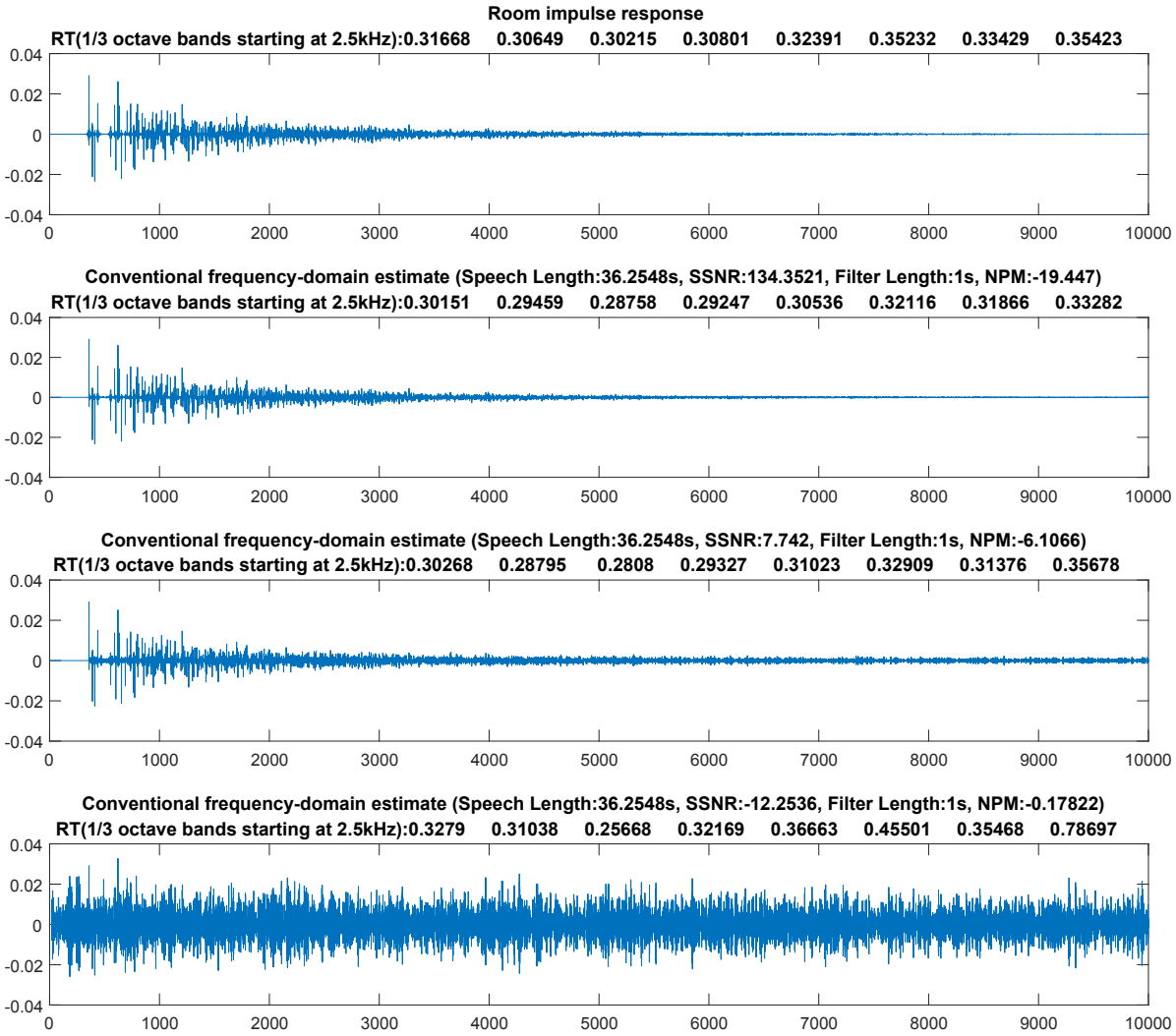


Figure 3.3: A simple conventional frequency-domain method

Frequency-domain estimator using nonstationarity									
SSNR	**NPM**	**RT Error (1/3 octave bands starting at 2.5kHz)**							
-32.246	-0.016	0.126	0.064	0.073	0.049	0.065	0.039	0.058	0.032
-27.250	-0.018	0.093	0.058	0.053	0.080	0.056	0.007	0.054	0.046
-22.245	-0.039	0.060	0.028	0.024	0.040	0.031	0.033	0.023	0.030
-17.250	-0.084	0.034	0.003	0.019	0.003	0.001	0.006	0.015	0.024
-12.250	-0.224	0.009	0.029	0.055	0.038	0.037	0.034	0.015	0.010
-7.254	-0.830	0.049	0.065	0.084	0.063	0.071	0.066	0.050	0.004
-2.256	-2.411	0.063	0.076	0.093	0.084	0.087	0.095	0.086	0.041
2.748	-4.533	0.066	0.082	0.100	0.090	0.097	0.109	0.103	0.068
7.742	-7.455	0.070	0.085	0.102	0.091	0.098	0.109	0.110	0.098
12.749	-9.903	0.069	0.086	0.104	0.092	0.099	0.110	0.112	0.124
17.754	-11.299	0.070	0.086	0.104	0.093	0.100	0.111	0.113	0.116
22.756	-11.709	0.069	0.086	0.104	0.093	0.099	0.110	0.114	0.117
27.759	-11.910	0.070	0.086	0.104	0.092	0.100	0.112	0.114	0.117
32.741	-11.956	0.070	0.086	0.104	0.092	0.100	0.111	0.114	0.118
37.739	-11.983	0.070	0.086	0.104	0.092	0.100	0.111	0.114	0.118
42.753	-11.994	0.070	0.086	0.104	0.092	0.100	0.111	0.114	0.118
47.738	-11.995	0.070	0.086	0.104	0.092	0.100	0.111	0.114	0.118

Table 3.4: Effects of background noise on FD estimator using nonstationarity

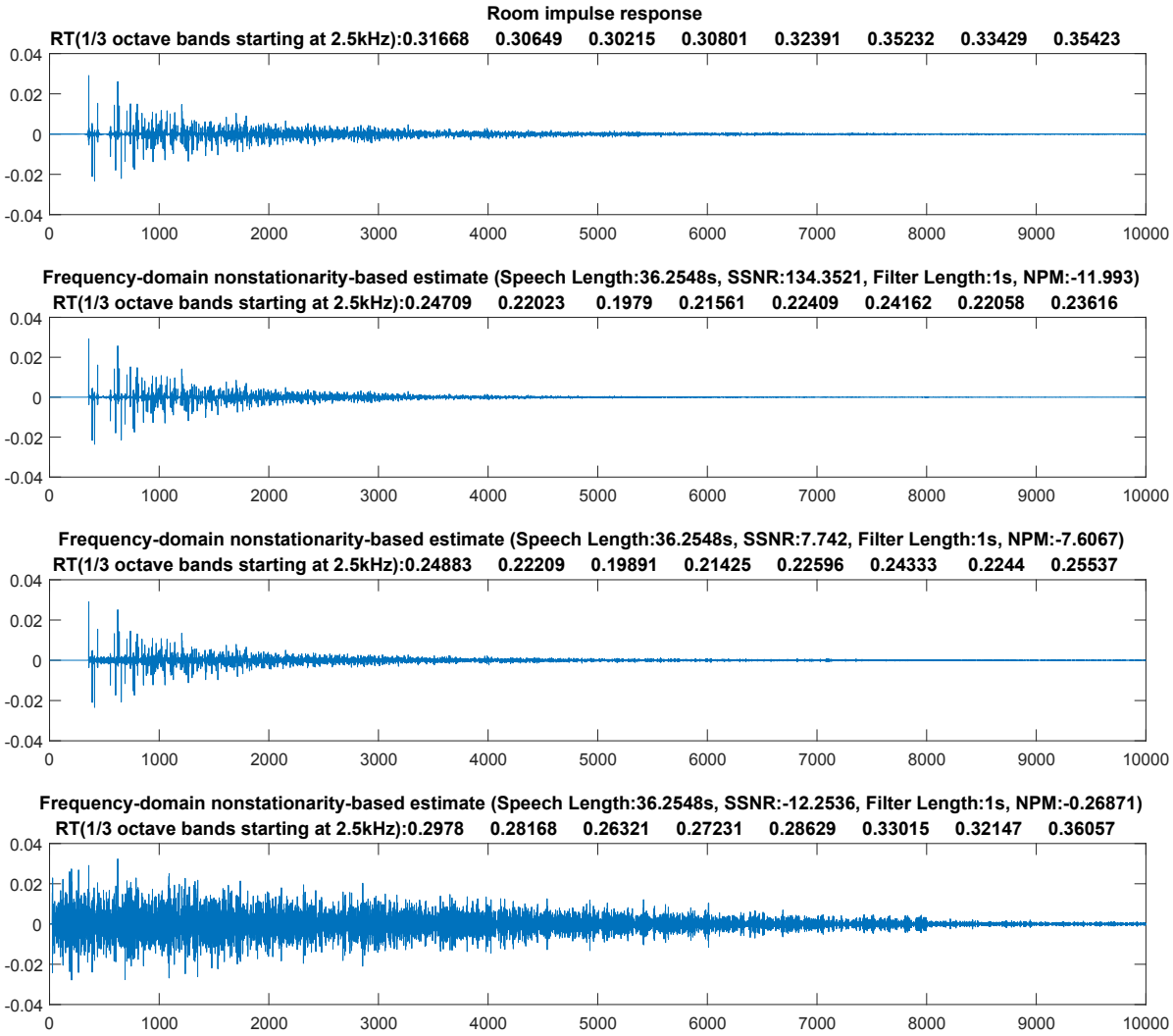


Figure 3.4: Frequency-domain estimator using nonstationarity

To evaluate the performance of these two algorithms, two evaluation metrics were calculated. These two metrics were the NPM measure and the absolute error in the estimated RT at different octave frequency bands given by the non-linear method. The amount of noise added to the synthetic recording was measured by the segmental signal-to-noise ratio (SSNR) metric. The SSNR is similar to the SNR except it works on short segments of the signal calculating the SNR value for each segment. The SSNR is then given by the average SNR of these smaller sections (equation 3.1).

$$\text{SSNR} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{P_{\text{signal}}^{(m)}}{P_{\text{noise}}^{(m)}} \tag{3.1}$$

When there is very low noise (SSNR=135) then the conventional method performs better, however, when noise is present the nonstationarity estimator produces a better result when comparing performance using the NPM measure. This is, unfortunately, overshadowed by the fact that the conventional method is still able to produce a better estimate of the RT.

3.2.2 Effects of filter length (FFT resolution)

The length of the filter that is selected in the estimation algorithms is crucial to analyse due to its trade-offs. The longer the filter is, the better the FFT resolution, however, the length of the filter also increases the computation required to calculate the estimate. Figures 3.5 & 3.6 and tables 3.5 & 3.6 show how the two FD approaches are affected by reducing the filter length when the speech length used is 36.2548s and the SSNR is 8.

Conventional frequency-domain estimate									
Filter Length(s)	**NPM**	**RT Error (1/3 octave bands starting at 2.5kHz)**							
0.500	-7.361	0.026	0.031	0.044	0.034	0.037	0.055	0.039	0.043
0.750	-6.834	0.024	0.021	0.027	0.023	0.020	0.045	0.029	0.015
1.000	-6.254	0.009	0.018	0.020	0.015	0.017	0.015	0.013	0.040
1.250	-5.873	0.011	0.005	0.021	0.012	0.015	0.027	0.007	0.035
1.500	-5.690	0.007	0.016	0.013	0.015	0.008	0.007	0.008	0.005

Table 3.5: Effects of filter length on conventional FD estimate

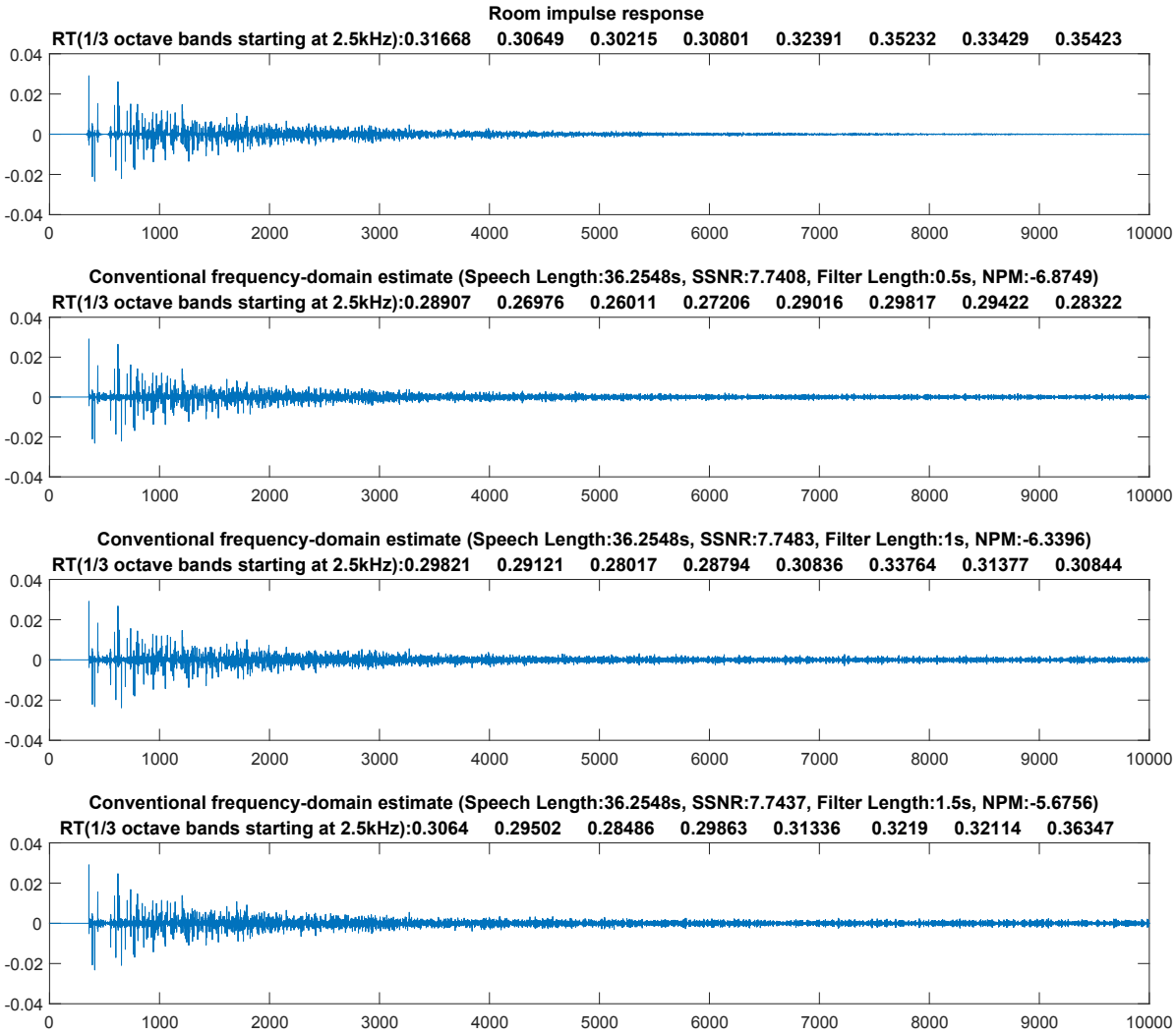


Figure 3.5: A simple conventional frequency-domain method

The results from figure 3.5 show that if the length of the filter used in the simple FD method is increased, the estimator performs better. The point of interest, however, is that the estimate of the RT given by the estimator, when the filter length is set to 500ms, is still fairly accurate.

Frequency-domain estimator using nonstationarity									
Filter Length(s)	**NPM**	**RT Error (1/3 octave bands starting at 2.5kHz)**							
0.500	-6.227	0.147	0.151	0.169	0.169	0.173	0.184	0.183	0.203
0.750	-7.608	0.103	0.112	0.137	0.118	0.127	0.132	0.138	0.135
1.000	-7.714	0.068	0.085	0.102	0.091	0.099	0.110	0.110	0.111
1.250	-7.324	0.044	0.040	0.073	0.070	0.079	0.086	0.081	0.068
1.500	-6.835	0.031	0.014	0.049	0.045	0.055	0.068	0.053	0.007

Table 3.6: Effects of filter length on FD estimator using nonstationarity

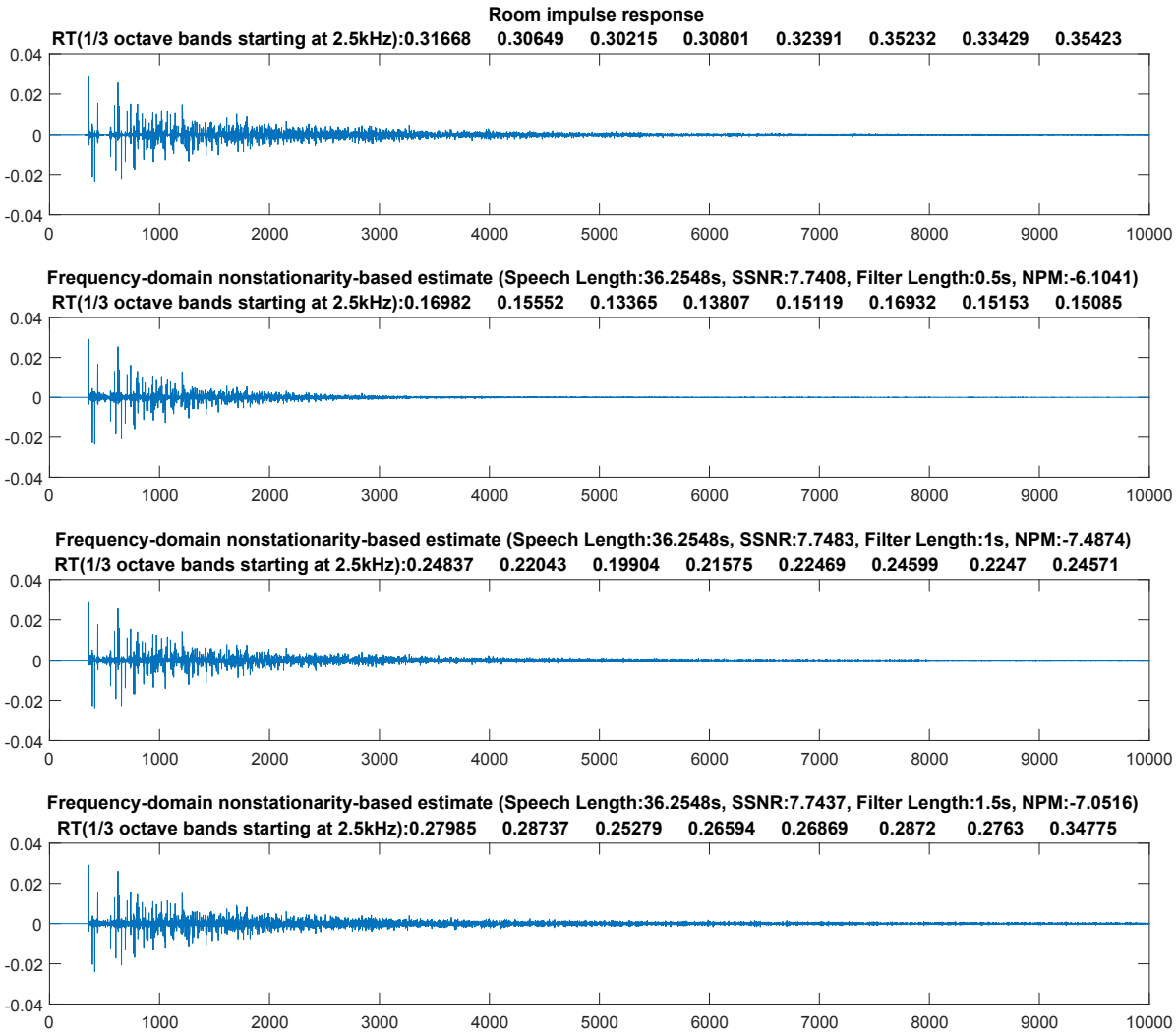


Figure 3.6: Frequency-domain estimator using nonstationarity

It can again be observed that the nonstationarity method performs better when comparing with the NPM measure, however, the RT is better estimated by the simple conventional method. This is due to the fact that the NPM measure is biased towards larger coefficients and, therefore, is not really affected by the reverberation tail whereas the RT very much depends on the reverberation tail. The simple conventional method can perform well under noisy conditions by virtue of the fact that, even though the noise floor is higher, the calculation of the RT ($T60$) measure depends on the slope of the energy decay curve down to the noise floor. The effects of the filter length will be explored further when looking at classification of rooms using real data. If a short length can be exploited then this will massively speed up the time taken to classify each room.

3.2.3 Effects of speech length

This project is ultimately trying to prove whether a robot could identify a room using only sound. It is, therefore, important to compare how the length of the speech signal used effects the RIR estimation. The best case scenario is one where the output speech signal is very short meaning, for example, the loudspeaker would have to output very little in order to determine the RIR. Figures 3.7 & 3.8 and tables 3.7 & 3.8 show how the two FD approaches are effected by reducing the filter speech when the filter length is set to 1s.

Conventional frequency-domain estimate										
Speech Length(s)	**SSNR**	**NPM**	**RT Error (1/3 octave bands starting at 2.5kHz)**							
2.023	13.620	-0.077	0.020	0.005	0.038	0.600	0.244	0.338	1.274	1.429
3.372	9.553	-1.063	0.008	0.018	0.064	0.006	0.006	0.011	0.005	0.062
11.145	8.986	-1.981	0.012	0.016	0.026	0.010	0.003	0.008	0.020	0.057
12.440	5.234	-1.427	0.001	0.002	0.028	0.023	0.007	0.005	0.043	0.044
36.255	7.746	-6.381	0.018	0.015	0.024	0.017	0.016	0.012	0.020	0.013

Table 3.7: Effects of speech length on conventional FD estimate

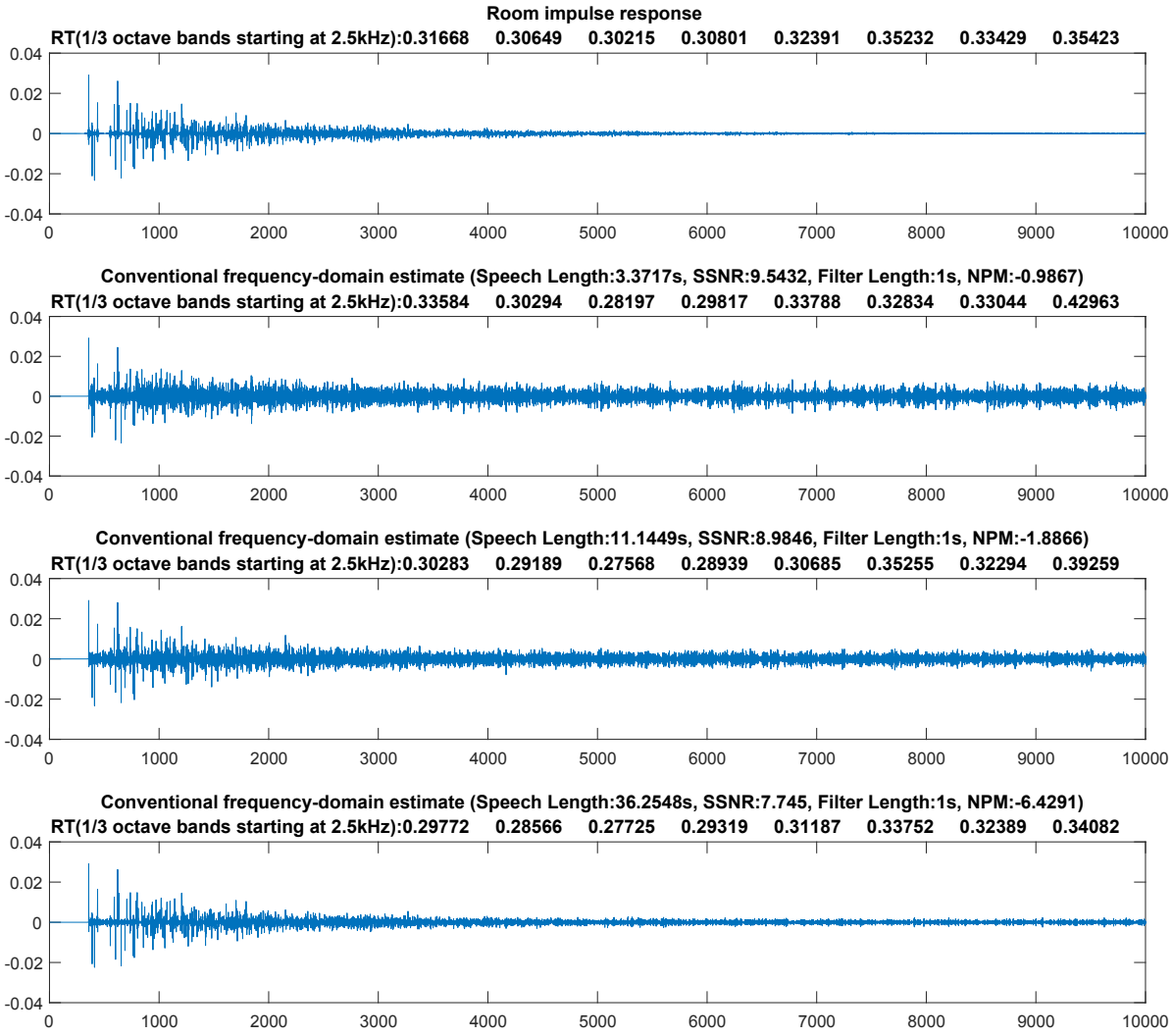


Figure 3.7: A simple conventional frequency-domain method

Frequency-domain estimator using nonstationarity										
Speech Length(s)	**SSNR**	**NPM**	**RT Error (1/3 octave bands starting at 2.5kHz)**							
2.023	13.620	-0.500	0.046	0.194	0.147	0.315	0.237	0.273	0.385	0.297
3.372	9.553	-3.232	0.018	0.014	0.026	0.016	0.001	0.004	0.001	0.045
11.145	8.986	-3.679	0.071	0.072	0.059	0.039	0.059	0.041	0.009	0.024
12.440	5.234	-3.099	0.039	0.006	0.029	0.035	0.045	0.060	0.049	0.003
36.255	7.746	-7.708	0.070	0.086	0.105	0.091	0.099	0.108	0.111	0.100

Table 3.8: Effects of speech length on FD estimator using nonstationarity

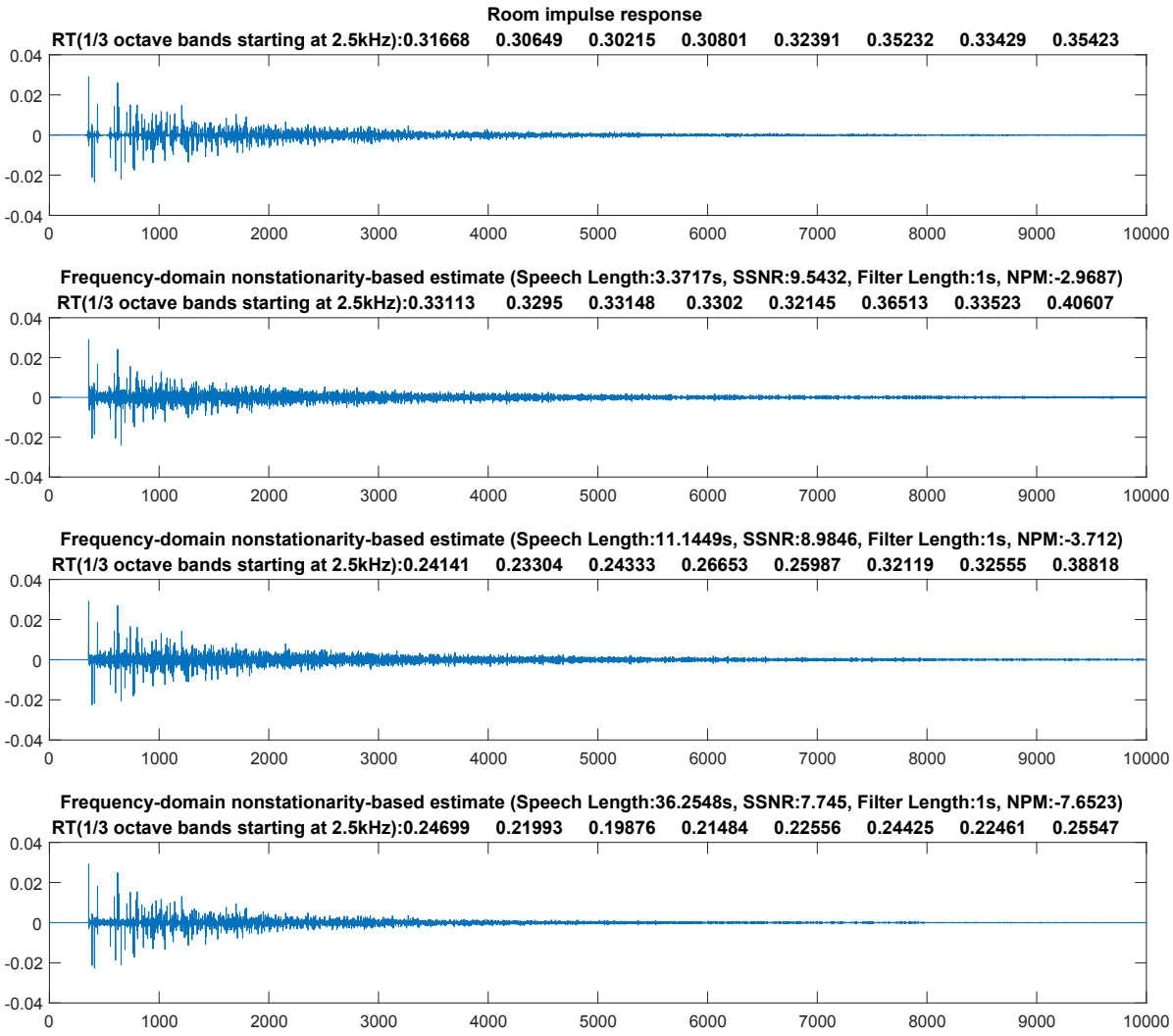


Figure 3.8: Frequency-domain estimator using nonstationarity

This is the main area where the nonstationarity frequency-domain estimator is superior to the simple conventional method in terms of the NPM measure. When a short speech signal is implemented, less frequency components of the room will be excited meaning the estimate is likely to be less accurate as noise contamination has a greater effect.

3.2.4 Conclusion

The conventional FD method performed better when estimating the RT in almost every scenario when using synthetic data. For this reason the conventional FD approach is the one that will be explored further in this project on real data.

3.3 ACE corpus data

The ACE corpus contained 14 RIRs taken in 7 rooms in two positions. These RIRs were important to check the validity of the current algorithms on real data where many factors are not controllable. The ACE corpus RIRs were deployed in a similar way to the synthetic RIRs. First the RT of each RIR was calculated directly. The results obtained can be seen in figure 3.9. These results could then be compared with the estimates calculated using speech signals with and without the presence noise.

The 2 positions for each room where chosen for a particular reason; in one position the microphone was set up a short distance away from the speaker and in the other case, the microphone was set up a long distance away from the speaker. The expected result is one where the RTs for both positions is similar. In theory, they would be the same due to the fact that the RT of a given room should be the same at any position, however, errors are always introduced when taking measurements.

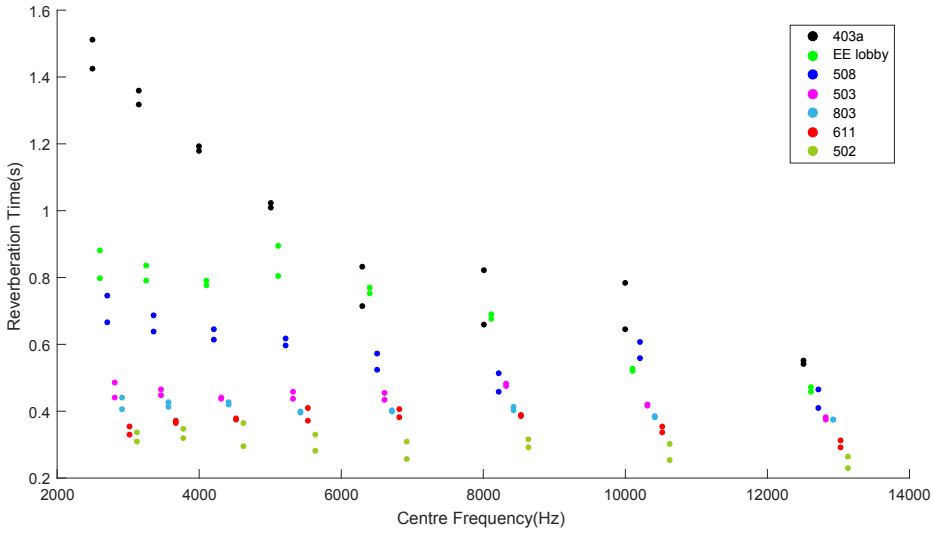


Figure 3.9: Directly calculated RTs from ACE corpus RIRs

Figure 3.9 shows the directly calculated RTs at third octave frequency bands for each room in 2 positions. This result is pleasing as it matches the expected result; the RTs of the two positions are relatively similar. These RT estimates can now be used as the base case when evaluating the estimates obtained from speech signals.

3.3.1 Room impulse response estimation using speech signals

The results of using the conventional FD estimator with all long speech samples (data set 10) can be seen in figure 3.10. In this setup no noise was added, the result being that each speech estimate is nearly identical to the directly calculated RT value.

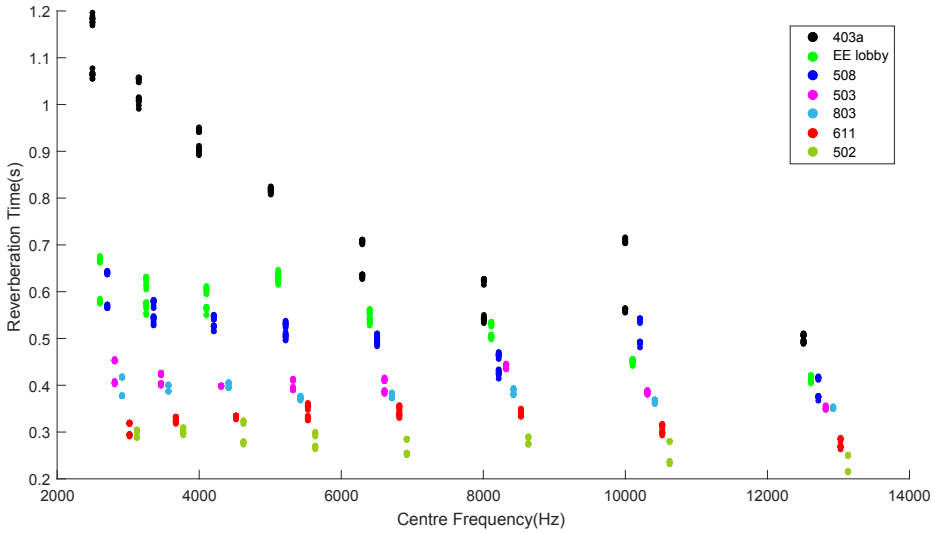


Figure 3.10: RTs calculated using speech by the conventional FD method

Classification is possible with the result seen in figure 3.10. A confusion plot can be seen in figure 3.11 which is derived from the output of a naive Bayes classifier when leave-one-out cross-validation is performed. The confusion plot shows that 100% classification can be achieved which is to be expected if all the rooms have varying RTs and there is no noise contamination.

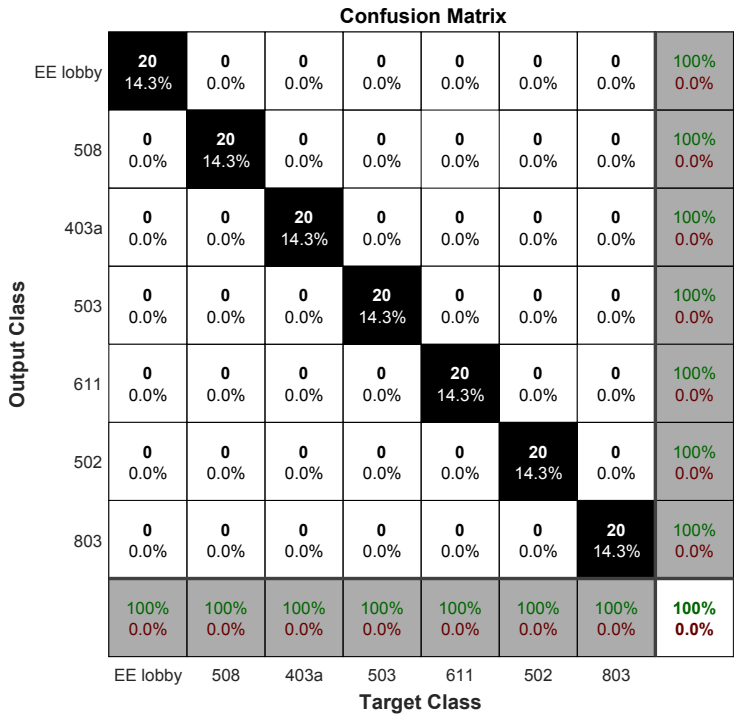


Figure 3.11: Confusion plot of classification with no added noise

3.3.2 Synthetic background noise

The algorithm has been shown to work using real room impulse responses. It is now prudent to show the effects of noise before looking at real recorded speech signals. This is on the grounds that in any recording, background noise is always present, therefore, to show that the algorithm is practical, it needs to be seen to be working within a noisy environment.

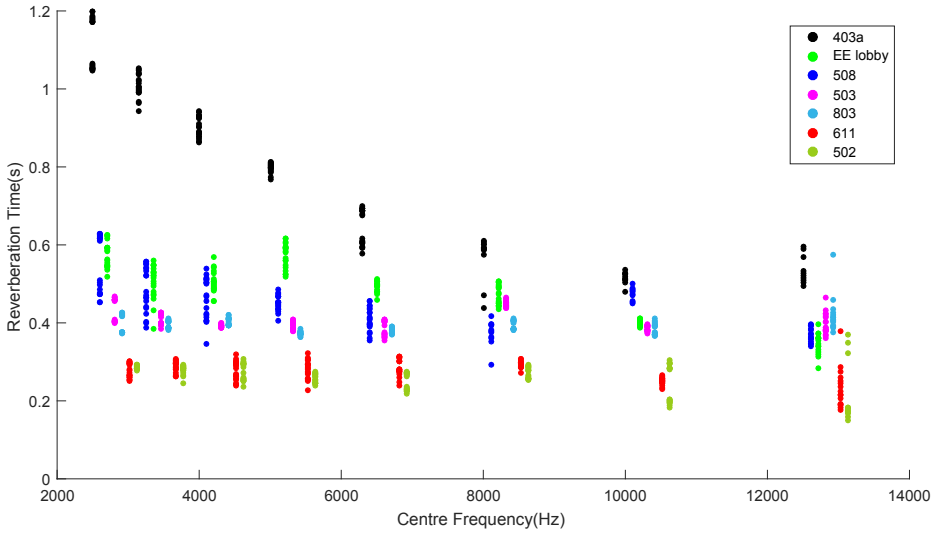


Figure 3.12: RTs calculated using speech by the conventional FD method

The results of the RTs calculated with noisy recordings (SSNR: 3.6) can been seen in figure 3.12. It is clear to see the effect of background noise in the results as there is a noticeable increase in the variance of the RTs calculated in each position .

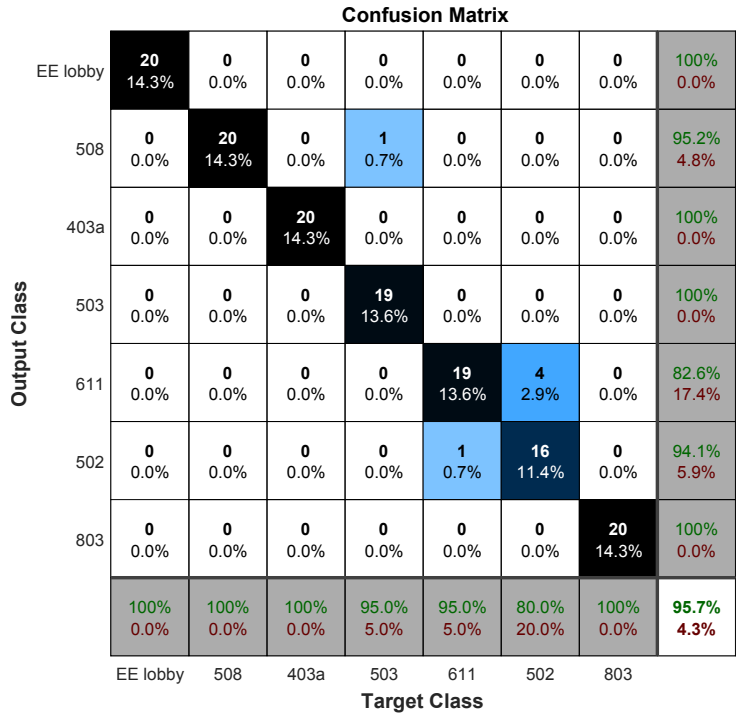
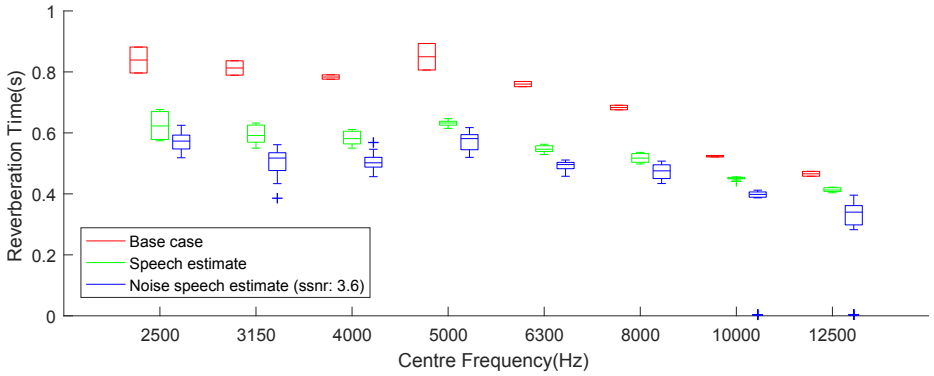


Figure 3.13: Confusion plot of classification with added noise (SSNR: 3.6)

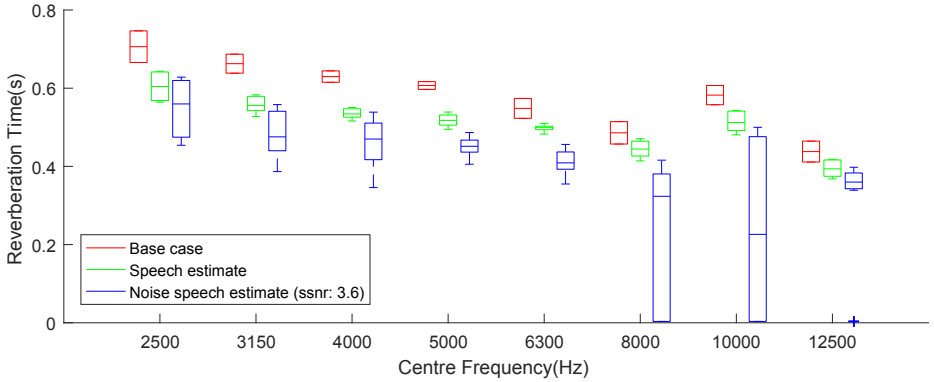
The classification achieved by a naive Bayes classifier when leave-one-out cross-validation is adopted is also reduced from 100% of data points correctly identified to only 95.7% correctly identified. This conclusion can be seen in the confusion plot in figure 3.13. This is still a very high classification rate in the presence of noise and demonstrates the robustness of the method that has been selected.

Distribution of T60 estimates

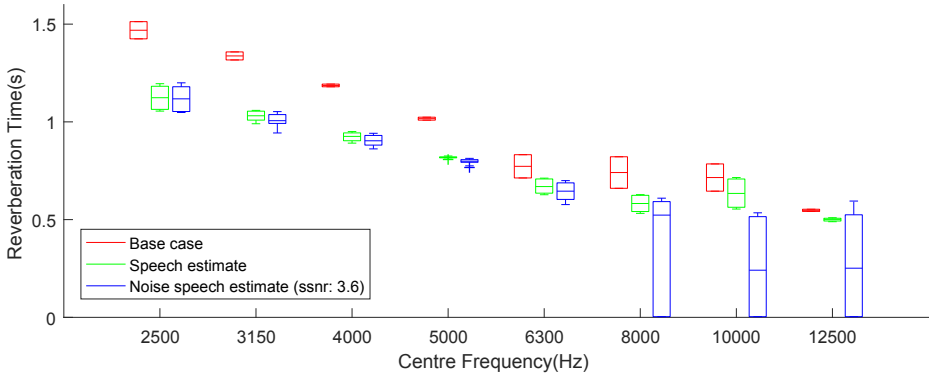
Figure 3.14 shows the distribution of T60 estimates at each third octave band for each room. The $T60$ estimates obtained using speech signals are mostly underestimates when compared with the base case, however, the process of adding white Gaussian noise mainly impacts the variance of the result.



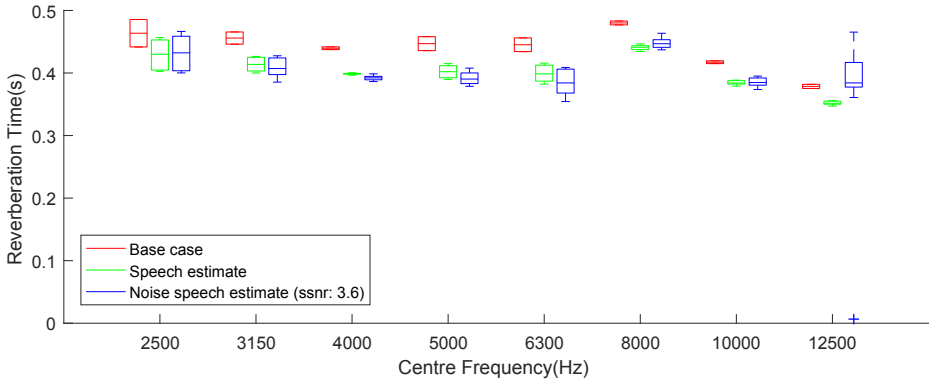
(a) Distribution of T60 estimates for the EE lobby for each third octave band



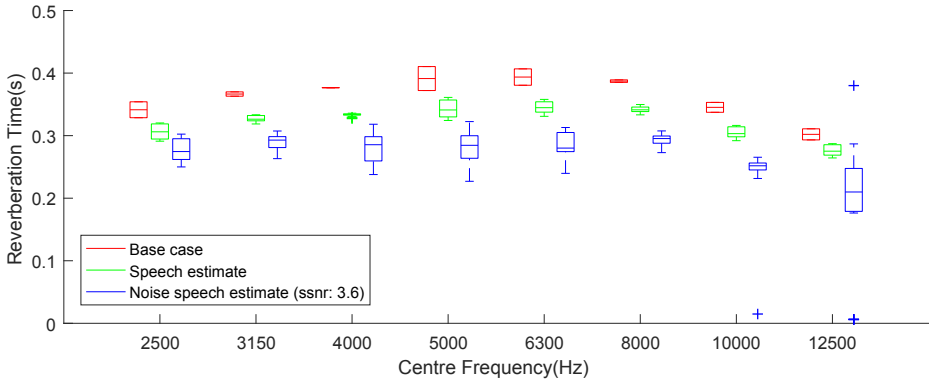
(b) Distribution of T60 estimates for room 508 for each third octave band



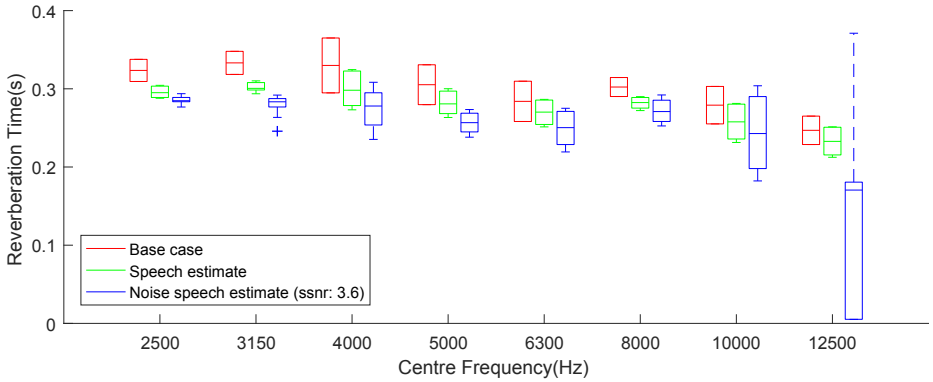
(c) Distribution of T60 estimates for room 403a for each third octave band



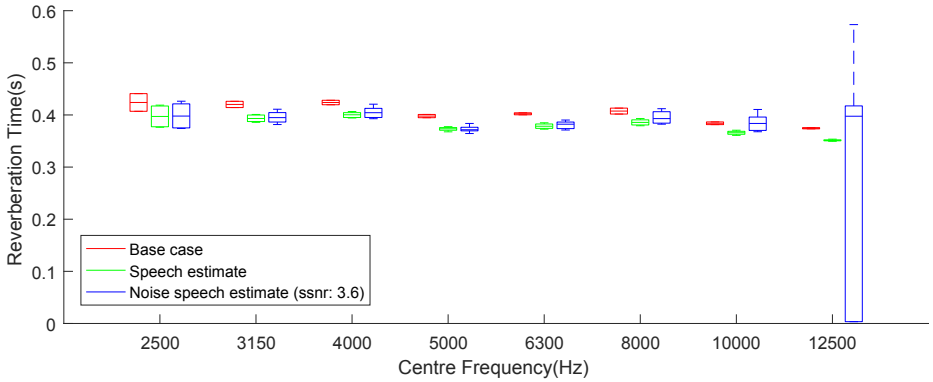
(d) Distribution of T60 estimates for room 503 for each third octave band



(e) Distribution of T60 estimates for room 611 for each third octave band



(f) Distribution of T60 estimates for room 502 for each third octave band



(g) Distribution of T60 estimates for room 803 for each third octave band

Figure 3.14: Distribution of T60 estimates for each third octave band

3.4 Experimental data

To prove that this project has achieved its objectives, the algorithms need to be seen to be working with real data. For this reason data from 5 rooms was collected in 4 different positions. A basic blueprint of the setup can be seen in figure 3.15 which shows the positions of the microphone and speaker when the data was collected. See appendix C.1 for images of the setup in each room for each position.

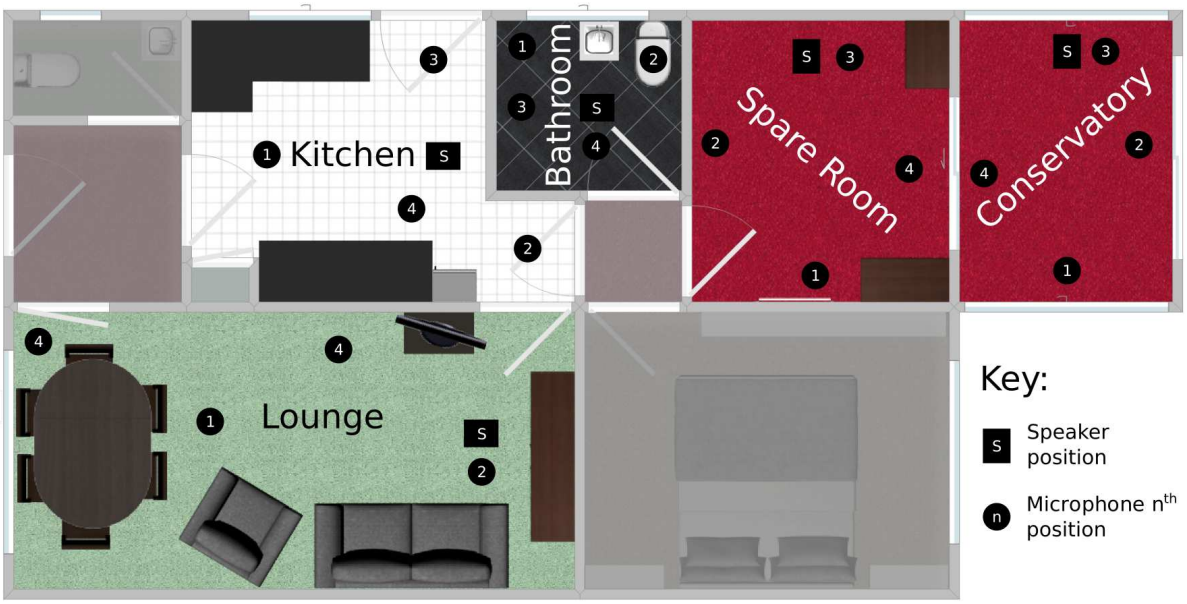


Figure 3.15: Floor plan

3.4.1 Exponential sine-sweep method

To obtain a near perfect estimation of each RIR for all the positions, the exponential sine-sweep method was deployed. The results of this approach can be seen in full in appendix C.3. These accurate estimations of the RIRs will be useful when evaluating the estimates obtained using speech signals.

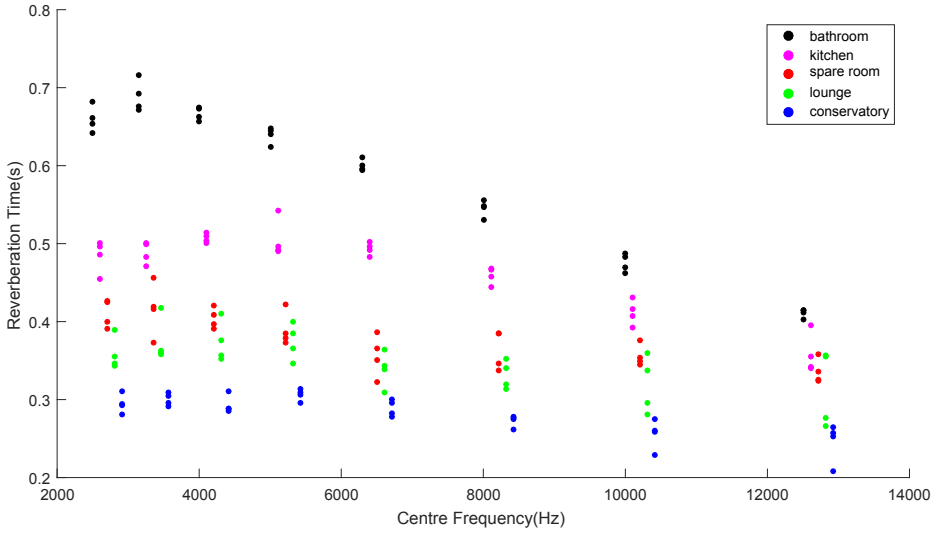


Figure 3.16: Directly calculated RTs from RIRs generated by sine-sweep method

The results of the calculated RTs, at different octave frequency bands, acquired from the RIRs gathered from the sine-sweep method are displayed in figure 3.16. These results show that classification will be possible, however, the results also show that the lounge and spare room have very similar RTs. This similarity, therefore, will make it difficult to distinguish between these two rooms, highlighting one of the major disadvantages of using the reverberation time for identification; which is a lot of rooms have similar reverberation times even when measuring the $T60$ metric at octave frequency bands.

3.4.2 Room impulse response estimation using speech signals

This section will explore the implementation and results of the estimation approaches when applied to experimental data. It will ultimately prove if identification of rooms is possible in the particular house where the data was collected.

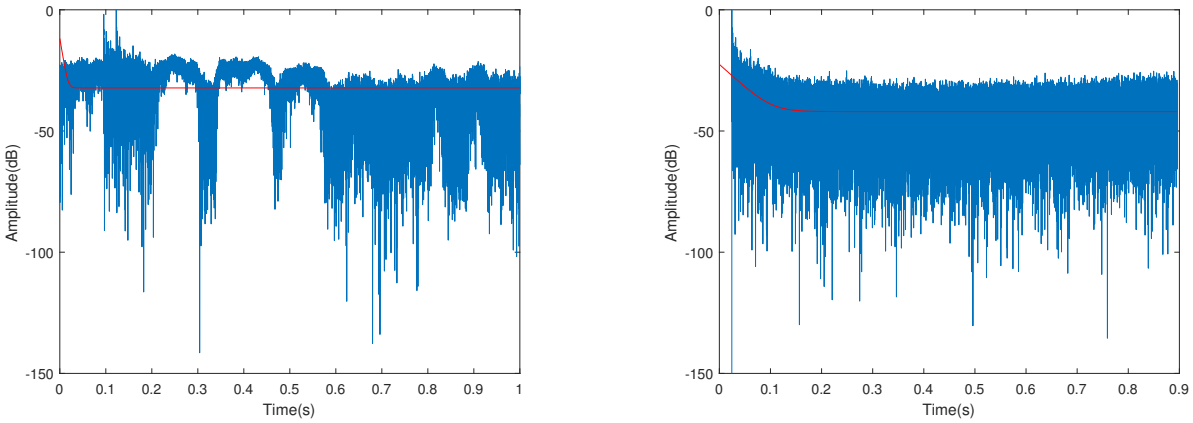
The testing of real data made apparent some simple improvements, to the algorithms implementations, needed to be addressed.

Reducing the sampling frequency

First, that the sampling frequency was set to 48000Hz which was not required and increased the computation time. This was resolved by having the clean and the recorded signal decimated by 3, to a new sampling frequency of 16000Hz, which had no noticeable effect on the estimates. It was implemented simply using the MATLAB decimate function, that includes lowpass filtering the input to guard against aliasing effects after downsampling.

Process cropping output from estimation methods

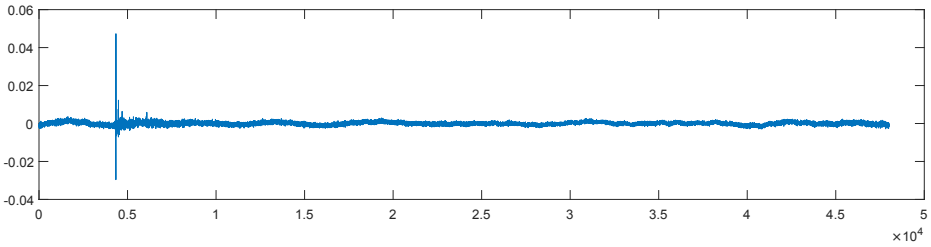
Second, the adjustment to remove unwanted frequency components in the RIR estimate. The output from the conventional FD estimator, which can be seen in figure 3.18a, contains corrupting low frequency components. This artefact, due to the limited amounts of speech at these lower frequencies, of the estimation method needs to be removed which is why the output had to be passed through a high pass filter (HPF).



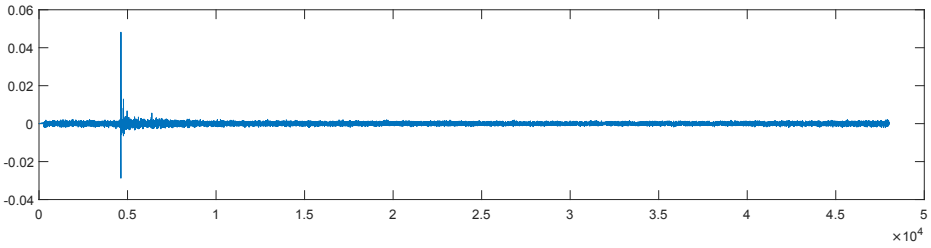
(a) $T60$ estimate of figure 3.18a (b) $T60$ estimate of figure 3.18c

Figure 3.17: Effect of cropping output of estimation methods on RT measure

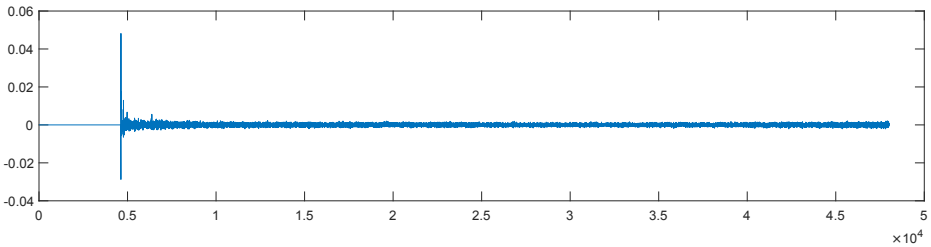
The result of this filtering is shown in figure 3.18b. Finally, the output was cropped, as seen in figure 3.18c, to improve the accuracy of the reverberation time (RT) estimation given by the non-linear approach. This can be seen in figure 3.17.



(a) Output from the conventional FD method



(b) Output after high pass filtering

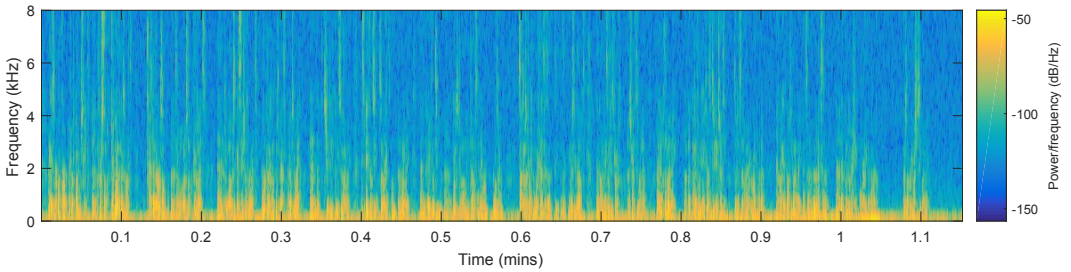


(c) Cropped output from HPF

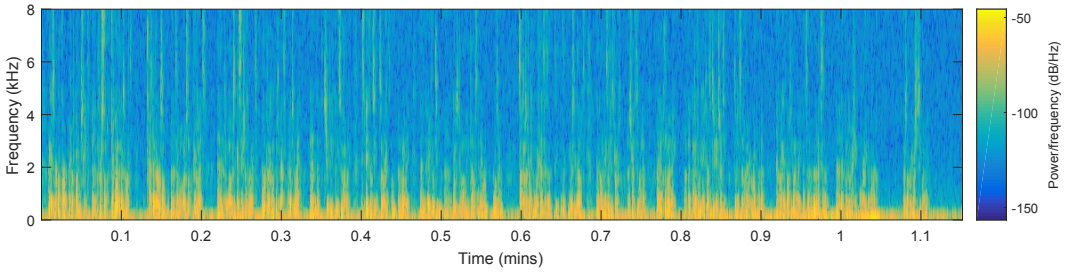
Figure 3.18: Process of cropping the output from estimation methods

Ideal situation

To first verify that the room identification approach chosen works in a real world scenario, the ideal case was tested. The ideal case is one where there is no noise added to the recorded speech signal. Figure 3.19 shows the spectrograms of the clean and recorded speech signal in a given room. The recordings were all taken in an empty house where background noise could be seen as negligible; the only noise that was subjectively heard when carrying out the data collection was external noise from passing cars and the refrigerator. In figure 3.19 the lounge recording in position 2 was used in conjunction with the long female speech data acquired from female 1 (F1 in appendix A). This configuration was chosen arbitrarily, however, to make future comparisons clear this example will be used to show the effect of different amounts of noise on a typical speech signal.



(a) Clean speech signal



(b) Recorded speech signal

Figure 3.19: Spectrograms of speech signals in an ideal situation

The reverberation time (RT) values estimated in this ideal scenario by the conventional FD method can be observed in figure 3.20. These results bear a close resemblance to the result obtained by the exponential sine-sweep method in figure 3.16 which demonstrates that speech signal can be used to accurately calculate the RIR in a practical sense. The same problems occur in differentiating between rooms that possess similar RTs, however, this is not a problem with the implementation but a problem in using the RT at octave frequency bands to identify a room.

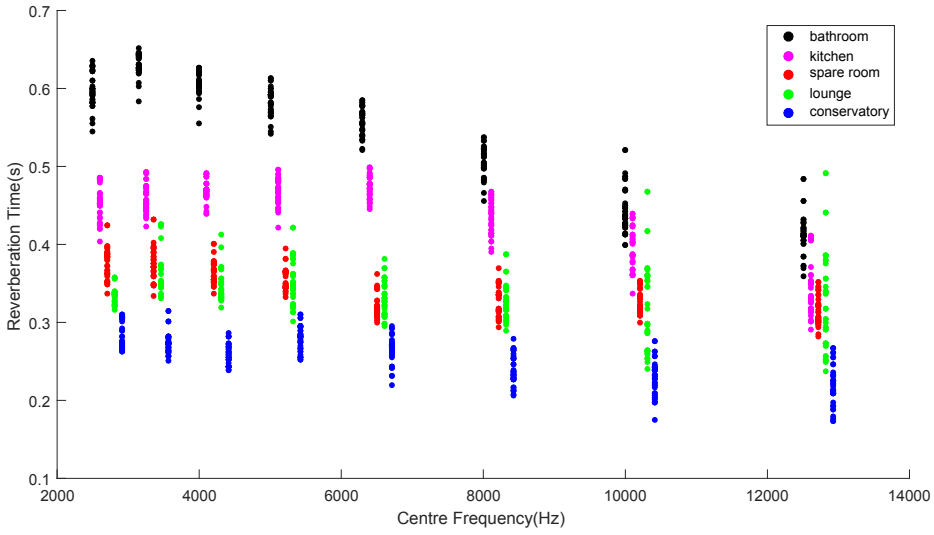


Figure 3.20: RTs calculated using speech by the conventional FD method

The classification achieved, using real recordings, by a naive Bayes classifier when leave-one-out cross-validation is used is shown in figure 3.21. The classification rate is 99.5% which means it is possible to obtain almost 100% classification by performing this method if the background noise is at a minimum level. This result, for example, shows that a robot programmed with this technique

could enter one of the five rooms in the house where the data was collected and identify its location only using sound.

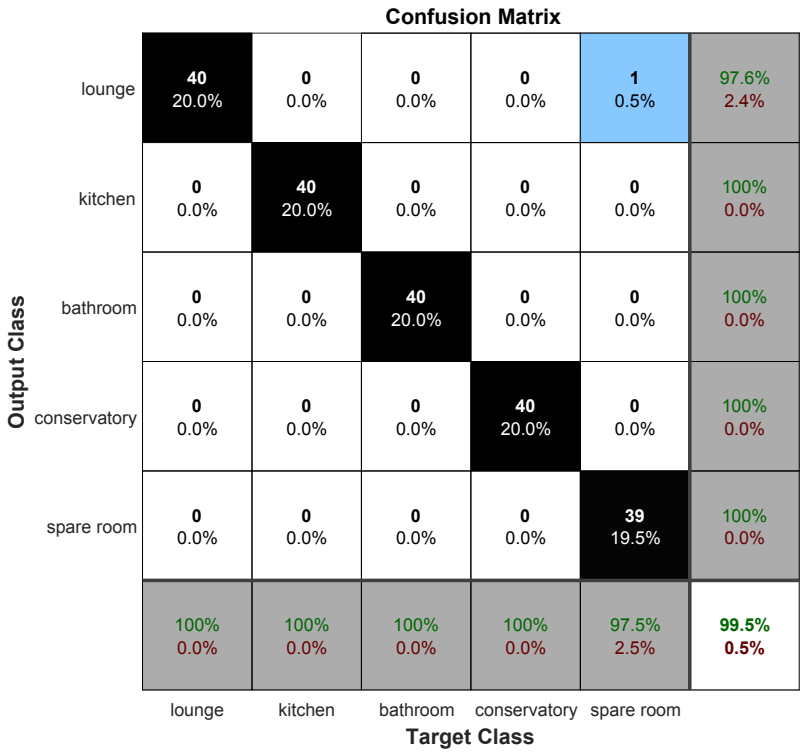


Figure 3.21: Confusion plot of classification with no added noise

3.4.3 Effects of background noise

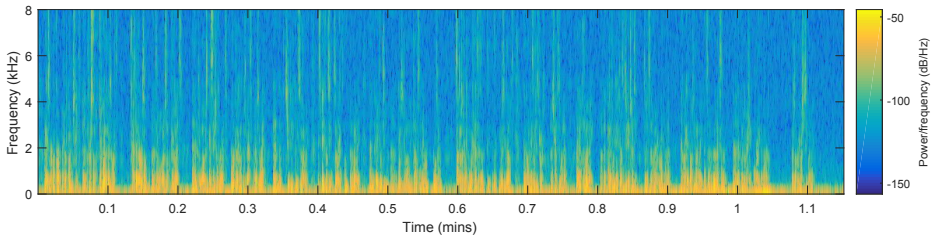
The next important part of the algorithm's assessment is to check that it performs well even after noise has been added to the speech recording. The next two parts of this section explore the effects of background noise, first when the SSNR is set to 9 and secondly when the SSNR is set to 5.

In both cases, the results of when all five rooms have been classified and the results when only four rooms have been classified are shown. This is due to the problem mentioned earlier of the lounge having a similar RT to the spare room, therefore, when noise is added, these rooms become even harder to differentiate between. By virtue of this fact the results can be greatly improved if only one of these rooms is ultimately used in the classification. This is why, in this particular section, the results that exclude the lounge are also included as a comparison.

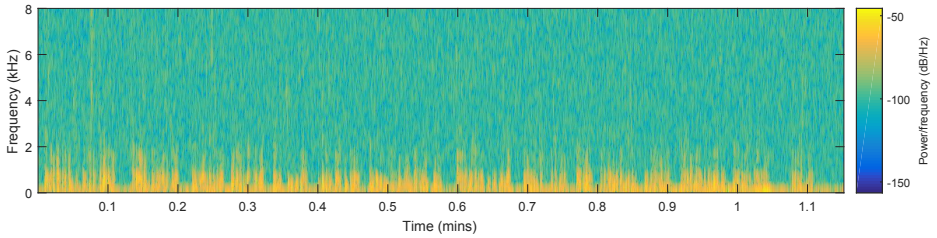
Conventional FD method (SSNR: 9)

This section shows the results when the SSNR was set to 9 and the filter length used was 1000ms. Figure 3.22 is the same as figure 3.19 except the recorded signal is now contaminated by background noise.

Figures 3.23 and 3.24 show the estimated RTs calculated in this noisy scenario. It is interesting to note that the noise clearly impacts the RTs estimated at high frequency bands more than those estimated at lower frequency bands. This increased variance in the results due to the noise inevitably makes room identification more challenging.



(a) Speech signal



(b) Speech signal with added noise (SSNR: 9)

Figure 3.22: Spectrograms of speech signals in a noisy situation (SSNR: 9)

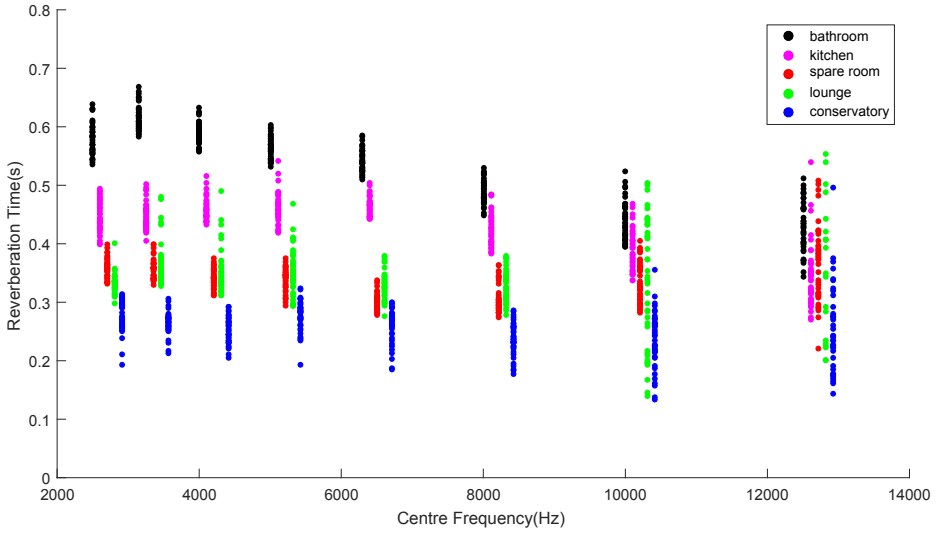


Figure 3.23: RTs calculated using speech by the conventional FD method (SSNR: 9, Rooms: 5))

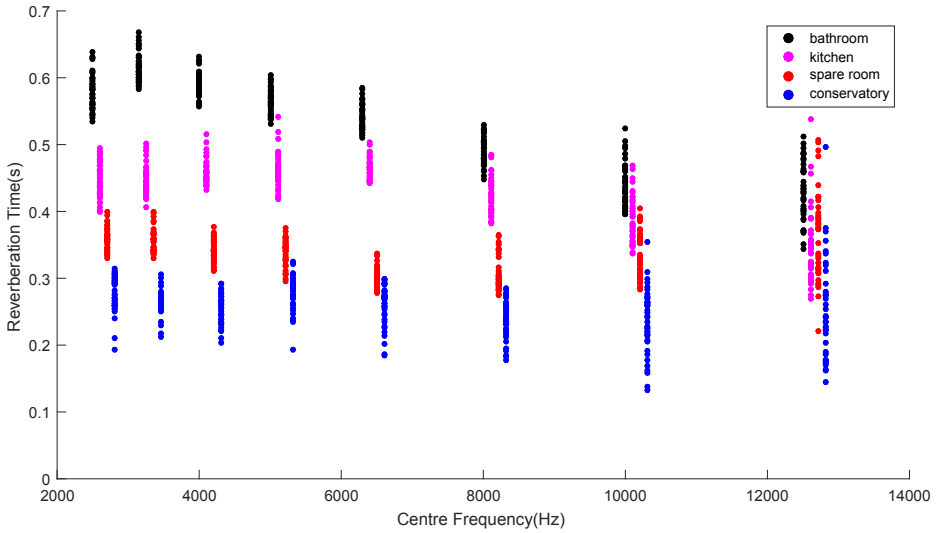
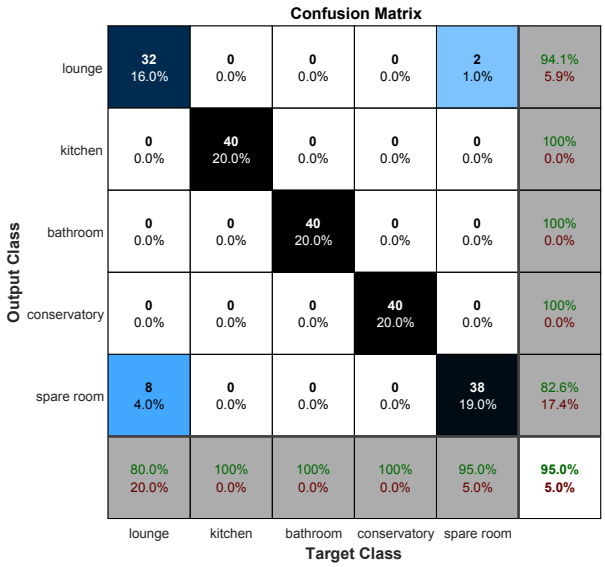
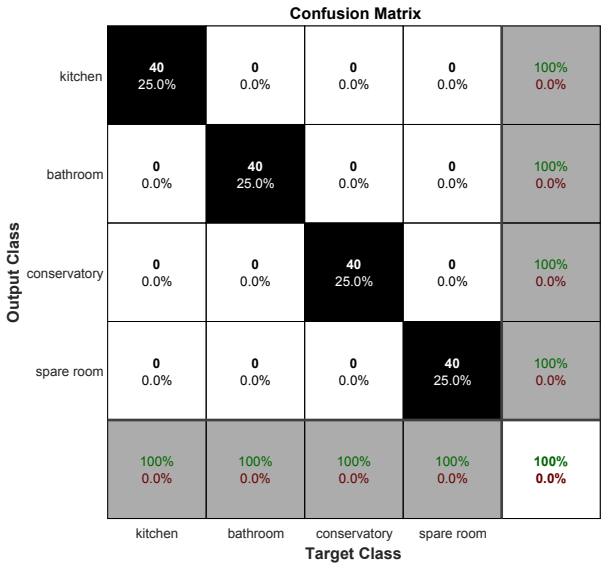


Figure 3.24: RTs calculated using speech by the conventional FD method (SSNR: 9, Rooms: 4)



(a) Confusion plot classifying 5 rooms



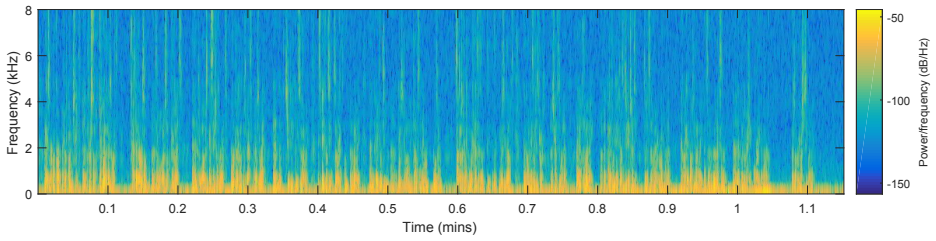
(b) Confusion plot classifying 4 rooms

Figure 3.25: Confusion plots - effects of noise (SSNR: 9.00)

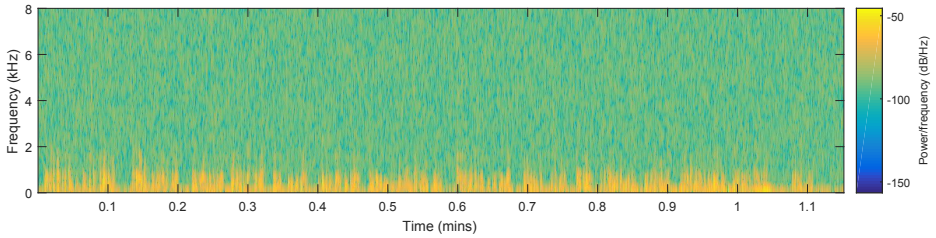
The two confusion plots both show a decrease in accuracy of classification as a result of noise, albeit the reduction is slight down to a 95% classification rate.

Conventional FD method (SSNR: 5)

This section displays the effects if the noise is increased still further to an SSNR of 5. The effect that noise has on the spectrogram of speech can be seen once again in figure 3.26.



(a) Speech signal



(b) Speech signal with added noise (SSNR: 5)

Figure 3.26: Spectrograms of speech signals in a noisy situation (SSNR: 5)

The effects of the noise on the RT estimates can been seen in figures 3.27 and 3.28.

The impact the noise has on the classification of the rooms can be seen in the confusion plots in figure 3.29 where the classification rate is down still further to 87.0% when classifying all 5 rooms. This is still not a disappointing result as a 96.3% classification rate can be observed if only 4 rooms are being classified excluding the lounge.

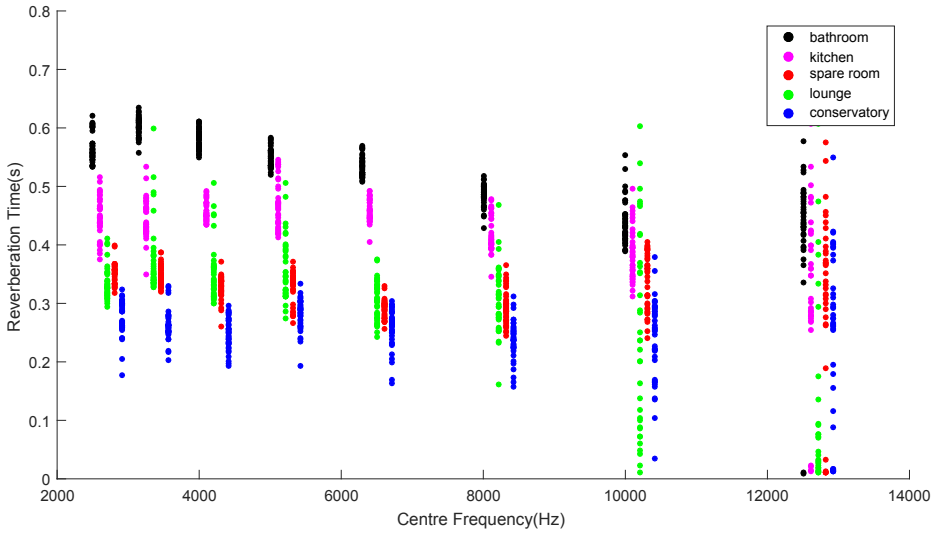


Figure 3.27: RTs calculated using speech by the conventional FD method (SSNR: 5, Rooms: 5))

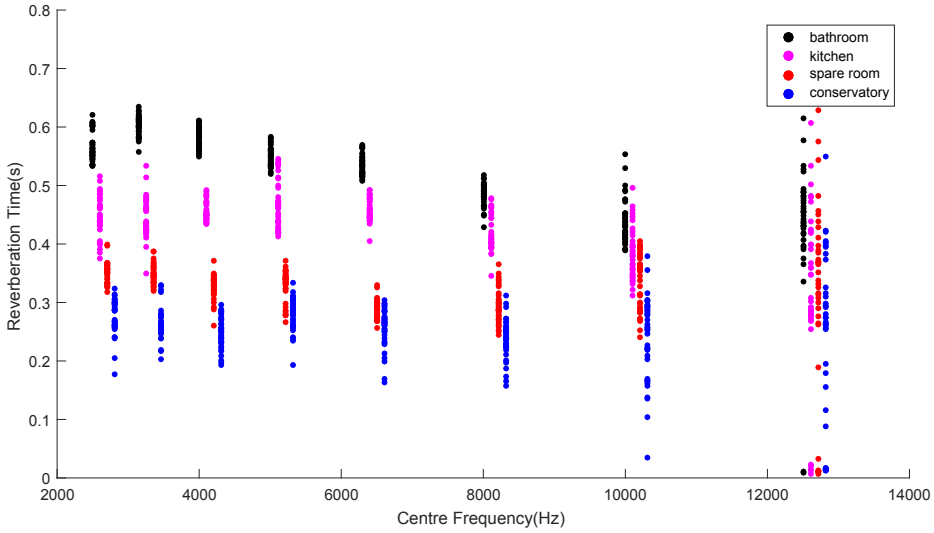
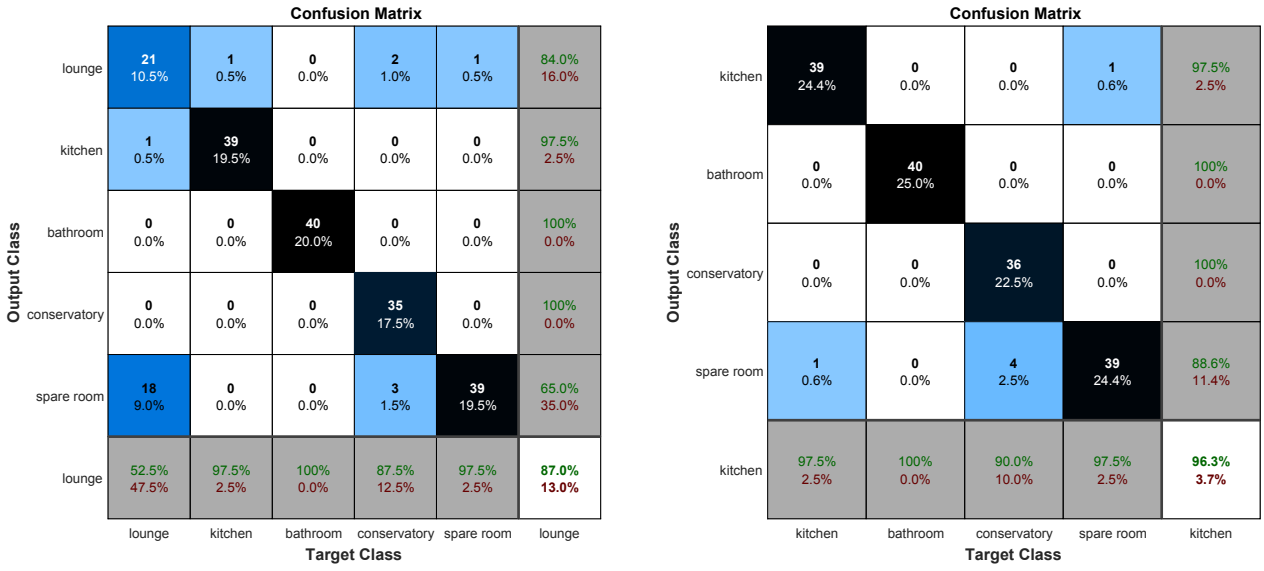


Figure 3.28: RTs calculated using speech by the conventional FD method (SSNR: 5, Rooms: 4))



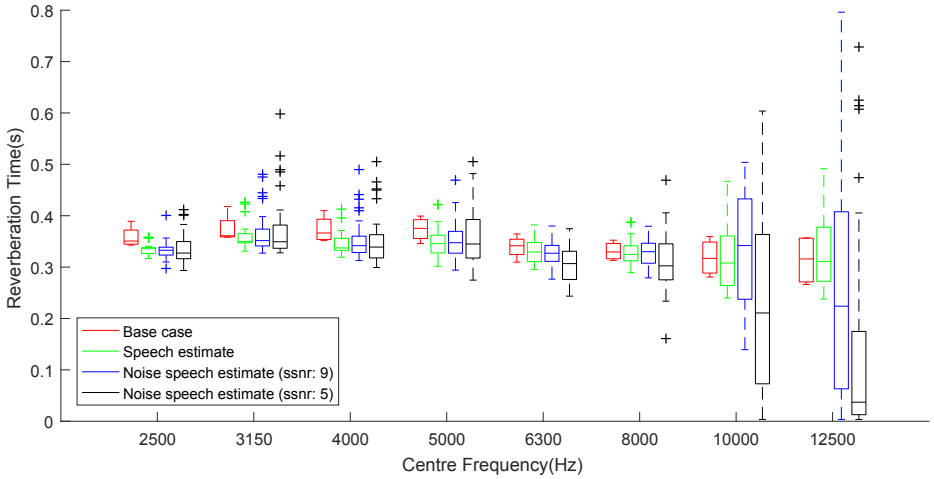
(a) Confusion plot classifying 5 rooms

(b) Confusion plot classifying 4 rooms

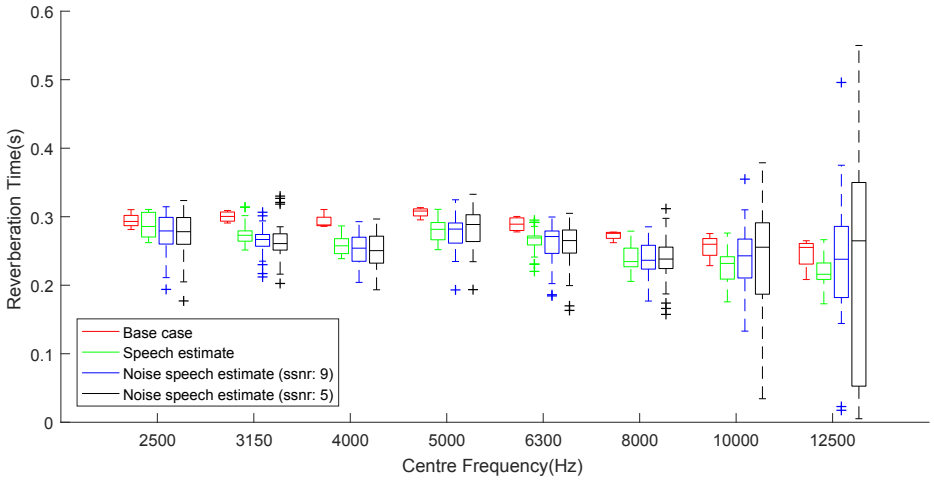
Figure 3.29: Confusion plots - effects of noise (SSNR: 5.00)

Distribution of T60 estimates

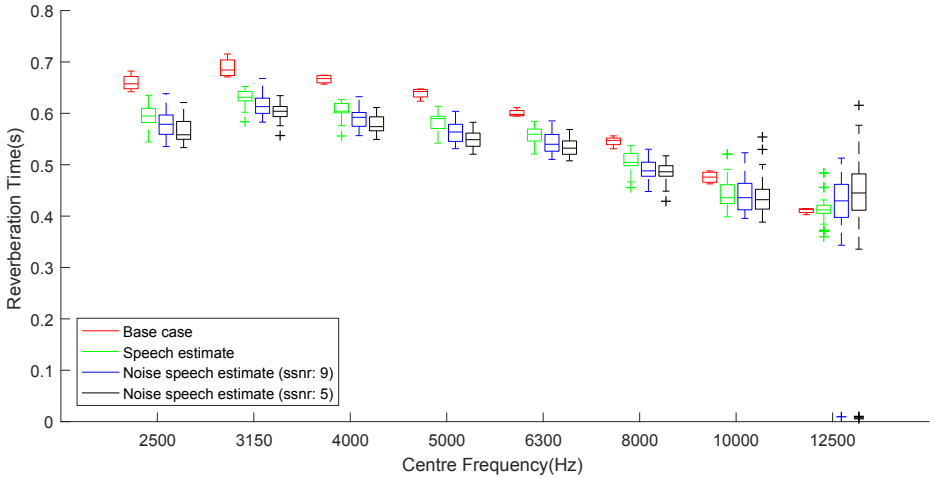
Figure 3.30 shows the distribution of T60 estimates at each third octave band for each room in the experimental data. The $T60$ estimates obtained using speech signals are very similar to the results gathered using the exponential sine-sweep method. The effect of adding white Gaussian noise can be seen in the increased variance of results.



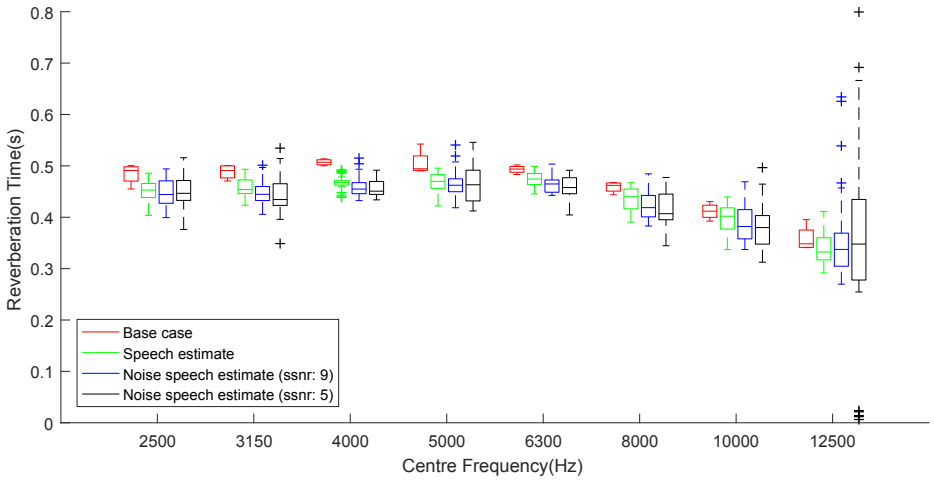
(a) Distribution of T60 estimates for the lounge for each third octave band



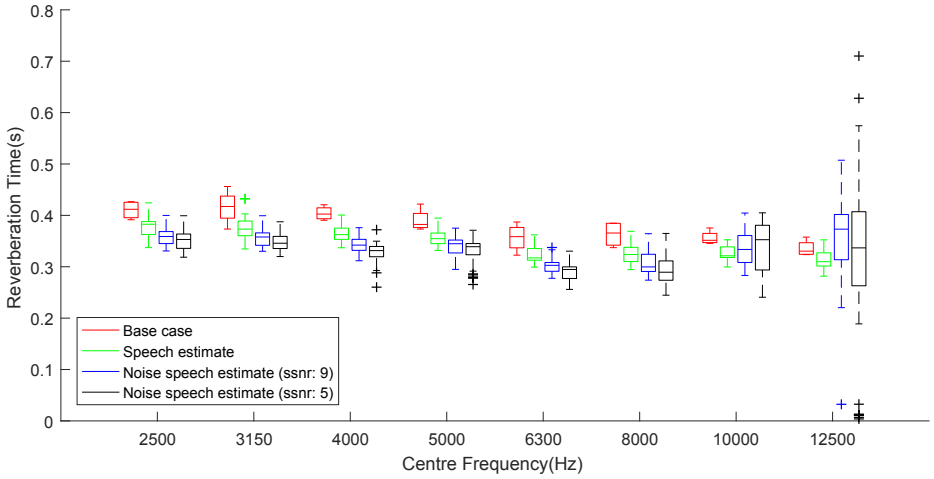
(b) Distribution of T60 estimates for the conservatory for each third octave band



(c) Distribution of T60 estimates for the bathroom for each third octave band



(d) Distribution of T60 estimates for the kitchen for each third octave band

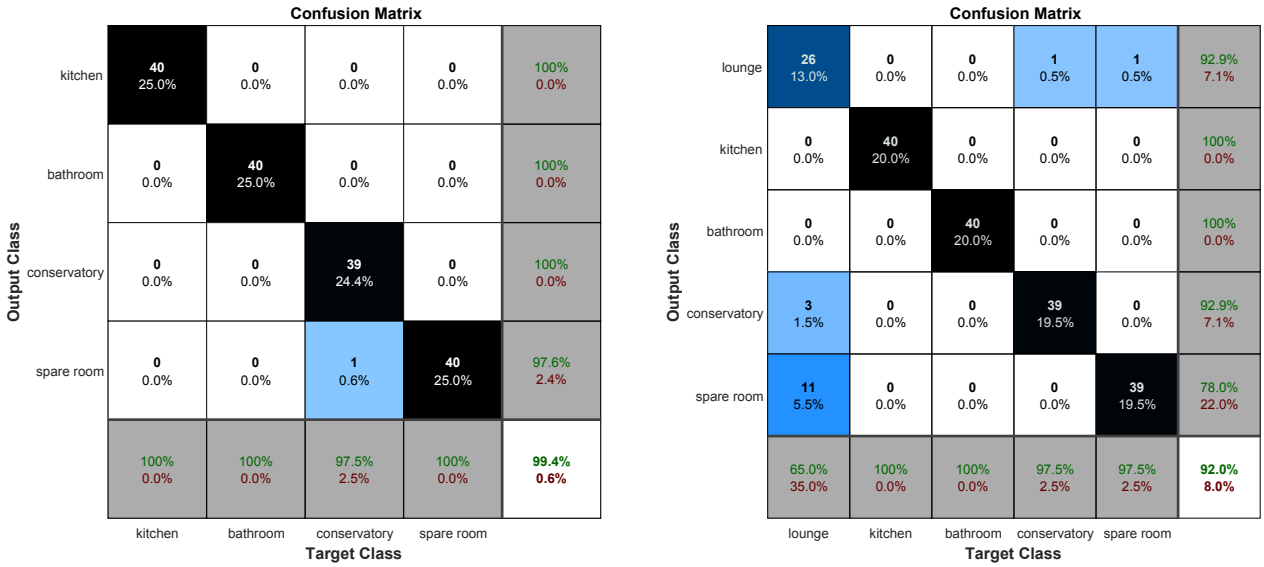


(e) Distribution of T60 estimates for the spare room for each third octave band

Figure 3.30: Distribution of T60 estimates for each third octave band

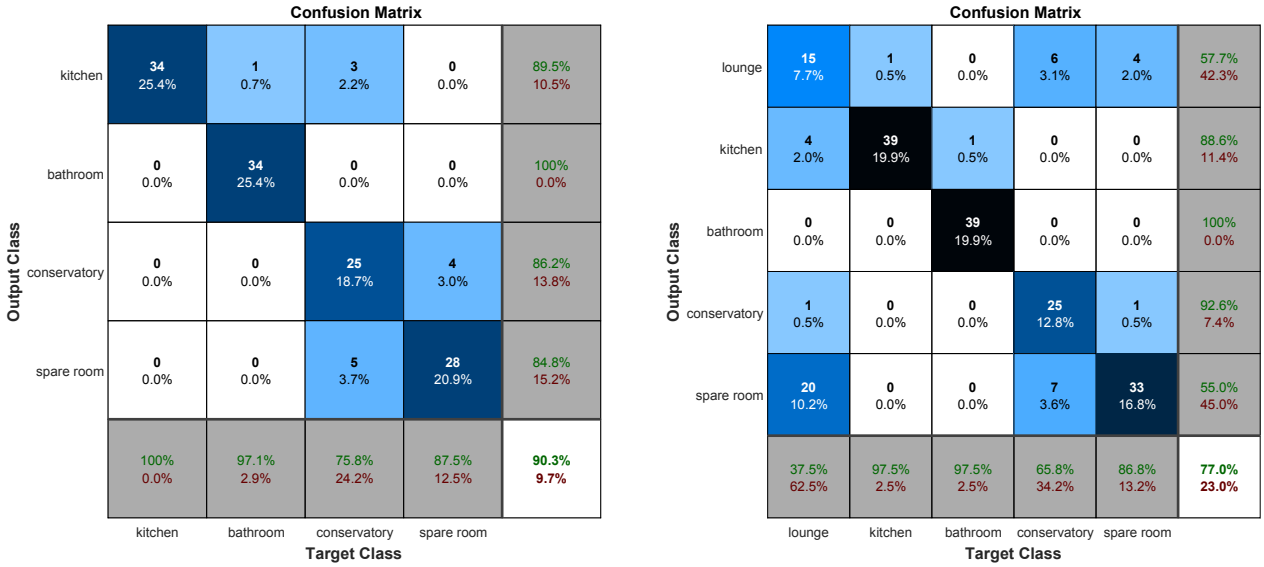
3.4.4 Using short speech signals

When calculating estimates of the RIR, generally the estimate is better when a longer speech signal is used as more frequencies are likely to be excited. This is, however, less than ideal in a practical situation as classification will take a much longer period of time. This is why estimates calculated using short speech signals need to be evaluated due to the fact that if reasonable classification can be achieved with shorter speech, this would be a far more exciting result.



(a) Confusion plot classifying 4 rooms (b) Confusion plot classifying 5 rooms

Figure 3.31: Confusion plots - comparing classification when identifying less rooms

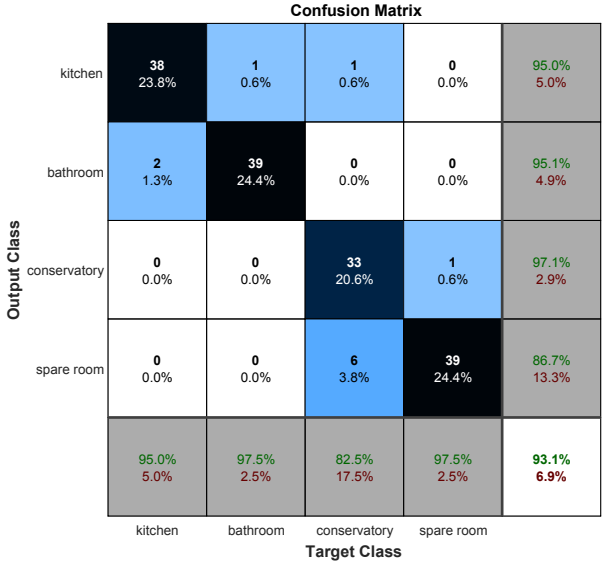


(a) Confusion plot classifying 4 rooms (b) Confusion plot classifying 5 rooms

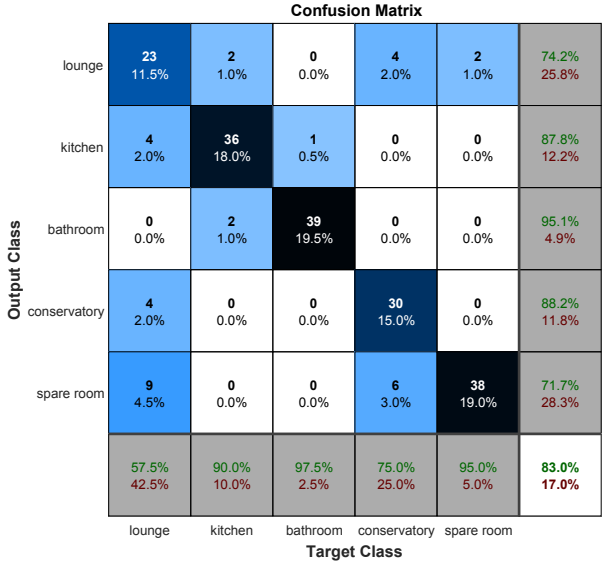
Figure 3.32: Confusion plots - effects of noise (SSNR:17.00)

3.4.5 Using short speech signals and a short filter length

To make classification extremely quick, ideally both a short speech signal and a short filter length would be implemented. The condition being that the filter length needs to be at least as long as the RIR. In this section, the deterioration of the estimates when highly optimising for speed of classification will be in focus. Figures 3.33 and 3.34 show the classification rates when short speech and a filter length of 600ms is chosen. They also show how these improvements to speed impact the robustness of the estimator when operating in a noisy environment.

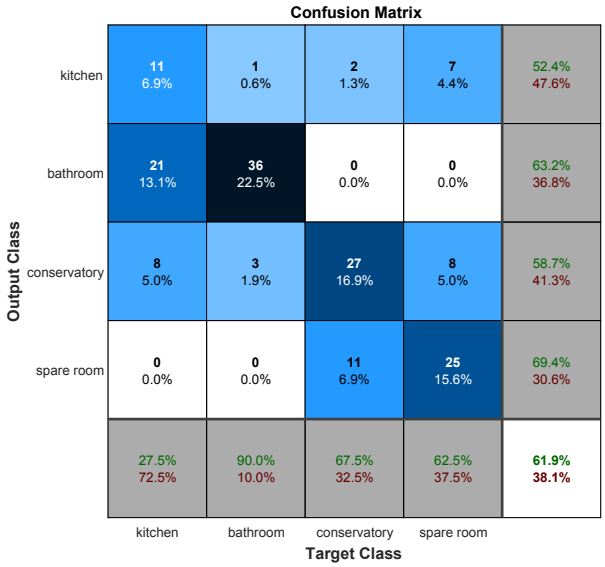


(a) Confusion plot classifying 4 rooms

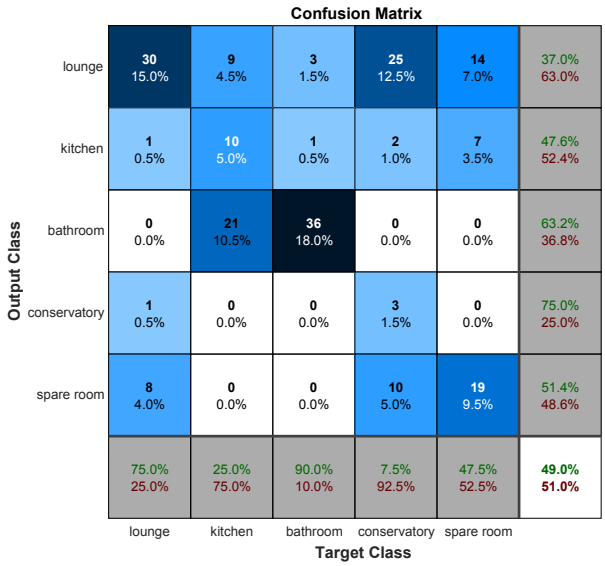


(b) Confusion plot classifying 5 rooms

Figure 3.33: Confusion plots - effects of filter length (L: 500ms)



(a) Confusion plot classifying 4 rooms



(b) Confusion plot classifying 5 rooms

Figure 3.34: Confusion plots - effects of noise (SSNR:17.00, L: 500ms)

4 Conclusion

4.1 Evaluation

The aim of this project was to investigate the hypothesis: 'Can a room be identified by its observable acoustic characteristics in a supervised scenario?'. The reverberation time (RT) was the characteristic that was explored due to previous research[8] showing success in using the RT for room identification. The focus of the project, therefore, was to expand on the existing work looking at improving the robustness of known methods in a supervised scheme. The setup that was analysed was one of a microphone and loudspeaker where the output from the loudspeaker was known.

Three approaches were implemented and tested on realistic synthetic speech signals generated from simulated ISM RIRs. It was observed at an early stage that time-domain estimation using least squares was not robust to noise and, therefore, this method was not carried further as it would not perform well in a practical situation where noise would be prevalent. The two other approaches worked in the frequency-domain, one being a typical conventional FD estimator and the other being an estimator that used nonstationarity. The simulated ISM RIR was used as a base case and the RT values at third octave frequency bands were calculated using a non-linear approach. The two frequency-domain methods were evaluated against the base case using two metrics, the absolute error in RTs and the NPM measure. The effects on the algorithms estimates due to noise; speech length and filter length, were studied. These two techniques were found to be robust to noise. Interestingly the nonstationarity estimator performed better under the NPM metric but was worse in terms of the absolute error in RTs. This had the consequence of the conventional method being used for future testing as it had the most robust RT estimates in a supervised situation.

The typical conventional FD estimator implemented was then tested using real RIR data from the ACE corpus to verify it works in a more complex setting. The real RIR data was used to generate the realistic recorded speech signals for testing the chosen method. The actual RIR used to generate the recording was again known in this case and, therefore, it was easy to evaluate the approach taken against a base case. The classification rate was found to be 100% when background noise was not contaminating the signal. The classification rate after large amounts of white Gaussian noise was added to the recorded signal (SSNR: 3.6) was found to be 95.7%. This was a promising result showing that the conventional FD estimator could perform well even under extreme conditions.

The last part of the project used experimental data collected using the setup described earlier of a microphone and loudspeaker. This was an important step in investigating the hypothesis as it was the first time the method was tested on real speech recordings. In this situation the RIR was not known so the exponential sine-sweep method was used to obtain an accurate base case for comparison purposes. The classification rate without noise was found to be 99.5% which proved the robustness of this approach in an ideal situation. Classification rates were also calculated after white Gaussian noise was added to the recorded signal. This caused a drop in accuracy,

however, two of the rooms had similar RTs and thus when only one of the two rooms was used in classification, the result was greatly improved. When the 4 rooms were classified and the SSNR was equal to 9.00, a classification rate of 100% was achieved. The classification rate dropped 96.3% when the noise was increased to an SSNR of 5.00. These results were obtained when using optimal parameters for robustness. Analysis was also performed to examine the outcome if speed of classification was the main concern instead of robustness. This analysis was carried out to investigate the limitations to the final chosen approach.

The hypothesis to this project was proven to be true. It is possible to identify a room using reverberation time in a supervised scenario. This was indeed still the case in the presence of noise highlighting the robustness of the method. This report ultimately shows how to implement a known FD RIR estimator with slight modifications within a new context. It shows how a single microphone and loudspeaker setup can be used for robust room identification with speech signals, instead of sine-sweeps, as a less intrusive way of obtaining the RIR.

4.2 Further work

To summarise, this project successfully proved the hypothesis it set out to investigate, however, if there was more time available further improvements could have been achieved. These improvements are stated below:

This project focused on evaluating the method using a setup of a single microphone and loudspeaker. This was due to time constraints. If this project was to be taken further, experimental data would also be collected from multiple speaker positions. It would also be advantageous to take recordings using a robot to investigate the approach applied in the context of a robot being able to identify a room using speech.

The background noise used to contaminate the speech signal throughout this project was white Gaussian noise. Ideally babble noise would be used as it is more realistic to a practical situation. This is one area where the results obtained could be improved if further work was performed.

To prove the hypothesis of this report more effectively, results from multiple homes should be taken, however, due to the time constraints of the project, the assumption that most houses contain rooms with differing reverberation time was invoked. If this project was to be taken further, more houses would need to be tested and 'research community' approved measurement equipment would have to be used. This, however, does not distract from the result obtained from what could be perceived as a typical UK house.

Appendices

A Speech test signals

This section shows a list of all the speech signals used in this report, listed in the following format [gender identifier, person index].

A.1 Female speech

A.1.1 Short speech length

F1 I live in The Hague.

F2 I live in Delft.

F3 I live in London.

F4 I live in Saint Ives.

F5 I live in Rotterdam.

A.1.2 Medium speech length

F1 I live in a rented apartment. It has two bedrooms, a kitchen, a large living room and also of course a bathroom.

F2 I live in a flat. It's a twelve story building and I live all the way on top. I have two bedrooms, a kitchen, a very big living room and a very nice view from my living room balcony.

F3 I live in a room with four walls. It has no ceiling. The rain comes in and nothing happens.

F4 Actually, I live in a shoe. That comes from a different nursery rhyme. It has two bedrooms and a kitchen and it also has windows.

F5 I live in a rented apartment. It has one bedroom, a kitchen and a living room.

A.1.3 Long speech length

F1 I get to work by car. I drive out of our street and at the first sub- traffic light I go to the right. After that there are some err several others so traffic lights and after a while there's a gas station at the right another traffic light after that. I go to the right and that street leave leads to the highway. It's the A13 so I get on the highway and the A13 and goes by Delft and later on Rotterdam. Sometimes I drop off my boyfriend at the the the at the University of Delft and other days I just drive and straight to Rotterdam. I get off the highway err I get from the A13 to the A20 so that's the highway around Rotterdam and I get off the highway at err Rotterdam Alexander and at that point I'm actually already in my work. There's not really much to see by the way.

F2 Usually I would walk to work its not very far from where I work so its basically just crossing two streets and walking through another so when I go to work err along the way first I have to get err through my elevator all the way down. I walk out of the flat. I walk past the daycare centre err past the bike shed. Then it gets tricky because I have to cross a very busy road especially if I leave in the morning and then I walk on and it's basically just offices so not very exciting to watch. Obviously when its early there will be many people biking around. Errm I'm basically already at the university campus so there is a lot of traffic and a lot of bikes that have no clue where they're heading so you have to be a little bit careful then when crossing the two bike lanes erm. While crossing that you will be in the Mekelpark so it's a nice park, which used to be a car road but now its just green. And then I walk through this other street. Erm Getting very close to the building where I work. It tends to be very windy out there so you have to be a little bit careful. But then you go past there I'm already basically at the entrance of my office.

F3 I am sitting in a room at the moment and it's actually completely fall of stuff. I can't hear very much, just my own voice. And actually the floor is also completely err, transparent it's only a wire mesh so I am kind of floating. It's pretty cool actually. There are lights so you can actually see all the way to the ceiling, the ceiling is also the same as the walls. Strange triangles peering out. Erm, it's kind of an inspirational space I guess. Everything is kind of a creamy yellow colour and I am suspended. It's a bit like erm a kind of sci-fi feeling. And next to me there are some really strange bags. Err, atrange bags that are full of sponges. These sponges are quite interesting because actually they make the whole thing look like a...a

strange art installation and I am actually sitting here in the middle probably part of this installation, talking to myself and wondering what's going to happen. There are ropes on the ceiling. These ropes seem to be suspended. Who knows were they are connected to, or what they are for. Maybe all sorts of things.

F4 I go to work every day and I walk. On the way to work I see a man. The man is accompanied by seven women. I assume that these are his wives because that's the way the nursery rhyme goes. Each of these wives has seven sacks and in each of the sacks there are seven cats or at least I've been told there are seven cats 'cause I've never actually seen the cats. The cats each in turn has seven rats. Erm, the rats I have also not seen. Erm, but I think that they...I really think that they are there. This is what I see every day on the way when I go to work.

F5 When I get to work, I first have to go to the tram station. Because I live in Rotterdam and I work in Delft, I first have to walk to the tram to get to the central station. Once I get to the tram, I get in there and see the...the apartments of other people living in Rotterdam. When I reach the central station, I cross a bridge first. It's a beautiful bridge that crosses a big river in Rotterdam. After getting out of the tram I reach the newly built central station. It has a lot of people and it's quite crowded. From there I can either choose to go by bus or go by train. If I go by train, it will be faster, but there might be delays along the way. So, typically I go by bus. By going by bus I can also pass some nicer scenery. So, for instance, I can see the cows and the sheep in the meadow. When I get to Delft, I can either take another bus or walk to my work. If I walk to work, I can get some exercise along the way as well. But the bus is faster again. When I get to work I can take the elevator to the eleventh floor to locate my office. I open the door and there I am, ready to work.

A.2 Male speech

A.2.1 Short speech length

M1 I live in The Hague.

M2 I live in the Hague.

M3 I live in Delft.

M4 I live in Delft.

M5 I live in Wellington.

A.2.2 Medium speech length

M1 I live in in a rented apartment with one bedroom a kitchen and a bathroom.

M2 I live in a rented apartment. I've got a great living room. A small sleeping room and a working room where we study a lot, and a small kitchen.

M3 I am renting an apartment from DUWO. It's a two-bedroom apartment having a shower, a toilet and a kitchen. Shower and toilet are actually separated.

M4 I live in a rented apartment, which has one bedroom, err, a living room and an open plan kitchen.

M5 Er in Wellington, I rent an apartment with six other people. Er we have one kitchen, er two toilets, a shower, a big living room and a bedroom for each of us.

A.2.3 Long speech length

M1 To get to work each day I go to central station where I pass through er the Malieveld. The train then passes through Rijswijk as well as Delft which I get off at. Err from there I walk along the the canal in the direction of the err EWI building at TU Delft. I turn and then cross the canal and then I walk downwards along the err the Bridge and then I generally pass the the different TU Delft buildings. Finally when I enter the building I get into the elevator which I take to the 11th floor. And then my office is at the end of the 11th floor.

M2 I go to work by tram. I take it from out of my house. It's about 30 seconds err to walk to the err to the tram and then we go to Delft. First we came across Rijswijk - a small village and then, we travel along the 'Schie'. It's a river it's erm from The Hague up to Delft and the we enter Delft, the beautiful Delft. We see a lot of great building place where they reinstall the new station. And every month you come across the place then you see something new. You see a new building, a great gap in the ground. It's quite fun to see. And then I ss I stop at Delft station and I am taking a walk for about twenty minutes up to the faculty, of the university and then I am al...almost there. I have to get in the elevator, press the button, one of the highest, the 17th floor, and when I arrive at this floor I walk to my working room where I have two nice colleagues - Nikolay and Jorge.

M3 I live rather close to my work. Actually, the horizontal distance I need to cover every morning is the same as the vertical distance in order to get to my living roo, err, to my working space. From my living room which is in the err, fifth floor, I need to get down to the ground level and then simply walk about 500 meters in the direction of the university as I live directly, behind it. From there I have to take the elevator to the 17th floor so you can directly see that the horizontal distance which I need to cover is equivalent to the vertical distance. On my way to work I go through the Mekelpark, which is the park in-between all the faculties of the university. There every day I can see the trees, which are to the left and to the right of me, and it makes me always think like what would have been this walk to work differently when I would live ten years ago. Ten years ago there was no park but there was a street, which was very busy. Nowadays, the park is very idyllic, which I really enjoy. Therefore, my way to work already gets me in a good mood while I have to get to the elevator.

M4 I usually go to work by bicycle. Erm, I...I cycle through the city center of Delft. Which is a very nice er, old European town. I first pass the, the leaning Old Church, after which, I...I pass by the, the fish market, which claims to have been there since er, since the early fourteenth century. Er from there, there's a little passage er, next to the town hall and out on the Market Square that er erm, and there I face the New Church, which is new, well at least relative to the Old Church. It's still a very old building. Er, from there on I...I go through er Beestenmarkt and Bastiaanplein over the bridge and down onto the university campus area. I park my bike in...in the, in the bicycle shed of our building and then I take the lift up to the seventeenth floor where...where my office is.

M5 To get to my work, I...I go out my front door, and beside us is a Harley Davidson shop, which is quite noisy. And then er, I'd go to the end of the street and turn left and into the main street of Wellington, which is called Courtney Place. Er, and on Courtney Place I pass things like a, there's a big er, cinema or theatre called the Embassy erm, and there's a whole bunch of stores that sell like shops and like fancy restaurants and things like that. Er it's usually quite busy 'cause everyone's hitting to work at the same time. Erm, so then I walk along Courtney Place until I come to a place called Cuba Street, which is sort of like a bit more cultural. There's like a lot of cafes and maybe musicians busking, on the street. Erm, once I go past Cuba Street I...I go up a whole bunch of steps called the Dixon Street Steps. There's like a hundred and forty three of them or something like that. Erm, then you rise about fifty, sixty meters from the street below, so it's quite tiring. Erm, yeah, then you sort of wind up 'round a couple of streets past er the, like the hostels where the students live. Erm, and then you end up on campus and from that you go past the law faculty and and the...the arts department, the humanities section as well. Until you come to sort of the science and engineering block and then, I would go up to the engineering building. And this is about ten stories tall, I think. And from the windows you get a really, nice view of Wellington, ah, the ocean and also Mount Victoria. It's not really a mountain though. Ha.

B ACE corpus Room impulse responses

B.0.1 Room impulse responses

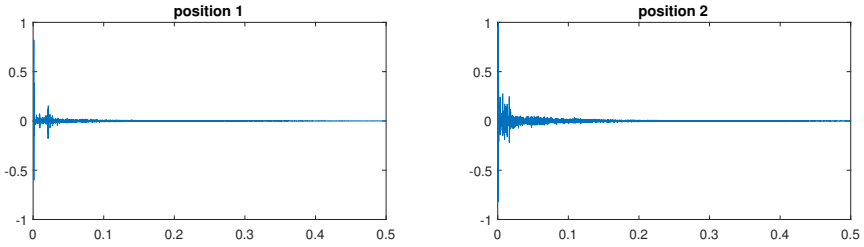


Figure B.1: EE lobby - room impulse responses

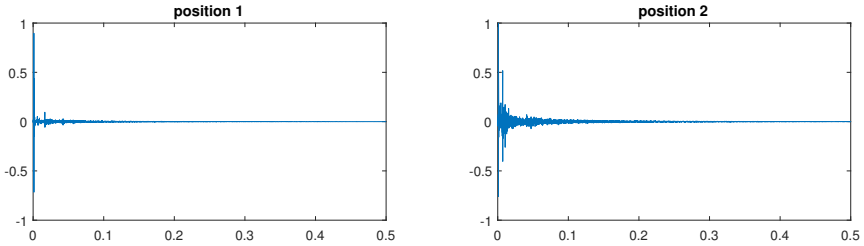


Figure B.2: Room 508 - room impulse responses

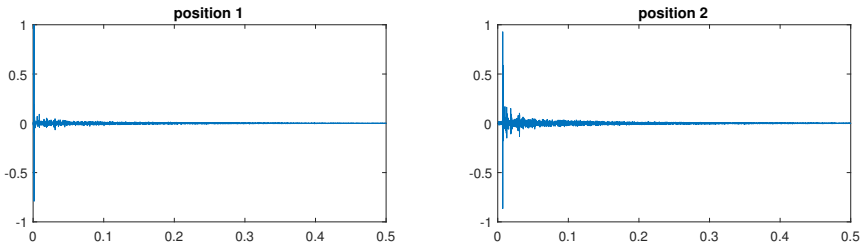


Figure B.3: Room 403a - room impulse responses

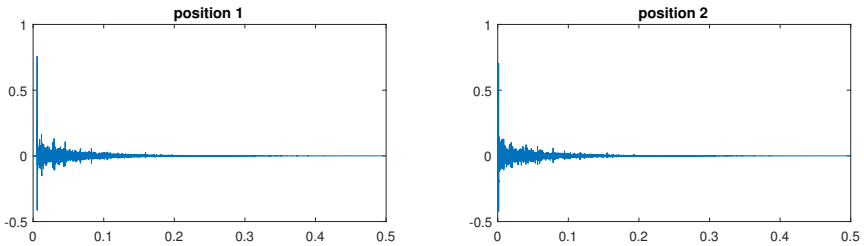


Figure B.4: Room 503 - room impulse responses

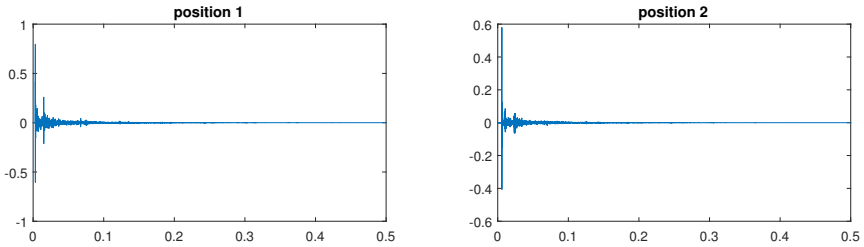


Figure B.5: Room 611 - room impulse responses

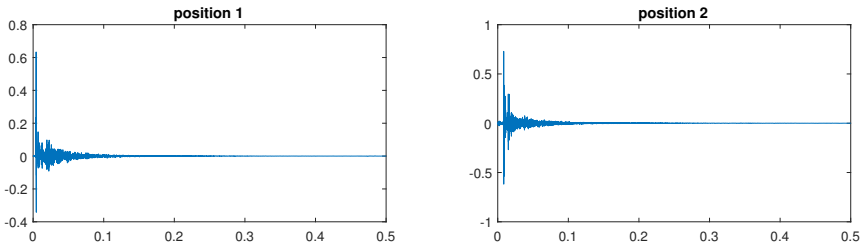


Figure B.6: Room 502 - room impulse responses

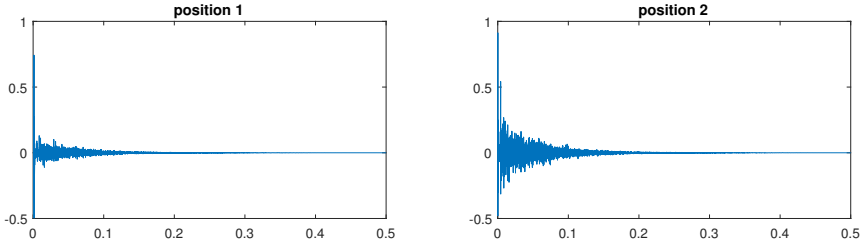


Figure B.7: Room 803 - room impulse responses

C Experimental data

C.1 Data collection setup

(a) Position 1

(b) Position 2

(c) Position 3

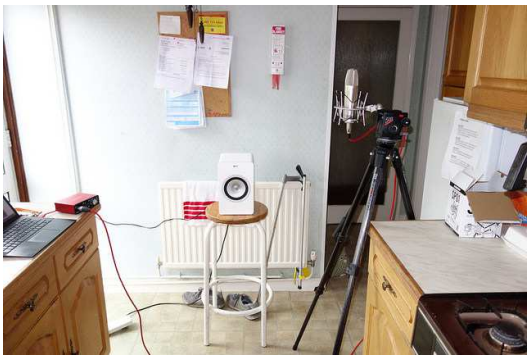
(d) Position 4

Figure C.1: Lounge

(a) Position 1

(b) Position 2

(c) Position 3



(d) Position 4

Figure C.2: Spare room

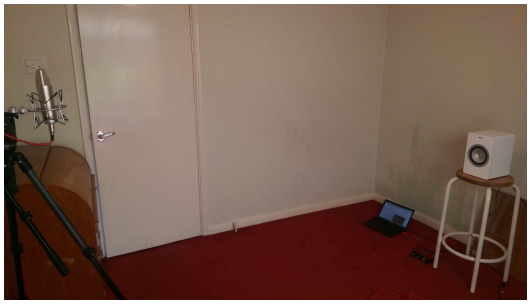
(a) Position 1

(b) Position 2

(c) Position 3

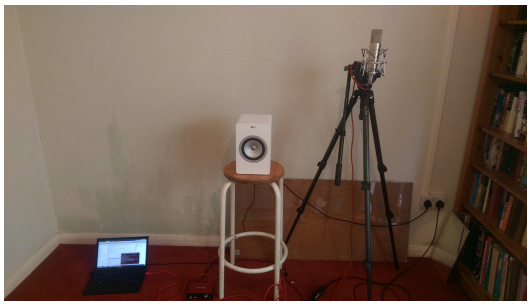
(d) Position 4

Figure C.3: Bathroom

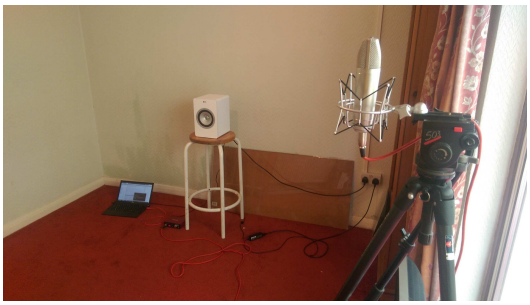


(a) Position 1

(b) Position 2

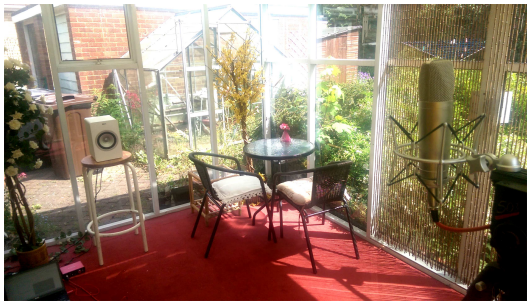


(c) Position 3



(d) Position 4

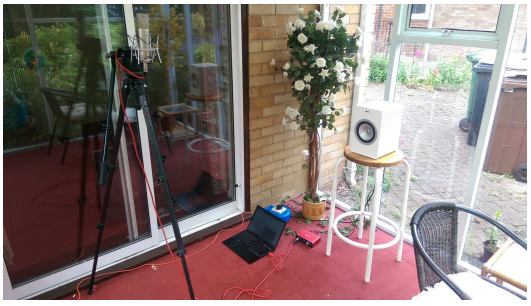
Figure C.4: Spare room



(a) Position 1

(b) Position 2

(c) Position 3



(d) Position 4

Figure C.5: Conservatory

C.2 Measuring equipment

(a) M-Audio Nova condenser microphone[1]

(b) KEF X300A speaker[2]

Figure C.6: Measuring equipment

[1] www.m-audio.de/download/file/fid/989

[2] http://www2.kef.com/uploads/files/en/x/download/x_information_en.pdf

C.3 Exponential sine-sweep method

C.3.1 Room impulse response estimation

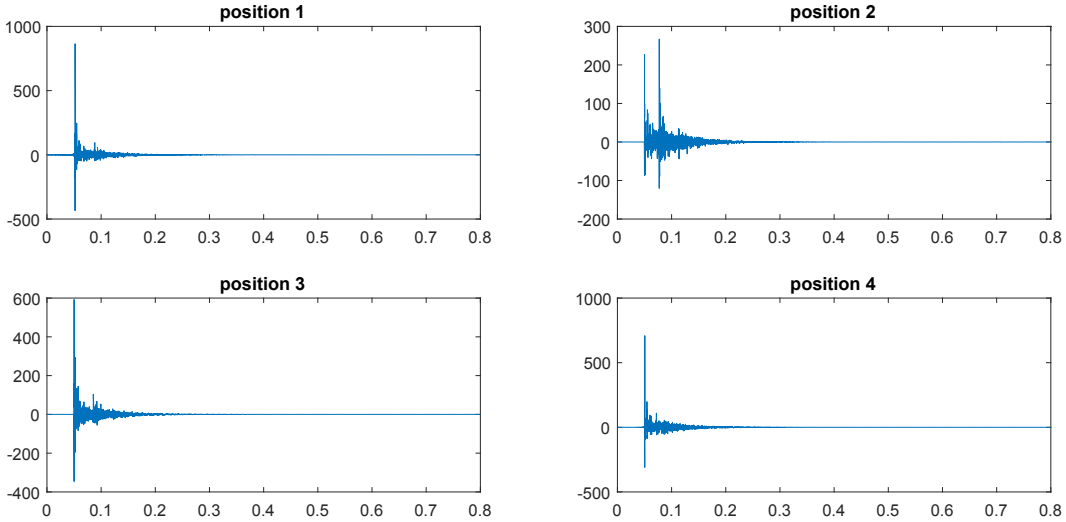


Figure C.7: Lounge - room impulse responses

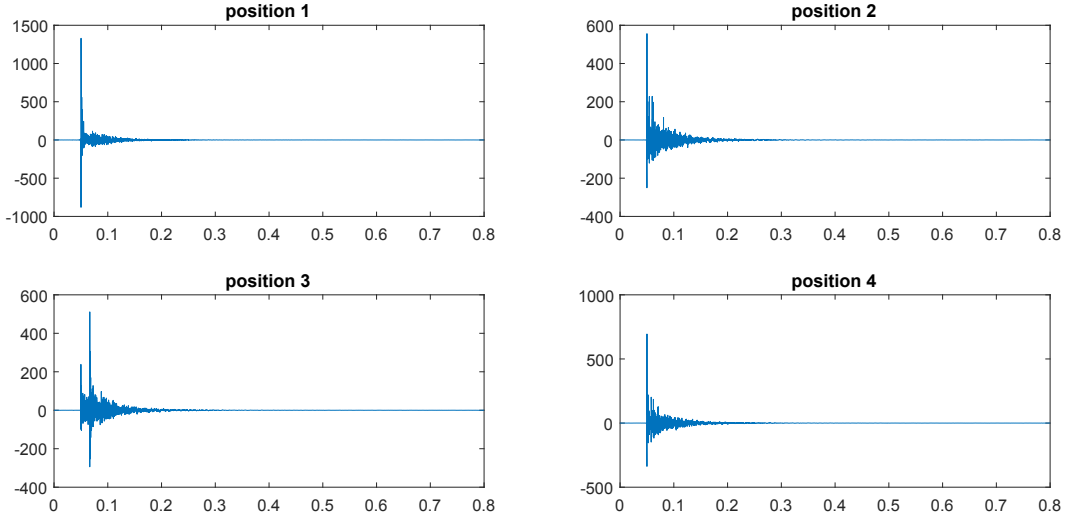


Figure C.8: Conservatory - room impulse responses

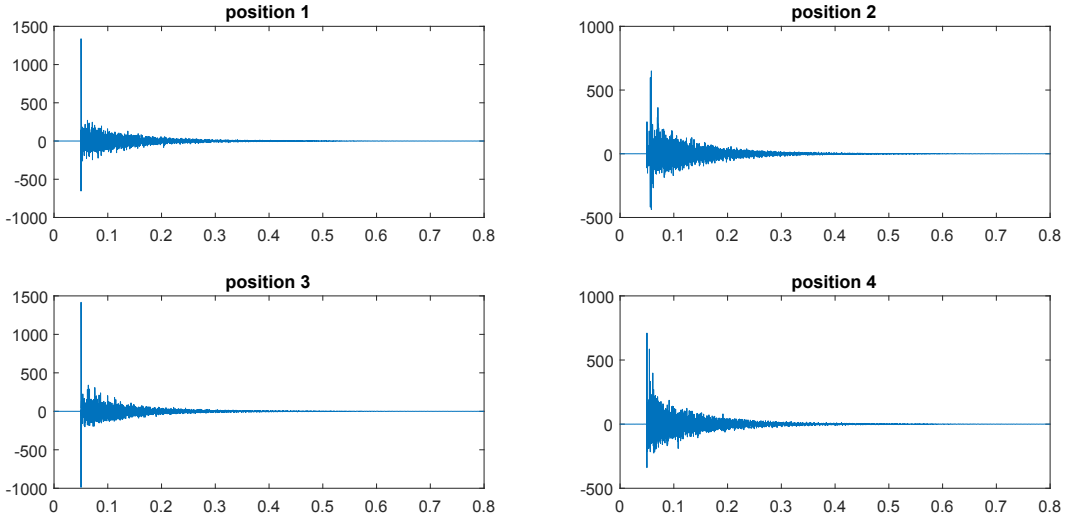


Figure C.9: Bathroom - room impulse responses

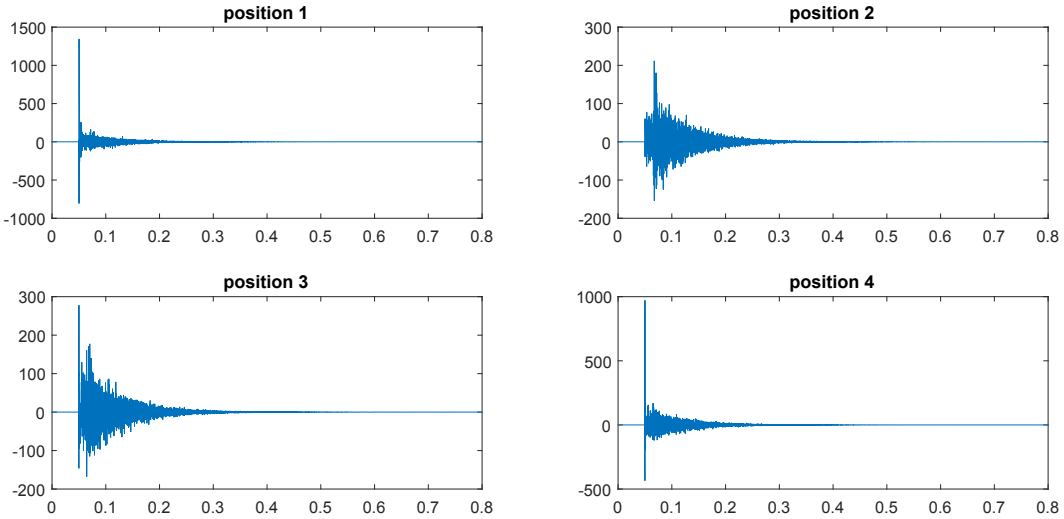


Figure C.10: Kitchen - room impulse responses

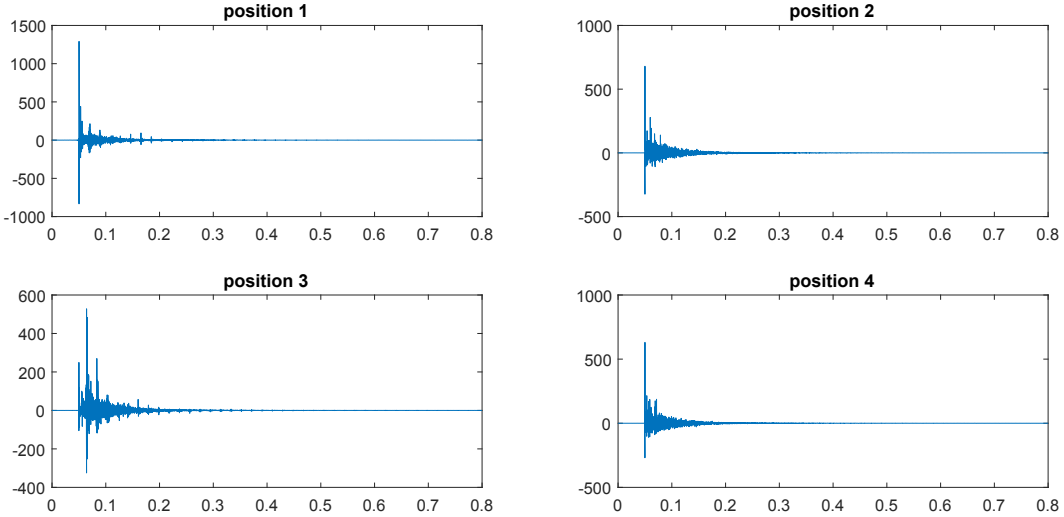


Figure C.11: Spare room - room impulse responses

C.3.2 Reverberation time estimation

Reverberation Time(s)									
Center Frequency(Hz)	*2000*	*2500*	*3150*	*4000*	*5000*	*6300*	*8000*	*10000*	*12500*
Lounge									
Position 1	0.18	0.35	0.36	0.36	0.37	0.34	0.34	0.34	0.36
Position 2	0.20	0.39	0.42	0.41	0.40	0.34	0.31	0.28	0.27
Position 3	0.17	0.35	0.36	0.35	0.35	0.31	0.32	0.30	0.28
Position 4	0.18	0.34	0.36	0.38	0.39	0.36	0.35	0.36	0.36
Conservatory									
Position 1	0.16	0.29	0.30	0.29	0.31	0.28	0.27	0.26	0.25
Position 2	0.15	0.29	0.31	0.29	0.31	0.30	0.28	0.26	0.26
Position 3	0.15	0.28	0.30	0.31	0.30	0.28	0.26	0.23	0.21
Position 4	0.15	0.31	0.29	0.29	0.31	0.30	0.28	0.28	0.26
Bathroom									
Position 1	0.33	0.64	0.67	0.66	0.65	0.61	0.55	0.47	0.41
Position 2	0.36	0.65	0.68	0.66	0.64	0.60	0.53	0.46	0.40
Position 3	0.35	0.66	0.72	0.67	0.64	0.60	0.55	0.49	0.41
Position 4	0.36	0.68	0.69	0.67	0.62	0.59	0.56	0.48	0.41
Kitchen									
Position 1	0.24	0.46	0.48	0.50	0.50	0.48	0.44	0.41	0.35
Position 2	0.25	0.50	0.50	0.50	0.49	0.50	0.47	0.42	0.34
Position 3	0.26	0.50	0.47	0.51	0.54	0.49	0.46	0.39	0.34
Position 4	0.26	0.49	0.50	0.51	0.49	0.50	0.47	0.43	0.40
Spare Room									
Position 1	0.22	0.43	0.46	0.42	0.42	0.37	0.38	0.38	0.36
Position 2	0.20	0.40	0.37	0.39	0.38	0.32	0.34	0.35	0.34
Position 3	0.21	0.42	0.42	0.41	0.39	0.39	0.38	0.35	0.32
Position 4	0.20	0.39	0.42	0.40	0.37	0.35	0.35	0.35	0.32

Table C.1: Reverberation time estimation using the exponential sine-sweep method

D Code listing

D.1 Exercise 1

```
function [g_FD, G_FD] = TD(length,sclean,srecorded)
% time-domain estimation using least squares
% inputs: sclean: output signal from loudspeacker
%         srecorded: signal from microphone
%         length: length of the estimated impulse response
% outputs: g_FD:  estimated room impulse response
%          G_FD: estimated room transfer function

length=length-1;

Y = zeros(length+1,1);
y = xcorr(sclean, sclean, length);
Y = Y + flipud(y(1:length+1));

s = xcorr(srecorded, sclean, length);
S = s(length+1:2*length+1);

g_FD = block_levinson(S, Y);
G_FD = fft(g_FD);
```

Listing D.1: Time-domain estimation using least squares

```
 function [g_FD, G_FD]=FD(length,sclean,srecorded,shift)
% conventional frequency-domain estimator
% inputs: sclean: output signal from loudspeacker
%         srecorded: signal from microphone
%         length: length of the estimated impulse response
%         shift: shift between windows in the STFT
% outputs: g_FD:  estimated room impulse response
%          G_FD: estimated room transfer function


% STFT
Sclean=stft(sclean',length,shift);
Srecorded=stft(srecorded',length,shift);

G_FD=mean(conj(Sclean).*Srecorded,2)./mean(conj(Sclean).*Sclean,2);
g_FD=real(ifft(G_FD));


function S = stft(s, nlength, winshift)

N = ceil((length(s)-nlength)/winshift)*winshift + nlength;
s = [s zeros(1,N-length(s))];

window=(N-nlength)/winshift;
S=zeros(nlength, window + 1);

for j=0:window
    S(1:nlength,j+1)=s(j*winshift+1:j*winshift+nlength)'.*boxcar(nlength);
end
S=fft(S);
```

Listing D.2: Conventional frequency-domain estimator

```
function [g_FD, G_FD] = NSFD(sclean,srecorded,T_sg,FLE_l,FLE_r)
% frequency-domain estimator using nonstationarity
% Estimates auto and cross-spectra of sclean and y within sub-periods specified
% by 'periods'.
% inputs: sclean: output signal from loudspeacker
%         srecorded: signal from microphone
%         periods: length of each segment (the spectrum is calculated
%         seperatly: in each segment
%         leng: 2*leng+1 is the window length in Blackman- Tukey method for spectral-estimation
%         (the window type is specified inside this program - Hamming).
% outputs: g_FD: estimated room impulse response
%          G_FD: estimated room transfer function

% Coded based on Sharon Gannot implementation

length = 1*(FLE_l + FLE_r);
window = hamming(2*length+1)';
Nfft = 2^nextpow2(2*length+1);

Sxx = zeros(size(T_sg,1),Nfft);
Syx = zeros(size(T_sg,1),Nfft);

for bl = 1 : size(T_sg,1)
  T = T_sg(bl,1):T_sg(bl,2);
  x1 = sclean(T);
  y1 = srecorded(T);

  maxlag = length;

  r_xx = xcorr(x(T),maxlag,'none').*window;
  r_yx = xcorr(srecorded(T),sclean(T),maxlag,'none').*window;
  K_xx = zeros(1,Nfft);
  K_xx(1:maxlag+1) = r_xx(maxlag+1:2*maxlag+1);
  K_xx(Nfft-maxlag+1:Nfft) = r_xx(1:maxlag);
  K_yx = zeros(1,Nfft);
  K_yx(1:maxlag+1) = r_yx(maxlag+1:2*maxlag+1);
  K_yx(Nfft-maxlag+1:Nfft) = r_yx(1:maxlag);

  Sxx(bl,:) = fft(K_xx,Nfft);
  Syx(bl,:) = fft(K_yx,Nfft);
end

theta = zeros(2,Nfft/2+1);

for w = 1:Nfft/2+1
  [theta(:,w)] = [Sxx(:,w) , ones(size(Sxx(:,w)))] \ Syx(:,w);
end

g_FD  = [theta(1,1:Nfft/2+1) conj(theta(1,Nfft/2:-1:2))];
Svx = [theta(2,1:Nfft/2+1) conj(theta(2,Nfft/2:-1:2))];
H = ifft(g_FD,[],2);
G_FD = real([H(Nfft-FLE_l+1:Nfft) H(1:FLE_r+1)]);
```

Listing D.3: Frequency-domain estimator using nonstationarity

Bibliography

[1] Nitin Sawhney and Pattie Maes. *Situational Awareness from Environmental Sounds*. 1997.

[2] Lang Tong, Guanghan Xu, and T. Kailath. "Blind identification and equalization based on second-order statistics: a time domain approach". In: *IEEE Transactions on Information Theory* 40.2 (Mar. 1994), pp. 340–349. ISSN: 0018-9448. DOI: 10.1109/18.312157.

[3] O. Shalvi and E. Weinstein. "System identification using nonstationary signals". In: *IEEE Transactions on Signal Processing* 44.8 (Aug. 1996), pp. 2055–2063. ISSN: 1053-587X. DOI: 10.1109/78.533725.

[4] S. Gannot, D. Burshtein, and E. Weinstein. "Signal enhancement using beamforming and nonstationarity with applications to speech". In: *IEEE Transactions on Signal Processing* 49.8 (Aug. 2001), pp. 1614–1626. ISSN: 1053-587X. DOI: 10.1109/78.934132.

[5] I. Cohen. "Relative transfer function identification using speech signals". In: *IEEE Transactions on Speech and Audio Processing* 12.5 (Sept. 2004), pp. 451–459. ISSN: 1063-6676. DOI: 10.1109/TSA.2004.832975.

[6] Z. Koldovský, J. Málek, and S. Gannot. "Spatial Source Subtraction Based on Incomplete Measurements of Relative Transfer Function". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.8 (Aug. 2015), pp. 1335–1347. ISSN: 2329-9290. DOI: 10.1109/TASLP.2015.2425213.

[7] X. Li et al. "Estimation of the Direct-Path Relative Transfer Function for Supervised Sound-Source Localization". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11 (Nov. 2016), pp. 2171–2186. ISSN: 2329-9290. DOI: 10.1109/TASLP.2016.2598319.

[8] A. H. Moore, M. Brookes, and P. A. Naylor. "Roomprints for forensic audio applications". In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Oct. 2013, pp. 1–4. DOI: 10.1109/WASPAA.2013.6701854.

[9] Jonathan S. Abel et al. "Estimating Room Impulse Responses from Recorded Balloon Pops". In: *Audio Engineering Society Convention 129*. Nov. 2010. URL: http://www.aes.org/e-lib/browse.cfm?elib=15594.

[10] Takayuki Hidaka, Noriko Nishihara, and Leo L. Beranek. "Relation of acoustical parameters with and without audiences in concert halls and a simple method for simulating the occupied state". In: *The Journal of the Acoustical Society of America* 109.3 (2001), pp. 1028–1042. DOI: 10.1121/1.1340649. eprint: http://asa.scitation.org/doi/pdf/10.1121/1.1340649. URL: http://asa.scitation.org/doi/abs/10.1121/1.1340649.

[11] Heinrich Kuttruff. *Room Acoustics, Fifth Edition*. 2009.

[12] Jont B. Allen and David A. Berkley. "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4 (1979), pp. 943–950. DOI: 10.1121/1.382599. eprint: http://dx.doi.org/10.1121/1.382599. URL: http://dx.doi.org/10.1121/1.382599.

[13] Patrick M. Peterson. "Simulating the response of multiple microphones to a single acoustic source in a reverberant room". In: *The Journal of the Acoustical Society of America* 80.5 (1986), pp. 1527–1529. DOI: 10.1121/1.394357. eprint: http://dx.doi.org/10.1121/1.394357. URL: http://dx.doi.org/10.1121/1.394357.

[14] Eric A. Lehmann and Anders M. Johansson. "Prediction of energy decay in room impulse responses simulated with an image-source model". In: *The Journal of the Acoustical Society of America* 124.1 (2008), pp. 269–277. DOI: 10.1121/1.2936367. eprint: http://dx.doi.org/10.1121/1.2936367. URL: http://dx.doi.org/10.1121/1.2936367.

[15] Sabine W.C. *Collected Papers on Acoustics*. 1992.

[16] Angelo Farina. "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique". In: *Audio Engineering Society Convention 108*. Feb. 2000. URL: http://www.aes.org/e-lib/browse.cfm?elib=10211.

[17] Angelo Farina. "Advancements in Impulse Response Measurements by Sine Sweeps". In: *Audio Engineering Society Convention 122*. May 2007. URL: http://www.aes.org/e-lib/browse.cfm?elib=14106.

[18] M. R. Schroeder. "New Method of Measuring Reverberation Time". In: *The Journal of the Acoustical Society of America* 37.3 (1965), pp. 409–412. DOI: 10.1121/1.1909343. eprint: http://dx.doi.org/10.1121/1.1909343. URL: http://dx.doi.org/10.1121/1.1909343.

[19] Poju Antsalo et al. "Estimation of Modal Decay Parameters from Noisy Response Measurements". In: *Audio Engineering Society Convention 110*. May 2001. URL: http://www.aes.org/e-lib/browse.cfm?elib=9992.

[20] J. Eaton et al. "Estimation of Room Acoustic Parameters: The ACE Challenge". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10 (Oct. 2016), pp. 1681–1693. ISSN: 2329-9290. DOI: 10.1109/TASLP.2016.2577502.

[21] D. R. Morgan, J. Benesty, and M. M. Sondhi. "On the evaluation of estimated impulse responses". In: *IEEE Signal Processing Letters* 5.7 (July 1998), pp. 174–176. ISSN: 1070-9908. DOI: 10.1109/97.700920.

[22] P. Domingos and M. Pazzani. *Machine Learning*. 1997.

[23] Ron Kohavi. "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143. ISBN: 1-55860-363-8. URL: http://dl.acm.org/citation.cfm?id=1643031.1643047.