Reducing Microphone Artefacts in Live Sound

Alice Clifford

Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary, University of London

Thesis submitted in partial fulfilment
of the requirements of the University of London
for the Degree of Doctor of Philosophy

2013

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. I acknowledge the helpful guidance and support of my supervisor, Dr Joshua Reiss.

Abstract

This thesis presents research into reducing microphone artefacts in live sound with no prior knowledge of the sources or microphones. Microphone artefacts are defined as additional sounds or distortions that occur on a microphone signal that are often undesired.

We focus on the proximity effect, comb filtering and microphone bleed. In each case we present a method that either automatically implements human sound engineering techniques or we present a novel method that makes use of audio signal processing techniques that goes beyond the skills of a sound engineer. By doing this we can show that a higher quality mix of a live performance can be achieved.

Firstly we investigate the proximity effect which occurs on directional microphones. We present a method for detecting the proximity effect with no prior knowledge of the source to microphone distance. This then leads to a method for reducing the proximity effect which employs a dynamic filter informed by audio analysis.

Comb filtering occurs when the output of microphones reproducing the same source are mixed together. We present a novel analysis of how the accuracy of a technique to automatically estimate the correct delay of the source between each microphone is affected by source bandwidth and the windowing function applied to the data.

We then present a method for reducing microphone bleed in the multiple source, multiple microphone case, both in determined and overdetermined configurations. The proposed method is extended from prior research in noise cancellation, which has not previously been applied to musical sound sources. We then present a method for simulating microphone bleed in synthesised drum recordings, where bleed enhances the realism of the output.

Through subjective listening tests and objective measures each proposed method is shown to succeed at reducing the microphone artefacts while preserving the original sound source.

Acknowledgements

Firstly I would like to thank my supervisor, Josh Reiss, for his advice and guidance throughout this journey. I also extend this thanks to Mark Plumbley and Mark Sandler for letting me join the wonderful community of people that make up the Centre for Digital Music at Queen Mary, University of London. I have enjoyed every minute of my time there.

Thanks go to everybody in the Centre for Digital Music for the discussions and advice in such a wide variety of topics. There are far too many people to mention by name but to pick out a few who were important in the completion of this thesis, many thanks go to Martin Morrell, Sonia Wilkie, Steve Hargreaves, Holger Kirchoff, Yading Song, Michael Terrell, Magdalena Chudy, Daniele Barchiesi, Emmanouil Benetos, Becky Stewart, Sam Duffy, Dan Stowell and Asterios Zacharakis. I would also like to thank Alan Boyd, Elias Kokkinis and Christian Uhle for their helpful advice.

Special thanks go to Enrique Perez Gonzalez for inspiring this research and for his support through the ups and downs of the past three and a half years. Thanks also go to my parents Janet and Colin Clifford. Without their unwavering support and belief in me I would not be where I am today. Thanks also go to the rest of my family for keeping me grounded and always putting things in perspective. I would also like to thank my close friends Katie, Aaminah, Lois and Kelly for providing much needed distractions from audio research.

Thanks also go to everyone at FXpansion Audio, particularly Henry and Angus, and to Queen Mary, University of London for funding my ImpactQM placement there.

I also acknowledge the EPSRC for providing the funding to support this research.

Vitaly puts on goggles, hooks himself into a computer on the sound truck, and begins tuning the system. Theres a 3-D model of the overpass already in memory. He has to figure out how to sync the delays on all the different speaker clusters to maximize the number of nasty, clashing echoes.

Neal Stephenson
"Snow Crash"

Contents

List of Figures

List of Tables

List of abbreviations

A-D	Analogue to digital
AEDA	Adaptive eigenvalue decomposition algorithm
ANC	Adaptive noise cancellation
BSS	Blind source separation
CTRANC	Crosstalk resistant adaptive noise cancellation
DAW	Digital audio workstation
DFT	Discrete Fourier transform
DSP	Digital signal processing
DUET	Degenerate unmixing estimation technique
EQ	Equalisation
FDCTRANC	Frequency domain crosstalk resistant adaptive noise cancellation
FFT	Fast Fourier transform
GCC	Generalized cross correlation
GCC-PHAT	Generalized cross correlation with phase transform
ICA	Independent component analysis
LMS	Least mean squares
MCWF	Multichannel wiener filter
MDW	Multisampled drum workstation
MIMO	Multiple input multiple output
MUSHRA	Multiple stimuli with hidden reference and anchor
PCC	Pearson's correlation coefficient
PHAT	Phase transform
RIR	Room impulse response
RLS	Recursive least squares
ROTH	Roth processor
RMS	Root mean square
SAR	Signal-to-artefact ratio
SCOT	Smoothed coherence transform
SDR	Signal-to-distortion ratio
SIR	Signal-to-interference ratio
TDE	Time delay estimation
TDOA	Time difference of arrival

Chapter 1

Introduction

1.1 Objectives

The aim of this research is to use signal processing to reduce artefacts that occur on microphone signals in a live sound context. We set out to answer the question of whether it is possible to reduce microphone artefacts with no prior knowledge of the sources or microphones, or the relative positions in an acoustic space. In this thesis microphone artefacts are defined as additional sounds or distortions that occur on the output of a microphone signal that alter the intended sound in an undesirable way.

This will be achieved by replicating the processes a human sound engineer undertakes to reduce these artefacts or by developing new signal processing methods that would ordinarily not be achieved by a human. This will be achieved by the following objectives:

- Comb filtering, proximity effect and microphone bleed that occur from using single and multiple microphones with single and multiple sources will be reduced using delay estimation, dynamic filtering and noise cancellation.
- Manual solutions that exist require a certain level of expertise in microphone placement. Many artefacts that occur are due to lack of knowledge in this area. Therefore any solution found will be able to be used by an amateur and will not require prior knowledge of the source and microphone configuration.
- As this research is aimed at live sound, any proposed method should be able to run in real time.
- As the artefacts are due to physical properties in the space, research into the reduction will take into account the physical properties of each artefact.

- Processing methods should preserve the target source with the least amount of additional distortion.

As we are concerned with researching methods for artefact reduction which make no assumptions about the source or microphone, we assume in this research that the only sources we are concerned with are intended sources that would be found in a live music performance. We therefore do not take into account external noise sources in this research and we also assume a low reverberation environment. It is known that noise and reverberation effect the performance of audio signal processing techniques and in this research we will focus on how the methods we research are affected by various sources. By not taking noise into account we will get a clearer idea of the performance of each method. In order to have a consistent reference across the research we also assume the complex radiation patterns of instrument sources are localised.

1.2 Motivations

Microphone artefacts are intrinsic to microphone design and sound engineering techniques exist to reduce them. These techniques are learnt from experience, which many amateur sound engineers and musicians do not have. Many of the artefacts can be attributed to the physical properties of the microphone and the space. There is little that can be done by the user to change the hardware of the microphone and often nothing can be done about the space the microphone is placed in. Limited studies have been conducted into how to reduce the appearance of artefacts using signal processing, which would require extra software with little or no input from the user.

Many modern microphones have some form of signal processing built in, commonly polar pattern switching and bass roll off. Recently, microphone manufacturers have begun producing more digital microphones, which have a built in analogue to digital converter tuned for the microphone. This shows that signal processing is already being used in microphone technology, but only where its implementation can be predicted by testing of the microphone.

More advanced signal processing could be included to reduce known artefacts that occur between the source and the microphone. This would mean a novice would still be able to get a high quality signal from the microphone, regardless of their expertise in microphone placement, and hear an expected output from the microphone. This in turn increases the clarity and quality of the microphone output, leading to an easier task for the sound engineer and ultimately a better experience for all people experiencing a music performance.

1.3 Research context

The research presented in this thesis fits within the umbrella of intelligent mixing tools, first presented by Perez Gonzalez and Reiss [2008a,b,c, 2007, 2009, 2010] and extended more recently by Giannoulis et al. [2013], Ward et al. [2012], Mansbridge et al. [2012a,b] and Maddams et al. [2012]. The aim of the intelligent mixing tools research is to provide tools to aid sound engineers, particularly amateur sound engineers, in providing an adequate baseline multitrack mix to enable them to spend more time on the creativity of mixing. This previous work is mostly concerned with the technical and aesthetic areas of mixing, such as level balancing, panning and automatic enhancement, rather than correcting problems in the recording process.

The research presented in this thesis strives to tackle the technical problems that occur specifically when using microphones and often poor microphone placement. The results can also be objectively measured.

Although this thesis is concerned with live sound, there are many other applications for the research. It is possible that aspects of the research can take an offline approach, which could be implemented in a recording studio environment. For example, offline approaches offer the flexibility of analysing a whole song and choosing the best course of action that would provide the optimal result over all the time. It was chosen to investigate live sound, where real time approaches could be established or at least implement block based approaches, since this is an open area of research. Live sound situations are often less controlled acoustics environments and it is likely the configuration will change over time therefore approaches need to be able to adapt to this. Studio production will generally be recorded in a controlled environment with acoustic control to tailor the reverberation and reduce some of the artefacts described here, such as bleed, in static conditions.

In live sound these artefacts are more often a problem due to the concert environment, for example the inability to adequately separate instruments, and possibly the lack of experience of the sound engineers involved. In smaller venues, they may even be the musicians themselves. Because of this it is likely there will be little knowledge of microphone placement techniques and artefacts are more likely to occur.

There are other, non-musical applications for the research outlined here. Theatre and broadcast environments suffer similar artefacts, along with any multiple source, multiple microphone situation, such as video conferences, which also suffer from noise and echoes [Habets and Benesty, 2013].

There is also scope for applying the research to audio forensics to improve the quality and intelligibility from audio evidence, or to gain extra information

such as location of sources and microphones with delay estimation research. It is also possible to apply the techniques to medical audio, such as heart sound recordings, for example removing crosstalk and aligning recordings [Hedayioglu et al., 2011].

1.4 Thesis structure

The microphone artefacts which are investigated in the research presented in this thesis are the proximity effect, comb filtering and microphone bleed. As there is only a small overlap between the approaches used to reduce each artefact, Chapters 3, 4 and 5 each contain a literature review and background of the state of the art in each field. Chapter 2 can be considered the background chapter for the overall thesis and contains information on how and why different microphone artefacts occur and introduces each area we discuss in the remainder of the thesis.

A chapter by chapter breakdown of the structure is as follows.

Chapter 1 - Introduction

In this chapter we outline the objectives and motivations of the research and outline the thesis contributions.

Chapter 2 - Background

This chapter provides a background in audio and microphone technology. From this microphone artefacts are categorised into environmental, positional and internal. We then describe in detail the cause and effect of the microphone artefacts that are investigated in this thesis.

Chapter 3 - Proximity effect detection and correction

In this chapter we propose a novel method for the detection and correction of the proximity effect. The novel detection algorithm uses spectral flux to detect low frequency changes in the signal that can be attributed to the proximity effect. A dynamic filter is then implemented to correct for theses effects.

Chapter 4 - Comb filter reduction

In this chapter we investigate using the GCC-PHAT delay estimation technique to reduce comb filtering in single source, multiple microphone configurations with arbitrary musical sources. A novel analysis of the effect of signal bandwidth and DFT window shape on the accuracy of the GCC-PHAT is provided.

Chapter 5 - Determined microphone bleed reduction

In this chapter we present a novel method for reducing microphone bleed in the determined multiple source, multiple microphone case. The method is based on a crosstalk resistant noise canceller from telecommunications research that has not previously been applied to musical instrument signals. It is extended by applying a multiple source version of the GCC-PHAT delay estimation technique from the previous chapter to centre the adaptive filters. The proposed method is shown to outperform the previous method in anechoic conditions in terms of both bleed reduction and preservation of the target source. It is also compared to a similar noise cancellation based technique, as well as the blind source separation technique DUET.

Chapter 6 - Overdetermined microphone bleed reduction using selective FDCTRANC

This chapter extends the bleed reduction research in the previous chapter by applying it to the overdetermined case, where there are more microphones than sources. This is done first by performing CTRANC in the frequency domain to improve results in reverberant conditions and reduce the computational cost. In listening tests the frequency domain implementation is shown to outperform a similar noise cancellation method. The proposed method is then extended to the overdetermined case by introducing a selection stage to determine which microphones are reproducing the same target source in order to suppress the bleed reduction algorithm between them. The selection process is shown to provide an improvement in a variety of configurations in terms of interference reduction and preservation of the target source.

Chapter 7 - Microphone bleed simulation in multisampled drum workstations

In this chapter we outline a novel method for simulating bleed between microphones specifically in drum kit recordings where each drum has been recorded separately. This is included as an example of conditions where microphone bleed can enhance an otherwise dry recording to improve the realism. In listening tests, participants are shown to be unable to distinguish the simulated recordings from real recordings with statistical significance.

Chapter 8 - Conclusions and future perspectives

In this chapter we summarise the achievements of the thesis. We explore how the research conducted has achieved the objectives and suggest potential further

work.

1.5 Thesis contributions

The main contributions presented in this thesis are:

Chapter 3

- A method for detecting and correcting the proximity effect in directional microphones without knowledge of the microphone or source to microphone distance.

Chapter 4

- Novel analysis of the GCC-PHAT method of delay estimation with regards to incoming signal bandwidth and DFT window shape.
- A recommendation of best practise when using the GCC-PHAT for arbitrary musical signals, which extends the knowledge of how window shape affects the accuracy of the GCC-PHAT.

Chapter 5

- Adaptation of a method of noise cancellation from telecommunications, not previously applied to musical instrument sources, applied to determined source, microphone configurations by combining CTRANC with centred adaptive filters.
- A novel method for multiple source delay estimation.

Chapter 6

- Extension of determined Crosstalk Resistant Noise Cancellation (CTRANC) to the frequency domain (FDCTRANC) and outlining problems with this method.
- Introducing an iterative method of FDCTRANC.
- Extension of FDCTRANC to the over-determined case, using a selection stage to indicate whether each other microphone is primarily reproducing the same target source or an interfering source for the microphone under test.

Chapter 7

- Novel method of microphone bleed simulation using available audio samples in a multiple microphone drum recording.

1.6 Related publications by the author

The work presented in this thesis has been published and presented in a number of journals and conferences:

1.6.1 Journal Articles

Alice Clifford and Josh Reiss, *Using delay estimation to reduce comb filtering of arbitrary musical sources*, Journal of the Audio Engineering Society, 2013

Alice Clifford and Josh Reiss, *Reducing comb filtering on different musical instruments using time delay estimation* in the Journal of the Art of Record Production, Volume 5, 2010

1.6.2 Conference Papers

Nicholas Jillings, Alice Clifford and Josh Reiss, *Performance optimization of GCC-PHAT for delay and polarity correction under real world conditions* in Proceedings of the 134th Audio Engineering Society Convention, Rome, Italy, 2013

Alice Clifford, Henry Lindsay-Smith and Josh Reiss, *Simulating microphone bleed and tom-tom resonance in multisampled drum workstations* in Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK, 2012

Alice Clifford and Josh Reiss, *Detection of the proximity effect* in Proceedings of the 131st Audio Engineering Society Convention, New York, USA, 2011

Alice Clifford and Josh Reiss, *Microphone Interference Reduction in Live Sound* in Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France, 2011

Alice Clifford and Josh Reiss, *Calculating time delays of multiple active sources in live sound* in Proceedings of the 129th Audio Engineering Society Convention, San Francisco, USA, 2010

1.6.3 Invited Speaker/Panellist

Audio Engineering Society 130th Convention. *Automixing and artificial intelligence*. Workshop panellist. 2011

1.6.4 Magazine articles

Alice Clifford and Paul Curzon *Combing for sound*, Audio!, Issue 2

Chapter 2

Background

In this chapter we present the background to the research presented in this thesis. We explain the purpose and function of a microphone and how it is used in music production and performance. We then discuss how different artefacts on microphone signals are caused, why they are undesirable and why they may need to be removed.

2.1 Microphone technology

Before sound reinforcement, live performance relied on the performer's ability and the acoustics of the performance space to carry the sound from the stage to the audience. After the invention of microphones, amplifiers and loudspeakers, a performer could be amplified to be heard more clearly and by more people in larger, less acoustically adequate spaces.

The first stage of this process is the microphone. A microphone is a transducer that converts sound pressure waves to an electrical current through vibration of a medium. The mechanism for this conversion varies but follows the same basic principle.

The most straightforward of microphones is the dynamic microphone [Eargle, 2004]. Dynamic microphones consist of a diaphragm attached to a magnet. When sound pressure waves travel from the sound source through air to the diaphragm, this causes the diaphragm to vibrate. This in turn moves the magnet within a coil, resulting in electromagnet induction and a varying current output. This is then fed into a microphone pre-amplifier and consequently to an amplifier to be played out of loudspeakers or sent into a sound card to be converted to a digital signal. Dynamic microphones are often used in live sound situations as they are inexpensive, robust and do not require additional power.

Other common microphone designs are condenser and ribbon microphones. In condenser microphones, also called capacitor microphones, the diaphragm acts as one plate of a capacitor. The vibration of the diaphragm changes the

distance between the diaphragm and a plate, which changes the voltage across the plates. Condenser microphones require additional phantom power to function but they are generally more sensitive than dynamic microphones.

Ribbon microphones consist of a thin metal plate which is suspended in a magnetic field. When sound pressure waves move the plate, this movement produces a current. More recently fibre optic and laser microphones have been developed, although these have yet to be widely adopted in music production.

Increasingly, microphones are being sold which are referred to as "digital" microphones. Although referred to as digital, these microphones still require a transducer to convert the sound pressure waves into an electrical signal. Quite often these microphones contain a dedicated Analogue-to-digital (A-D) converter therefore the microphone will have a digital output rather than analogue [Shapton, 2004]. This means that the A-D converter has been moved closer to the transducer. The advantages of this are that it allows the converter to be customised to the specific microphone and can also reduce noise as the distance the electrical analogue signal has to travel is much shorter. Custom DSP can also be used to optimise the bit depth of the conversion or to insert level control to avoid digital clipping [Eargle, 2004]. There is more that can be exploited from the digital microphone and additional processing that could be included which is tailored towards the specific microphone.

Recently digital microphones have become popular with home recordists, for example where looking for an easy way to record vocals for amateur podcasts. Digital microphones aimed at the consumer market have a USB connection which can be plugged straight into a computer to record, therefore removing the need for a dedicated sound card.

As well as the design of the microphone, an important characteristic of a microphone is the directionality. Generally microphones can be grouped into omnidirectional, which picks up sound from all directions, or directional, which rejects sound from certain angles around it. The area around a microphone from where it picks up sound is denoted as the pick up area.

Directionality is achieved by altering the amount of access the sound pressure wave has to the rear of the diaphragm. If the rear of the diaphragm is sealed, the diaphragm only responds to sound pressure waves that arrive to the front. This can be referred to as a pressure microphone as it response to absolute sound pressure at the front of the diaphragm and exhibits an omnidirectional directivity pattern. This means it picks up sound from all directions equally, although this varies with frequency. Omnidirectional microphones are often used for ambient recordings or to record multiple sources at once.

If both the front and rear of the diaphragm are open, the movement of the diaphragm is dependent on the difference in pressure between the front and rear

of the diaphragm and can be referred to as a pressure gradient microphone. A sound pressure wave arriving to the side of the diaphragm will result in an equal pressure at the front and the rear and thus there is zero gradient across it. This means any sounds arriving to the side of the diaphragm will be rejected and will not result in an output from the microphone. Pressure gradient microphones are thus directional.

A pressure gradient microphone which is completely open at the rear rejects sound from 90^o and 270^o angle and accepts sound at 0^o and 180^o equally, where 0^o indicates directly in front of the diaphragm. This is known as a Figure-8, or bidirectional, microphone. Different pick up patterns can be achieved by limiting the access to the rear of the diaphragm through the use of ports. Another common pick up pattern is cardioid, which rejects primarily from the rear and picks up sound predominantly from the front and some to the sides. The shape of the pick up pattern can be changed by changing the configuration of ports at the rear, to achieve hyper cardioid patterns, for example, which have a much narrower directionality.

Directional microphones can be used to improve the signal to noise ratio of a single sound source in a noisy environment by positioning the rejection areas of the microphone towards the noise source and the directional area towards the target source. A consequence of directionality is that a flat response has to be sacrificed due to the proximity effect, characterised by an undesired boost in low frequency energy as a source moves closer to the microphone, beyond what is expected.

Microphones can also be designed to enable switchable polar patterns, and thus the same microphone can be used for either directional or omnidirectional applications. This is common in dual diaphragm condenser microphones where the diaphragms are mounted back to back. A voltage is passed through the rear diaphragm to change its sensitivity, which in turn changes the response of the rear diaphragm to sound pressure waves, and thus also changes the directionality [Eargle, 2004].

2.2 Microphone artefacts

The most straightforward microphone configuration is a single source reproduced by a single microphone in free field or anechoic conditions, i.e. without reverberant surfaces. In ideal conditions this is described as

$$x[n] = \alpha s[n - \tau] \tag{2.1}$$

where x is the microphone signal, s is the sound source, α is change in amplitude due to air absorption as the source travels through air, τ is the delay due to distance and n is the current timestep.

In reality x contains many other sounds and distortions and it is not only a scaled, delayed version of the sound source. Anything other than this can be referred to as a microphone artefact.

We have classified the artefacts that can occur into three categories, which are explained here.

Internal

Internal artefacts refer to artefacts that occur due to the microphone itself. A microphone is not a transparent device. Microphones are physical, analogue devices and each has its impulse response and thus its own characteristics.

Each microphone has its own frequency response, often by design, which is dependent on source to microphone distance and angle. Some microphones are designed to have a very flat response which are often reference microphones which are used for testing other devices so the microphone has to be as transparent as possible. On the other hand, microphones designed for a specific purpose can have a distinctive frequency response that is far from flat. For example, the Shure SM58 has a distinctive peak in the 4kHz range as this microphone is aimed at the live, vocal market [Shure, 2013].

This means that the sound source may sound different when recorded using a microphone than it does in real life. This can be a desired effect and the reason a particular microphone is chosen, or it can be undesired if the choice of microphones are limited or an accurate reproduction of a sound source is required.

Environmental

The environment can cause artefacts which are external to the microphone. This generally refers to reverberation characteristics of the acoustic space and external noise.

Reverberation refers to the composition of reflections of the sound source off nearby surfaces [Howard and Angus, 2000, chap. 6]. This means that if the source and microphone are in a space with reflective surfaces, or any space that is not freefield conditions, then delayed versions of the source will arrive at the microphone after the direct sound and be summed together.

The opposite of a reverberant space is an anechoic space that suppresses room reflections. Anechoic recordings or very dry recordings can sound lifeless and lacking ambiance [Izhaki, 2007, chap. 23], and often on synthesised sounds

reverberation is added to enhance the realism and space in the recording. This is also applied to dry studio recordings. On the other hand if too much reverberation is present either artificially or naturally, the intelligibility of the sound source is reduced and the timbre can be changed by the comb filtering that occurs due to summation of delayed versions of the direct sound source.

Reverberation can be broken down into different parts. Early reflections refer to the first reflections that arrive at the microphone after the direct sound source and are considered to arrive at the microphone up to 80ms after the direct sound [Peters et al., 2011]. Often these reflections have only reflected off a few surfaces and allow us to perceive the size of a space. Early reflections off highly reflective surfaces can be high in amplitude, sometimes nearly equal amplitude to the direct sound, which can cause more extreme comb filtering.

Other environmental factors are external uncorrelated noise in an environment which is not the sound source, such as air conditioning units or in the live sound situation, audience noise.

In a real reverberant environment (2.1) can be extended to

$$x[n] = h[n] * s[n] + v[n] \tag{2.2}$$

where h is the room impulse response between the source and microphone which contains the room reverberation and v is external noise.

Positional

Positional factors refer to artefacts that result from the location and number of microphones and sources. So far we have referred to artefacts assuming a single source and microphone. In reality there may be more.

It is a common recording technique to record a single source with a number of microphones. For example taking stereo recordings of pianos, or recording an acoustic guitar with two microphones to record different aspects of the instrument. The problem with this is often the direct sound will arrive at each microphone at different times. When the microphone signals are mixed together this can cause comb filtering, which causes certain frequencies to be cut whilst others are boosted, changing the frequency composition of the source.

The configuration can also be extended to multiple sources, which is common in a live sound situation where all instruments are on the same stage or in a more "live" band recording where each instrument is in the same acoustic space. In this case, often a single microphone will be employed to reproduce a single microphone, but it likely that each microphone will pick up other interfering sources that are not the target microphone. These interfering sources can be referred to as microphone bleed, spill or crosstalk [Izhaki, 2007, chap. 18].

Extending this further, a multiple source, multiple microphone configuration may also contain single sources reproduced by multiple microphones as well as single sources reproduced by single microphones and ambient microphones to reproduce multiple sources in the space.

2.2.1 Summary

We have explained a number of microphones artefact and causes. Often these artefacts are a nuisance and it is desirable that they are either avoided or removed.

In this thesis we investigate reducing three microphone artefacts: the proximity effect, comb filtering and microphone bleed. These artefacts are particularly problematic in live sound where offline digital audio editing and processing techniques may not be used. Here we outline the background and causes of each artefact and why they are a problem in live sound. A signal model for each artefact is also described.

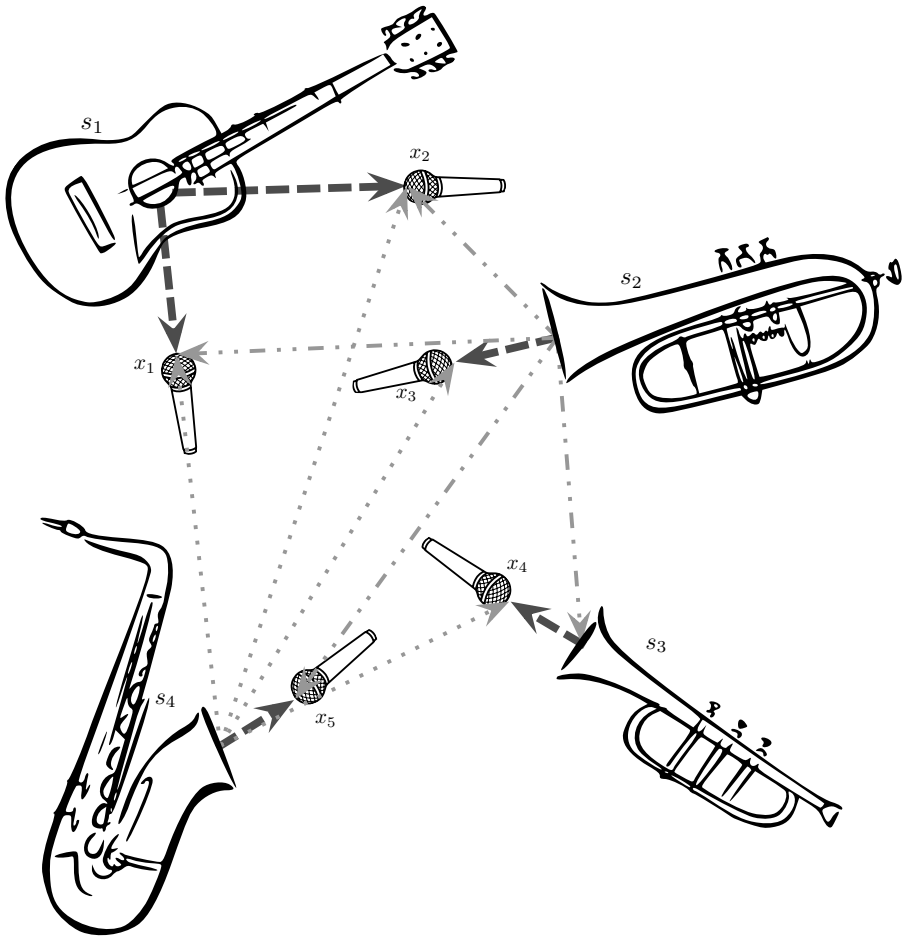


Figure 2.1: Typical configuration of sources and microphones in a live sound production.

2.2.2 General signal model

It is possible to describe all real microphone layouts with a general signal model. Consider an acoustic space with L sources being reproduced by M microphones, for example as depicted in Figure 2.1. The m^{th} microphone signal, x_m, can be described as

$$x_m[n] = \sum_{l=1}^{L} h_{lm}[n] * s_l[n] \tag{2.3}$$

where h_{lm} is the room impulse response (RIR) between source s_l and microphone x_m. Here $m = 1, \ldots, M$, where M is the number of microphones and $l = 1, \ldots, L$ where L is the number of sources. External noise is not included and the impulse response of the microphone is not taken into account. In anechoic conditions, h_{lm} is assumed to be a Dirac delta delayed by τ_{lm} at amplitude α_{lm} so (2.3) can be simplified to

$$x_m[n] = \sum_{l=1}^{L} \alpha_{lm} s_l[n - \tau_{lm}] \tag{2.4}$$

where α_{lm} is the amplitude change primarily due to air absorption between the source and microphone and τ_{lm} is the delay of the sound pressure wave leaving the source and arriving at the microphone at time n.

Different configurations can be described as determined, where $L = M$, underdetermined, where $L > M$, and overdetermined, where $L < M$.

2.2.3 Proximity effect

Even with the simplest microphone configuration described by (2.1) and shown in Figure 2.2, the choice of microphone can cause additional artefacts. It may be the case that this configuration is in a reverberant environment or an environment with external noise. As mentioned previously, a method to reduce this is to use a directional microphone and positioning the sound source in the pick up area and the external noise sources in the rejecting area.

The drawback of this is that all directional microphones exhibit the proximity effect.

The proximity effect is characterised by an artificial boost in the low frequency of the microphone output as the source to microphone distance decreases. The low frequency boost occurs due to the method used to enable directionality in microphones.

It has already been explained that directional microphones are also known as pressure gradient microphones. This is because the movement of the diaphragm which causes an output current is due to the difference in pressure either side

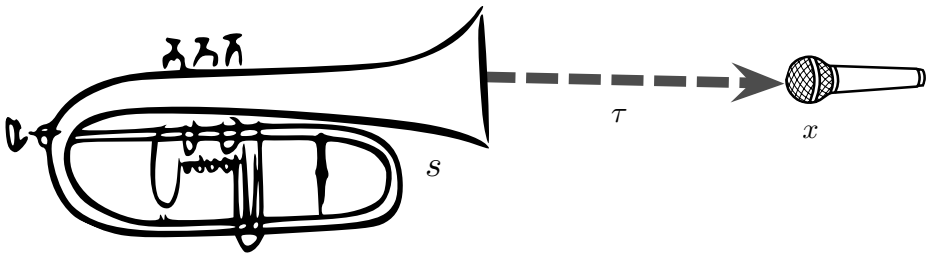


Figure 2.2: A common layout for reproducing a single source s with a single microphone x.

of the diaphragm. The difference in sound pressure is caused by a difference in amplitude of the pressure wave as it travels from one side of the diaphragm to the other. A pressure wave arriving at 0^o will travel the furthest to reach the rear of the diaphragm, therefore will exhibit the largest drop in amplitude and therefore the largest pressure gradient.

The output of a pressure gradient microphone can be considered a ratio between the sound source, which is close to the microphone, and the noise, which is at a further distance, which can be expressed as Signal-to-noise ratio (SNR). A high SNR indicates that the source is close to the microphone and a low SNR indicates it is further away.

A point source is modelled as a spherical wave and the amplitude drop in relation to distance is governed by the inverse square law. At larger distances, the spherical wave can be modelled as a plane wave [Howard and Angus, 2000]. Over the same distance from the same origin, a spherical wave will exhibit a greater drop in amplitude compared to the plane wave.

If the sound source of a microphone is modelled as a spherical wave as it is close to the microphone and the noise is modelled as a plane wave, the amplitude drop of the sound source between the front and rear of the diaphragm will be greater, resulting in a higher pressure gradient and thus a higher perceived amplitude than the noise modelled as a plane wave.

This ratio can be expressed as

$$P_R = \sqrt{1 + \frac{1}{k^2 r^2}} \tag{2.5}$$

where k is the wave number, $k = \frac{\omega}{c}$ and r is the distance from source to microphone [Etter, 2012]. This difference in SNR for different values of r is shown in Figure 2.3.

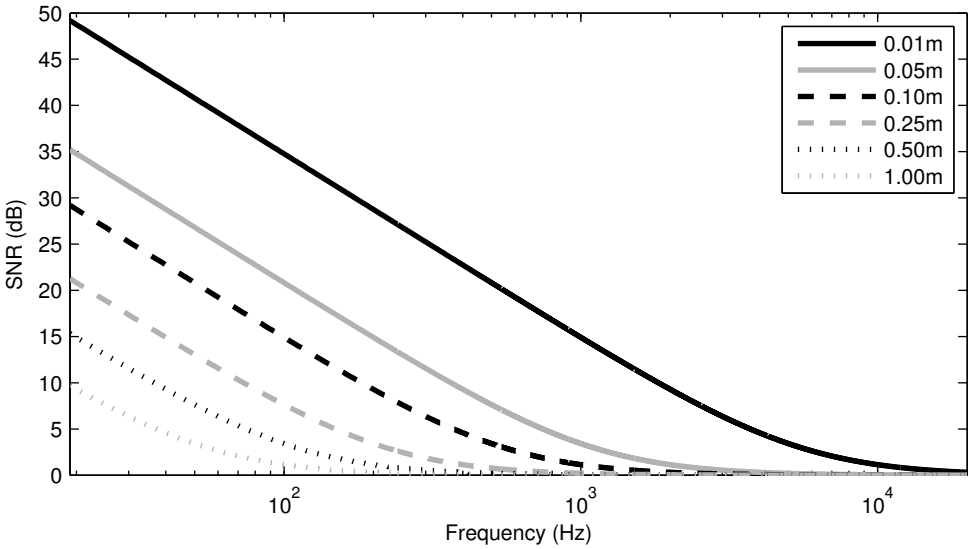


Figure 2.3: Pressure gradient ratio over frequency with changing source to microphone distance.

As frequency increases, the ratio reduces -6dB per octave, eventually reaching 0 as the frequency becomes large. This perceptually results in a boost at low frequencies, as the pressure gradient ratio is generally higher at lower frequencies.

The corner frequency, when the SNR reaches 0, can be calculated from (2.5) as

$$f_c = \frac{c}{2\pi r}. \tag{2.6}$$

Figure 2.4 shows how the corner frequency of the SNR roll off changes with source to microphone distance. The proximity effect occurs because the corner frequency increases as distance decreases.

In a live musical performance, musicians naturally move while performing. This movement changes the source to microphone distance and can therefore cause undesired tonal changes that cannot be corrected using equalisation.

The proximity effect is often considered with vocal performances where the vocalist is holding the microphone in their hand. This means the source to microphone distance changes rapidly and the tone of the microphone output will change.

Although here we consider the proximity effect to be an unwanted artefact,

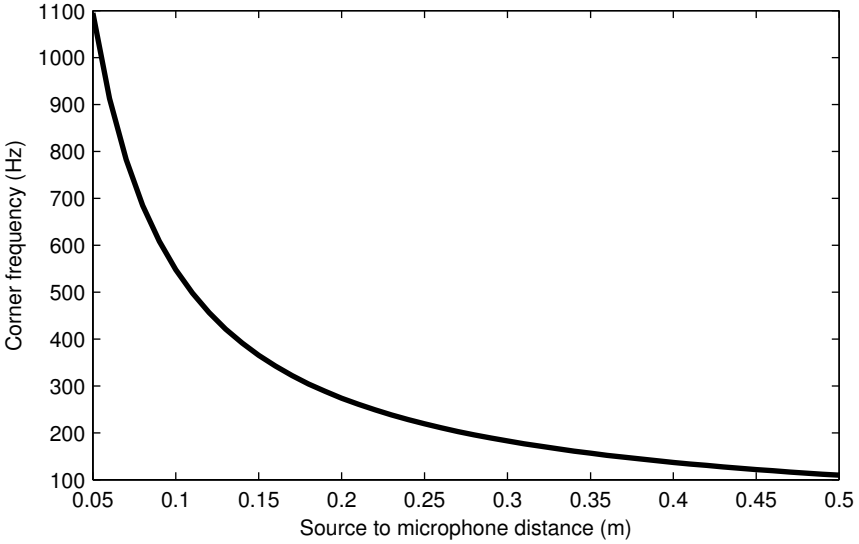


Figure 2.4: Pressure gradient ratio corner frequency with changing source to microphone distance.

there are certain times when it is used as a desired effect, particularly for vocalists. Trained vocalists will be aware of the proximity effect and the effect it has on the tone of their voice. It can be used to enhance low frequency content and produce a boomier, louder and more present sound [Savage, 2011].

2.2.4 Comb filtering

Quite often an instrument will produce a different sound depending on the angle of the listener or microphone. For example, a microphone positioned next to the sound hole of an acoustic guitar will produce a different sound to that at a microphone positioned next to the fingerboard, as in Figure 2.5. Or an engineer may want to reproduce the acoustic space around an instrument with a microphone a further distance from the instrument, but a closer microphone is also required to reproduce more delicate elements of the sound. In these situations, multiple microphones positioned around a single source gives the sound engineer flexibility to mix the microphone signals together in whichever way they desire.

The problem with this is that often the microphones are not equidistant from the sound source. This means that the sound arrives at each microphone at a different time. When the microphones are mixed together, this causes comb filtering.

Comb filtering occurs when any signal is summed with a delayed version of

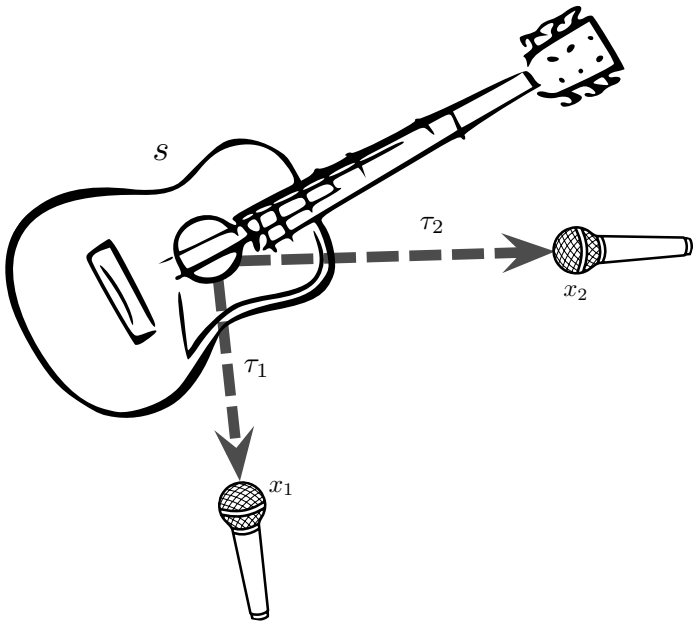


Figure 2.5: A common layout for reproducing a single source s with multiple microphones x_1 and x_2.

itself. In many areas of acoustics, such as sound system design, comb filtering is unwanted [McCarthy, 2006, chap. 2]. But comb filtering can also be a desired effect in the form of flanging or phasing audio effects [Huber and Runstein, 2005, chap. 6].

Comb filtering is so called due to the "comb" shaped frequency response it produces, as seen in Figure 2.6. It is characterised by the peaks and troughs associated with the filter which occur due to the cancellation and reinforcement of frequencies along the audible spectrum.

When a signal is delayed in time, all frequencies are delayed by the same amount. This results in a linear phase shift across the spectrum, causing some frequencies to cancel and others to reinforce. The period of this reinforcement and cancellation is directly related to the amount of delay that is occurring.

Amplitude differences between the microphone signals also changes the frequency response of the resulting comb filter. Equal amplitude will result in complete rejection at the troughs whereas if the delayed signal is of a lower amplitude than the direct signal, the filter will be less severe. Previous research

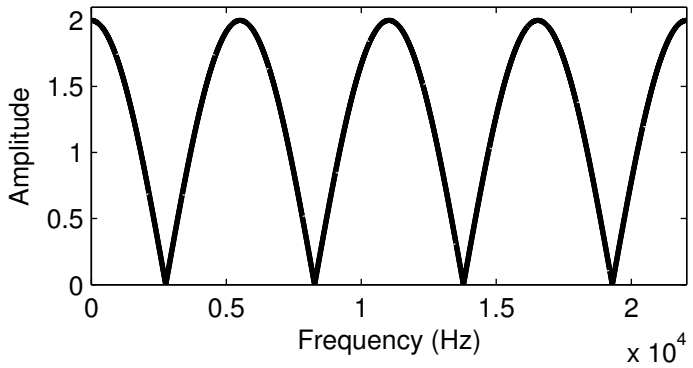


Figure 2.6: Transfer function of a comb filter with a relative delay of 8 samples at 44.1kHz sampling rate.

suggests comb filtering can be heard when the delayed signal is as much as 18dB lower in amplitude than the direct signal [Brunner et al., 2007].

In music production comb filtering can also occur when audio is duplicated, processed and mixed with the original signal, such as recording a guitar both direct and through an amplifier and microphone. Additionally it can occur when stereo recordings are mixed to monaural audio.

Differences in source to microphone delays can also occur when multiple microphones are used to reproduce multiple sources, for example in an ensemble performance where each instrument has a dedicated spot microphone. Microphone bleed can occur between the microphones and can also cause comb filtering if mixed. Similar problems can occur when a stereo microphone pair is used to reproduce an ensemble of instruments and the instruments have their own dedicated microphones. The sound from an instrument will arrive at the spot microphone and the stereo pair with different delays. With a large ensemble, many delays can occur.

Comb filtering due to multiple microphones reproducing the same source is detrimental due to the changes in frequency content that occurs. This can cause the source to sound characteristically "phasey" and often leads to a "thin" sound.

Signal model

A single source, s being reproduced by two microphones x_1 and x_2, as in Figure 2.5, can be described as

$$x_1[n] = \alpha_1 s[n - \tau_1] \tag{2.7}$$

$$x_2[n] = \alpha_2 s[n - \tau_2] \tag{2.8}$$

where n is the current time step, τ_1 and τ_2 are the delays associated with the sound source travelling from the source position to the position of x_1 and x_2 and α_1 and α_2 are associated amplitude changes. Uncorrelated noise and reverberation are not considered. When the microphones are summed to become y, in terms of s this is

$$y[n] = \alpha_1 s[n - \tau_1] + \alpha_1 s[n - \tau_2]. \tag{2.9}$$

It can also be stated that

$$x_2[n] = x_1[n - \tau] \tag{2.10}$$

assuming $\tau_2 > \tau_1$ where $\tau = \tau_2 - \tau_1$.

In the general case this is

$$x_l[n] = \alpha_l s[n - \tau_l] \tag{2.11}$$

where

$$y[n] = \sum_{l=1}^{L} \alpha_l s[n - \tau_l]. \tag{2.12}$$

2.2.5 Microphone bleed

We have discussed single source configurations that can cause the proximity effect and comb filtering. This assumes that there is only one source in a space and that other sources are noise.

In reality, especially in live sound, it is more likely there will be multiple sound sources in a single acoustic space. In this case it is plausible that each sound source has at least one dedicated microphone.

With multiple sources in an acoustic space it is probable that all sources can be heard from all positions. This means that any microphones positioned anywhere in the space will reproduce all sources. The position of each microphone relative to the sources will determine the amplitude of each source in the microphone output. If each source has at least one dedicated microphone, we can assume that each microphone is positioned closest to one sound source and other sources that are reproduced at lower amplitude can be referred to as microphone bleed, as in Figure 2.1.

A microphone reproduces sound that enters the area surrounding it which is described by its pick up pattern. When placing a microphone to reproduce a target sound source, it is placed to ensure the source is within this area. Sound from other sources may also enter this area and will also be reproduced, which can be referred to as interference.

Microphone bleed is a problem because any effects or processing applied to

a microphone signal with the intention of being applied to the target source will also be applied to any interfering sources. This will cause errors in mixing and result in a lower quality production. If the microphone signals with bleed are mixed, this can also cause comb filtering as multiple delayed versions of the same source are being summed. An interfering signal can also reduce the intelligibility of the target source by frequency masking [Howard and Angus, 2000, chap. 5]. It is therefore advantageous to reduce the amplitude or amount of this microphone bleed.

Signal model

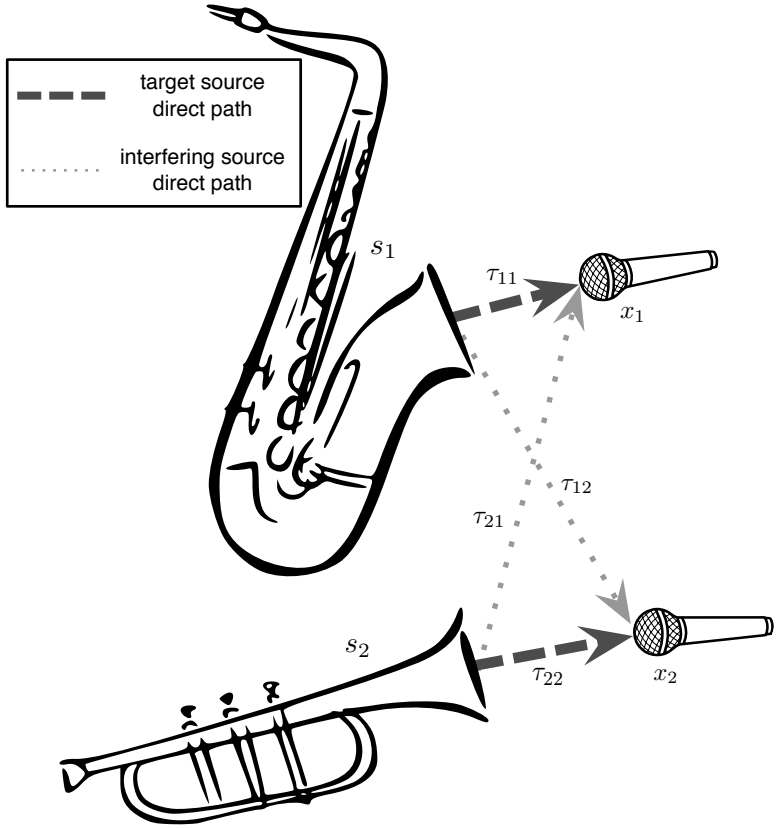


Figure 2.7: A configuration of two sources being reproduced by two microphones with the direct signal paths and equivalent delays shown.

Two microphones, x_1 and x_2, reproducing sources s_1 and s_2, as in Figure 2.7, can be described by

$$x_1[n] = \alpha_{11} s_1[n - \tau_{11}] + \alpha_{21} s_2[n - \tau_{21}] \tag{2.13}$$

$$x_2[n] = \alpha_{12} s_1[n - \tau_{12}] + \alpha_{22} s_2[n - \tau_{22}], \tag{2.14}$$

where τ_{lm} is the delay of source l to microphone m and α_{lm} is the amplitude change of source l to microphone m.

If the microphone signals defined in (2.13) and (2.14) are summed to the output y this becomes

$$y[n] = x_1[n] + x_2[n] \tag{2.15}$$

$$\begin{aligned} &= \alpha_{11}s_1[n-\tau_{11}] + \alpha_{12}s_1[n-\tau_{12}] + \\ &\quad \alpha_{21}s_2[n-\tau_{21}] + \alpha_{22}s_2[n-\tau_{22}] \end{aligned} \tag{2.16}$$

assuming

$$\tau_{11} < \tau_{21} \tag{2.17}$$

$$\tau_{22} < \tau_{12}. \tag{2.18}$$

Equation (2.16) shows that two versions of each source with different delays will be summed, thus causing comb filtering of both sources which is discussed in Section 2.2.4. The relative difference of the delay of each source arriving at each microphone is defined by

$$\tau_1 = \tau_{21} - \tau_{11} \tag{2.19}$$

$$\tau_2 = \tau_{12} - \tau_{22} \tag{2.20}$$

and the relative gain difference as

$$\alpha_1 = \alpha_{21} - \alpha_{11} \tag{2.21}$$

$$\alpha_2 = \alpha_{12} - \alpha_{22}. \tag{2.22}$$

2.3 Strategy

This thesis will be concerned with the following artefacts: the proximity effect, comb filtering and microphone bleed. These artefacts are of particular research interest because they are often encountered by sound engineers and are all caused by microphone positioning.

The following chapters discuss the research that has been undertaken in each area. In each case, a background of each particular subject area is provided, along with commonly used methods for reducing the artefacts. We then outline the literature concerned with reducing each artefact from a digital signal processing point of view and find ways of improving on existing research or conceiving new methods. Each correction algorithm is outlined in detail and then assessed on either simulated data or real recordings, depending on what

is appropriate and suitable, and evaluated either through objective measures, analysis or subjective listening tests. Research into a special case of microphone bleed is also presented which discusses situations where bleed may be desired, such as in simulated drum recordings. In this case we present a method for simulating the microphone bleed. Finally, we propose possible extensions to each method for future research.

Chapter 3

Proximity effect detection and correction

The most basic microphone configuration will consist of a single microphone reproducing a single sound source in an acoustic space. Assuming the positions of the sound source and microphone remain static, artefacts may come from sources external to the configuration, such as reverberation and external noise.

Artefacts can also come from the microphone itself in the form of the proximity effect, which is characterised as a perceptual boost in low frequency amplitude as the source to microphone distance decreases. The main consequence of the proximity effect is unstable frequency content since the low frequencies are boosted as the source to microphone distance decreases and excessive gain which can cause distortion and clipping on the microphone pre-amplifier.

In this chapter we present a method for detecting the proximity effect purely from analysis of the audio signal. We then present a variable gain low shelving filter to correct the low frequency boost.

3.1 State of the art

In Section 2.2.3 we outlined the causes of the proximity effect and how it affects mixing. In this section we discuss current methods and research for detecting and reducing the proximity effect.

In commercial products, the proximity effect is tackled in a number of ways. A class of condenser microphones consist of two diaphragms to provide selectable polar patterns. This can also be used to reduce the proximity effect by effectively enabling a cardioid polar pattern for high frequencies and a non-directional pattern for low frequencies, which will not exhibit the proximity effect [Shure, 2010]. Although this will reduce the amount of low frequency boost the presence of a non-directional capsule even at low frequencies will increase the amount of ambient noise in the microphone signal. The additional components required will also increase the cost of the microphone.

Other microphones include a bass roll off in an attempt to reduce the effect

but this can alter the sound in an undesirable way and remove low frequencies that may not be boosted by the proximity effect. A sound engineer can also apply equalisation (EQ) to the microphone signal to reduce the amplitude or completely cut low frequencies. If the source remains static, this equalisation will successfully reduce the effects of the proximity effect. But if the source to microphone distance changes, the parameters set by the engineer would no longer be valid. A multi band compressor can also be used with the lowest band set to cover the frequency band that the proximity effect tends to occur at, but this varies with each microphone. As with using a filter, sounds that may naturally contain a lot of low frequency information will also be affected.

The published research into the proximity effect is limited. Work by Dooley and Streicher [2003] provides an in depth examination of the technology and use of the bi-directional microphone but there is little explanation of the proximity effect. Torio and Segota [2000] and Torio [1998] model a directional microphone as a combination of a low and high pass first order filters with an overall gain control.

Nikolov and Milanova [2000, 2001] also present a model to describe the proximity effect. Josephson [1999] describes the effect and compares theoretical models to real data and Millot et al. [2007] present results of microphone tests showing the proximity effect.

The proximity effect can be thought of as being three dimensional, in terms of frequency, angle of incidence and distance [Torio, 1998]. Attempts to reduce the proximity effect by sound engineers are limited as they are unable to take into account the absolute distance of the source and microphone and the angle of incidence. If absolute distance data could be found then this could be coupled with microphone data and the proximity effect accurately corrected. A study by Etter [2012] investigates Automatic Gain Control with proximity effect compensation. This method utilises a distance sensor on the microphone. Although this gives accurate distance data, the distance sensor adds additional hardware and therefore cost and inconvenience. Ideally proximity effect correction can be achieved with any microphone as an input.

Methods for calculating source to microphone distance and angle use microphone arrays which require knowledge of the array and at least two microphones [Benesty et al., 2008b]. Work by Georganti et al. [2011] outlines a method to estimate the absolute distance between a single source and a single microphone by using statistical parameters of speech which inform a pattern estimator algorithm. The method is shown to perform for close distances but requires training of the algorithm and is only for speech sources.

Related work on detecting similar artefacts in microphones signals by Elko et al. [2007] attempts to detect and suppress pop noise caused by plosives in

recorded speech and follows a similar framework of detection and correction using knowledge of the physical properties of the artefact.

From this survey of related works it is apparent that the literature on detecting and reducing the proximity effect is limited and there does not exist an adequate solution. Practical solutions exist, but are more akin to removing the offending frequency range instead of attempting to correct the boost in low frequency amplitude. Automatic solutions have been proposed but they rely on accurate source to microphone distance data. We therefore propose a novel method of detecting and correcting for the proximity effect using spectral analysis and dynamic filtering.

3.2 Proximity effect in practice

Detection of the proximity effect first requires understanding and analysis of how it affects microphones under real conditions. Although distance based frequency responses are available for the majority of microphones from the manufacturer the available data can be limited and the manufacturer selects which information they disclose. We have included an analysis of a directional microphone here to show real, unbiased data.

We used a Genelec 8040 loudspeaker to output a white noise signal which was recorded using an omnidirectional reference microphone (DPA 4006) and cardioid condenser microphone (AKG C451) in the Listening Room at Queen Mary, University of London. Although not an anechoic room, carpet was placed under the microphones and loudspeaker to reduce reflections off the floor and the walls were treated with diffusive and absorbent material. Separate recordings were made at distances between 0.01m and 0.3m, each 10 seconds in duration. The microphones were recorded simultaneously and the amplitude of the microphone signals at the furthest distance was the same. The same equipment was used for all experiments described in this chapter.

The microphone recordings were low pass filtered with a 4th order Butterworth filter with a cut off frequency at 500Hz. Figure 3.1 shows the RMS amplitude for the filtered microphone recordings of each distance and microphone type.

At 0.01m there is a 9.38dB difference in amplitude between the two microphones. At 0.3m there is only a 0.95dB difference in amplitude. This higher difference at short source to microphone distance is due to the proximity effect in the cardioid microphone.

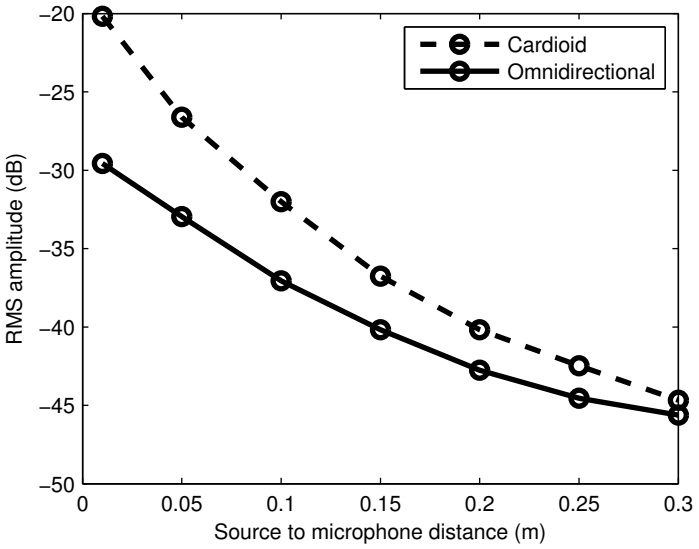


Figure 3.1: Gain low pass filtered white noise recorded with cardioid and omnidirectional microphones at distances between 0.01m and 0.3m.

3.3 Proximity effect detection

The proximity effect must first be detected in a microphone signal before correction can be applied. This is a question of whether the microphone is directional or omnidirectional, but also whether the microphone or the source is moving and if it is moving in a way which is causing the proximity effect to occur, i.e. at a close distance.

The type of microphone could be specified by the user but this relies on user knowledge of the different types of microphone, which some amateur engineers may not have. Some microphones also feature variable polar patterns, therefore knowing the model of the microphone is not an indication of which polar pattern is currently in use. We therefore require an automatic method to detect whether the proximity effect is occurring. The output of the detection should ideally be a binary decision as the proximity effect is determined by whether the microphone is directional or not. Once the proximity effect is detected, this will trigger a correction algorithm which will evaluate how much the proximity effect is affecting the incoming signal.

We want to be able to detect the proximity effect in a microphone without using extra hardware such as distance sensors. We therefore have to achieve detection through analysis of audio features of the microphone signal.

As there is little previous literature on detecting the proximity effect, we have to use knowledge of the properties of the proximity effect to select the

most appropriate features in the microphone signal to use as indicators. We take a heuristic approach in how to analyse the selected features.

However, analysing the low frequency amplitude of a microphone signal cannot be used to detect the proximity effect. This is because there are many occasions where a change in low frequency content is not due to the proximity effect and is due to other causes such as an instrument playing a lower note or the musician playing louder. It is expected that the low frequency amplitude will increase as the source to microphone distance decreases, regardless of the type of microphone being used. The difference with a directional microphone is that the low frequency amplitude will be artificially boosted. A detection algorithm has to be able to take these scenarios into account to avoid false positive results.

In Section 2.2.3 we have shown in Figure 2.4 that the corner frequency of the pressure gradient ratio roll off changes with changing source to microphone distance. At a distance of 5cm the corner frequency is around 1100 Hz which then decreases to around 500Hz at a distance of 10cm. The corner frequency then decreases at a slower rate as distance increases. If we assume the source is moving over time in front of the microphone this corner frequency will be changing within a range, which for a vocalist holding a microwave is likely to be up to 30cm. As we do not know the source to microphone distance, we will generalise that the proximity effect is a boost below 500Hz that has to be rectified.

In this approach no prior knowledge of the microphone or sound source is assumed and only the signal from the microphone is available. The aim of this approach is to detect when the proximity effect is occurring and therefore if the microphone used is directional.

3.3.1 Spectral flux

As the proximity effect is a spectral effect, analysis of spectral features can be used to inform the detection algorithm. A variety of spectral features exist, which are outlined by Lartillot and Toiviainen [2007] and are based on statistical measures of the frequency spectrum.

As we do not have a reference to compare the incoming signal with, if the source is static it is difficult to distinguish whether the proximity effect is occurring or if a boosted low frequency is due to other factors such as additional EQ or the content of the signal. We therefore need to exploit information if the source moves and analyse how the spectrum changes over time.

For this reason, spectral flux is a likely candidate as it is a measure of how data is changing over time, in this case spectral content, and is commonly used in onset detection [Bello et al., 2005]. It is calculated by taking the Euclidean

distance of the magnitude of subsequent frames of data. This is described by

$$\zeta[n] = \sqrt{\sum_{k=0}^{N-1} \left[|X[i,k]| - |X[i-1,k]|\right]^2} \tag{3.1}$$

where X is the microphone signal x in the frequency domain, k is the bin number where $k = 0, \ldots, N-1$, N is the data frame size and i is the current frame.

This is suitable for proximity effect detection because it is assumed that if the source moves and the proximity effect occurs, this will be shown in the spectrum. It is expected that the spectral flux of low frequencies of a signal experiencing the proximity effect would increase more as distance decreases than higher frequencies. This can be used as an indicator of the proximity effect, although we must take steps to ensure natural changes in frequency content of the incoming signal are not mistaken for the proximity effect, which will be detailed in the next section.

The limitations of using spectral flux are that it assumes the incoming signal is at constant amplitude or increasing in amplitude as the distance decreases. If the amplitude of the signal is decaying as distance decreases or the amplitude is constant as the distance decreases at the same speed as the algorithm is running, the spectral flux could remain constant. It is unlikely that either of these would occur but we assume that if it does, another movement event will occur which will trigger the detection algorithm.

3.3.2 Algorithm

The detection algorithm is performed on a frame by frame basis with frames of length N samples. When a new frame is received it is transformed into the frequency domain using the FFT. The frequency bins are then split into j bands of equal width up to 2kHz. Only frequency bins below 2kHz are used as most musical signals contain the majority of frequency energy below 2kHz [Katz, 2007]. We want to avoid analysing spectral content that is not from the target source. The spectral flux for each band ζ_j is then calculated by

$$\zeta_j[i] = \sqrt{\sum_{k=p_j}^{Q_j-1} \left[|X[i,k]| - |X[i-1,k]|\right]^2} \tag{3.2}$$

where Q_j is the maximum bin for the current band j and p_j is the minimum bin. The incoming signal is split into bands to smooth out any increases in amplitude which may be specific to a narrow frequency band due to the recorded instrument playing a lower note or external noise.

In the ideal case of white noise recorded with an omnidirectional microphone the spectral flux will be similar for all bands as all frequencies will exhibit an equal increase in amplitude as the distance decreases. To show this, Figure 3.2 shows the spectral flux over time for an omnidirectional recording of a white noise source. The frame size was $N = 2048$ at 44.1kHz sampling rate and the frequency bins have been split into four bins, each 25 bins in width up to $k = 100$, or to 2.15kHz. The distance between the source and microphone was varied in an oscillating motion over time. As the input is white noise at a constant amplitude output to a microphone at the same angle and position in front of the speaker, any amplitude changes are due to changes in distance. A positive gradient in spectral flux over time indicates the source to microphone distance is decreasing. Equally a negative gradient indicates the distance is increasing. This figure shows that with an omnidirectional microphone, the spectral flux for each band is similar.

Figure 3.3 shows the same experiment with a cardioid microphone. It can be seen that the lowest band exhibits higher spectral flux as the source to microphone distance is at its shortest. The frequency bands above this behave similarly to the omnidirectional microphone.

Therefore if a directional microphone is being used, lower bands will exhibit greater spectral flux over time as the distance decreases due to the proximity effect. This can therefore be used as a measure for detection.

The bands are then split into two sets of low and high frequency bands at 500Hz to encompass all bands which may be affected by the proximity effect. As we mentioned previously, the proximity effect is not uniform for all directional microphones. We then calculate the mean spectral flux for the low and high frequency sets. This is done to smooth out erroneous increases in low frequency amplitude due to other causes than the proximity effect. A large difference between the means will indicate the presence of the proximity effect.

The difference is indicated by Δ_p, where $\Delta_p = \zeta_L - \zeta_H$, ζ_L is the mean low frequency spectral flux and ζ_H is the mean high frequency spectral flux. Once Δ_p crosses a predefined threshold T, the proximity effect is detected. Thus

$$P = \begin{cases} 1 & \text{if } \Delta_p >= T, \\ 0 & \text{if } \Delta_p < T. \end{cases} \tag{3.3}$$

where 1 indicates the detection of the proximity effect and P is the detection function.

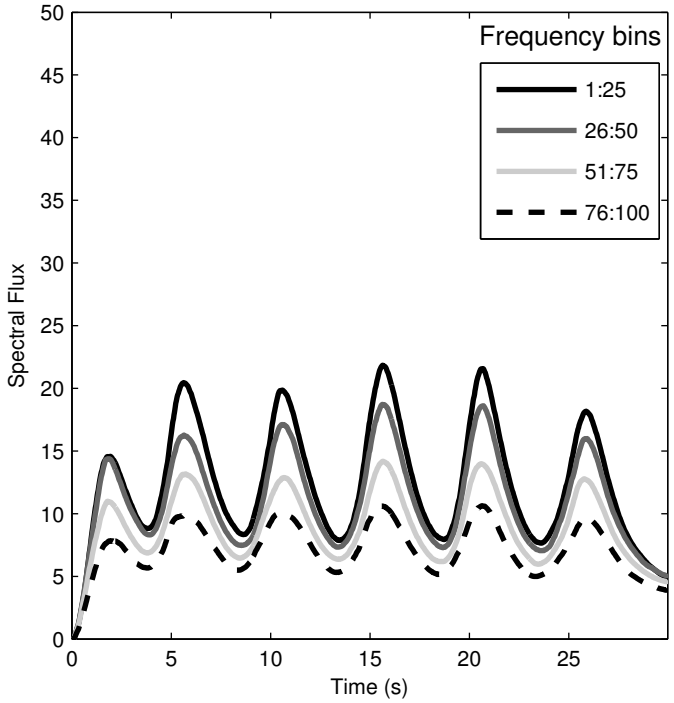


Figure 3.2: Spectral flux of three bands of white noise recorded with an omni-directional microphone with time varying distance.

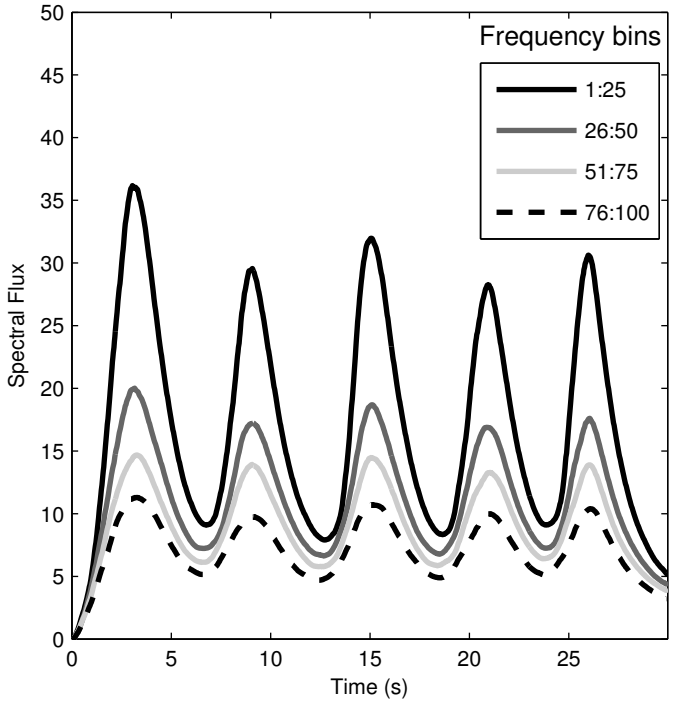


Figure 3.3: Spectral flux of three bands of white noise recorded with a cardioid microphone with time varying distance.

3.3.3 Evaluation

The detection algorithm was tested by recording white noise and a sample of male singing vocal in the same conditions as in Section 3.2. The distance between the source and microphone was periodically changed over time. Each microphone was recorded separately. In the evaluation we calculated the spectral flux in bands 10 bins in width up to $k = 100$, resulting in 10 bands in total.

The aim of the evaluation was to establish whether the algorithm is able to detect the proximity effect in directional microphones when the source to microphone distance of a moving source to a single microphone is short. Ideally we would want to know the exact source to microphone distance. This can be achieved using video analysis or hardware proximity sensors but size and cost limits the flexibility this can have [Etter, 2012]. Instead, we controlled all parameters to ensure that the only amplitude changes were due to source to microphone distance changes. Under these conditions an overall increase in amplitude is only attributed to a decrease in source to microphone distance.

Figures 3.4 - 3.7 show the output of the detection algorithm for a white noise and male vocal input source. The detector outputs 1 when the proximity effect is detected and 0 if it is not detected. The RMS level of the input signal is shown in each case to give an indication of the source to microphone distance. Any amplitude changes are attributed to the increase in amplitude as the source to microphone distance decreases as the microphone was moved in an oscillating motion in front of the loudspeaker. The maximum distance was approximately 0.5m and the minimum approximately 0.01m.

Figure 3.4 shows the output of the proximity effect detector using the omnidirectional microphone recording with a white noise source and Figure 3.5 shows the same for the cardioid microphone. The proximity effect was not detected in the omnidirectional recording, which is expected. The proximity effect on the cardioid microphone recording was accurately detected each time the source to microphone distance decreases.

Figures 3.6 and 3.7 show the proximity effect detection output for a male vocal source with an omnidirectional and cardioid microphone respectively. The algorithm successfully detected when the source to microphone distance decreased and caused the proximity effect in the cardioid microphone case. The proximity effect was not detected in the omnidirectional microphone case.

Although the proximity effect detection output is shown here varying over time, in reality it is a binary decision and if the proximity effect is detected at all this means that the microphone is directional and is exhibiting the proximity effect. We can then assume that if the proximity effect is detected at any point

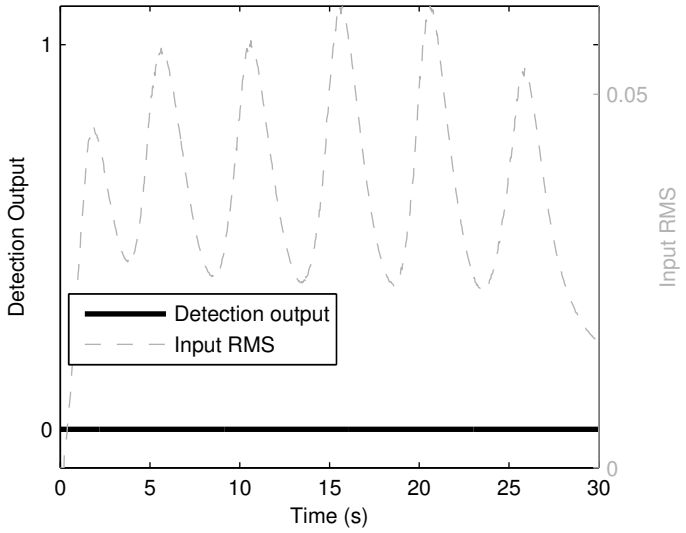


Figure 3.4: Proximity effect detection of a white noise signal recorded with an omnidirectional microphone.

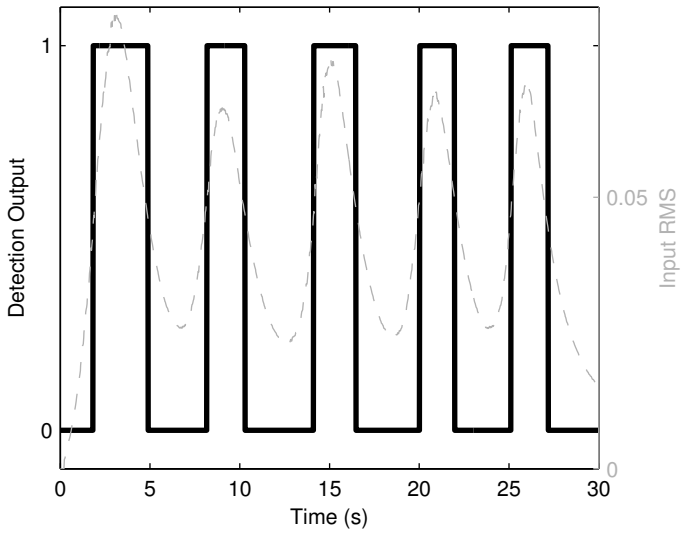


Figure 3.5: Proximity effect detection of a white noise signal recorded with a cardioid microphone.

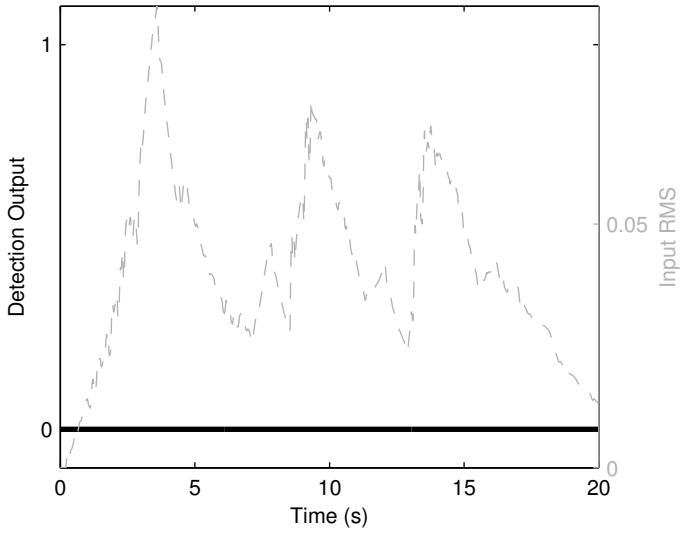


Figure 3.6: Proximity effect detection of a male vocal source recorded with an omnidirectional microphone.

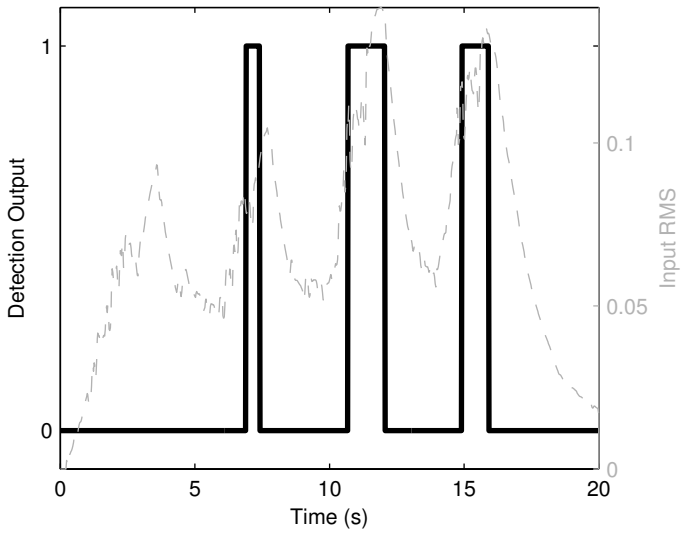


Figure 3.7: Proximity effect detection of a male vocal source recorded with a cardioid microphone.

in the audio sample, a correction algorithm should be triggered.

3.4 Proximity effect correction

Once the proximity effect is detected, correction is then required. This section outlines a method for correcting for the proximity effect through analysis of the incoming microphone signal.

As mentioned previously it is possible to use a multiband compressor to smooth out the proximity effect. The problem with this is that the parameters of the compressor are static and are usually set by the sound engineer at a fixed source to microphone distance during a sound check of a live performance. Therefore if the amount of movement or position changes, or if a different instrument uses that microphone, then the parameters will no longer be relevant. The parameters also need to be set by a trained sound engineer.

We therefore propose using a dynamic shelving filter with an adaptive gain based on analysis of the incoming audio which will allow the level dependence of the multiband compressor but the isolated low frequency equalisation of a static filter.

In a live sound situation the sound engineer will apply gain and EQ for the source signal at a fixed distance to get the desired sound. If we assume that this is the mean distance between the source and microphone throughout a performance, we can use this as a baseline to aim the correction towards. The goal is therefore to match the ratio between the high frequencies and low frequencies when the source to microphone distance decreases to that at the mean distance. Doing this will keep the tone of the sound source stable.

The method is performed on a frame by frame basis. The incoming microphone signal x is first transformed to the frequency domain using the FFT of size N to become X. The frequency bins are then split into two frequency bands at the cutoff point f_c in Hz as a bin number k_B, calculated by $f_C/(f_S N)$ where f_S is the sampling frequency.

The mean amplitude of each frequency band is then calculated by

$$\bar{X_L} = \frac{1}{k_B + 1} \sum_{k=0}^{k_B} |X_L[k]| \tag{3.4}$$

$$\bar{X_H} = \frac{1}{N - k_B - 1} \sum_{k=k_B+1}^{N-1} |X_H[k]|. \tag{3.5}$$

therefore when $x[n]$ is white noise, $\bar{X_L} = \bar{X_H}$.

The mean amplitude that we aim the correction towards is estimated by taking an accumulative average of the low frequency bins of the incoming signal,

X_L, up to the current time. This becomes the threshold of the dynamic filter, R.

The dynamic filter we employ is a low cut shelving filter with a cutoff point f_c equal to the crossover bandwidth point, in this case chosen as 500Hz. The gain G of the shelving filter is calculated using the ratio of R to the mean amplitude of the low frequency bins of the current frame of data, $\bar{X_L}$, described by

$$G = \begin{cases} -20\log_{10}\left(\frac{\bar{X_L}}{R}\right) & \text{if } \bar{X_L} > R, \\ 0 & \text{if } \bar{X_L} <= R. \end{cases} \tag{3.6}$$

So if the mean low frequency amplitude is less than the threshold, the filter is not applied. The filter equations are taken from [Zölzer, 2002, chap. 2] and the difference equations are defined by

$$y_1[n] = a_{B/C}x[n] + x[n-1] - a_{B/C}y_1[n-1] \tag{3.7}$$

$$y[n] = \frac{H_0}{2}\left[x[n] \pm y_1[n]\right] + x[n]. \tag{3.8}$$

The gain G in dB is adjusted by

$$H_0 = V_0 - 1, \text{ with } V_0 = 10^{G/20} \tag{3.9}$$

and the variable for cut frequency a_B for boost and a_C for cut are calculated by

$$a_B = \frac{tan((\pi f_c/f_s) - 1)}{tan((\pi f_c/f_s) + 1)} \tag{3.10}$$

$$a_C = \frac{tan((\pi f_c/f_s) - V_0)}{tan((\pi f_c/f_s) + V_0)}. \tag{3.11}$$

So once the low frequency amplitude goes above the cumulative mean low frequency amplitude, gain reduction takes places which is related to how far the low frequency amplitude of the current frame is above the mean. The processing can be applied separately to the analysis in a side chain approach.

3.4.1 Evaluation

There does not exist a precedent for evaluating a proximity effect correction algorithm, nor is there a standard metric for measuring the "amount" of proximity effect. In microphone specifications the proximity effect is shown described by showing the frequency response of the microphones at different distances and angles. This is also repeated in the literature [Olson, 1991].

We will therefore show in this evaluation that the proposed algorithm is

performing what we set out to achieve through an analysis of the processed audio.

As with the evaluation of the detection algorithm, ideally we want to evaluate the algorithm on audio that is recorded using a directional microphone where we have absolute control over the distance but this comes with problems with measurement, as mentioned previously. For the analysis of the correction algorithm it is important to be able to determine the precise distance to show how much the proximity effect is having an effect. We therefore simulated audio recorded with a directional microphone using models described in [Torio, 1998] to simulate the proximity effect using distance as the input variable. In this implementation we modelled the filters using first order Butterworth filters. A model of an omnidirectional microphone is not used because we assume the proximity effect has already been correctly detected and therefore the microphone is directional.

White noise

The correction algorithm was evaluated using white noise and a 20 second male vocal sample as input sources, typical of the type of signal which will often exhibit the proximity effect. A framesize of 2048 samples was used with a sampling rate of 44.1kHz.

Different types of time varying movement were analysed to establish how the algorithm handles different situations. These can be seen in Figure 3.8 showing time against source to microphone distance.

Figures 3.9 to 3.14 show the results of the correction algorithm using a white noise input source. The low frequency amplitude before and after correction and the threshold R are shown as a function of time. Due to convergence of the accumulative averaging, only the last 10 seconds of the audio sample is shown.

Figure 3.9 shows the low frequency amplitude before and after correction for the first movement vector where $\bar{Y_L}$ is the mean low frequency amplitude after correction. The source to microphone distance was kept static at 0.01m what was the most extreme example. Ultimately no correction occurred as the microphone was not moving therefore $\bar{Y_L} = \bar{X_L}$ and R remained at the same level. In this case the sound engineer would have already corrected for the proximity effect manually as the source is static.

Figure 3.10 shows the same movement as Figure 3.9 but at 0.5m. The results are the same.

Figure 3.11 shows the source to microphone distance slowly decreasing in a cosine movement. As the distance goes below 0.2m the correction began to reduce the level of low frequencies to the mean level. This shows the method

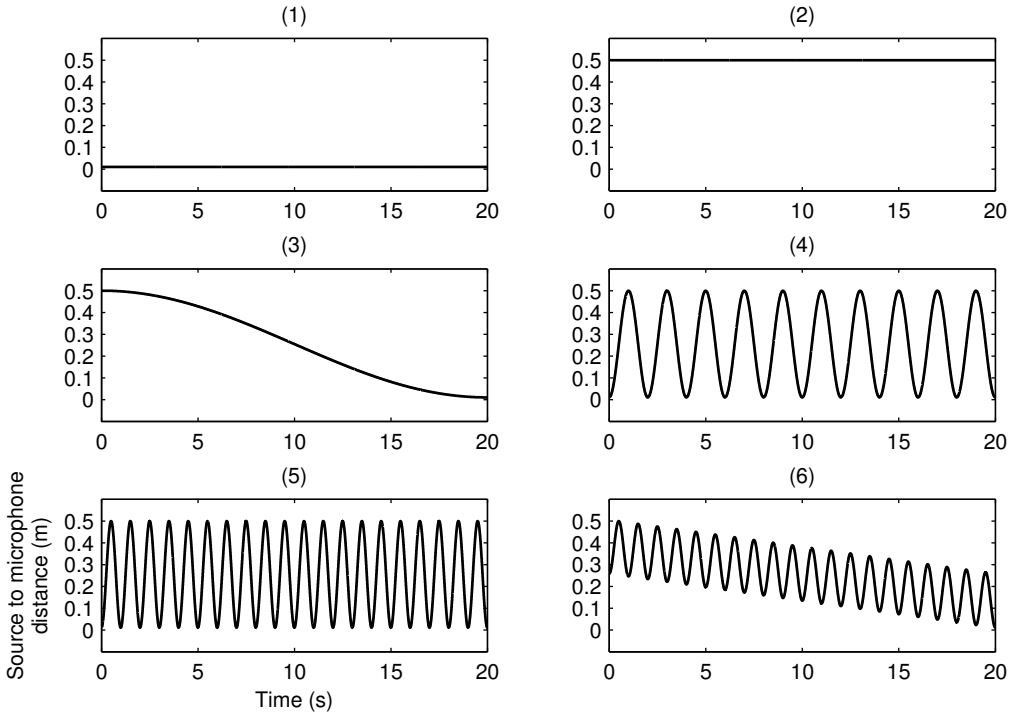


Figure 3.8: Movement vectors tested.

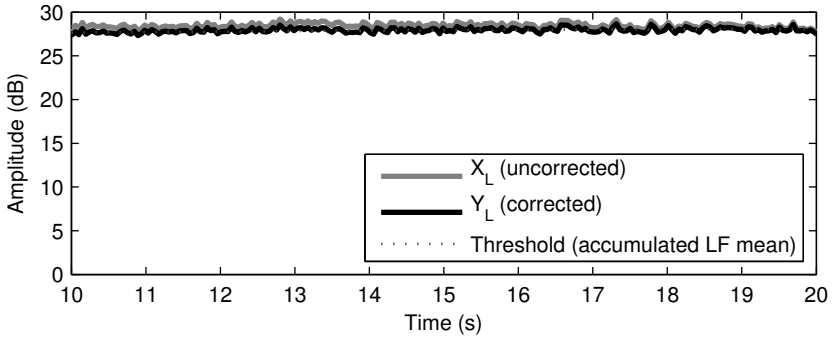


Figure 3.9: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(1) with white noise source.

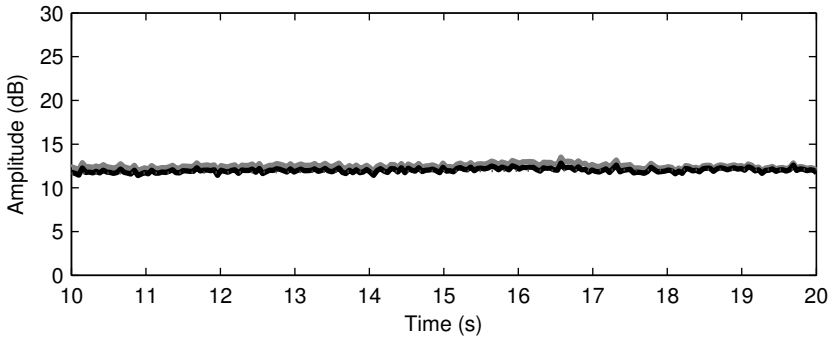


Figure 3.10: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(2) with white noise source.

was successfully reducing the proximity effect.

Figures 3.12 and 3.13 show sinusoidal movement at different frequencies. Due to the sinusoidal movement the expected low frequency amplitude, the cumulative average, was stable and the low frequency amplitude was successfully reduced towards this. The amount of reduction could be increased by adjusting the filter, but high levels of reduction will exhibit similar artefacts as over compression.

Figure 3.14 shows a more complex movement with a sinusoidal movement which gradually decreases the minimum and maximum distances. This is included to show the case if the average movement may change slowly over time. As the source to microphone distance is decreased, more reduction occurs.

We further analysed the data by calculating the Euclidean distance between the uncorrected and corrected low frequency amplitude and the mean. Figure 3.15 shows the results for each movement vector. This shows that the correction algorithm succeeded in the task of reducing the amplitude of the low frequencies towards the accumulated mean level.

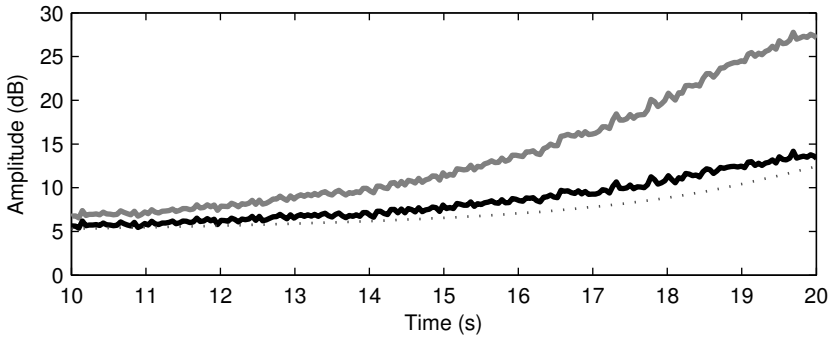


Figure 3.11: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(3) with white noise source.

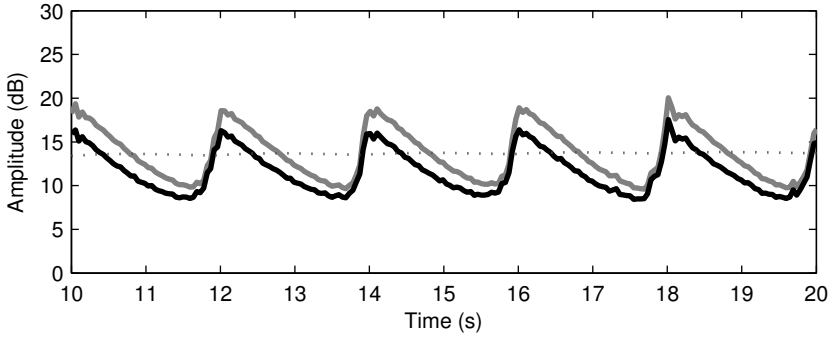


Figure 3.12: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(4) with white noise source.

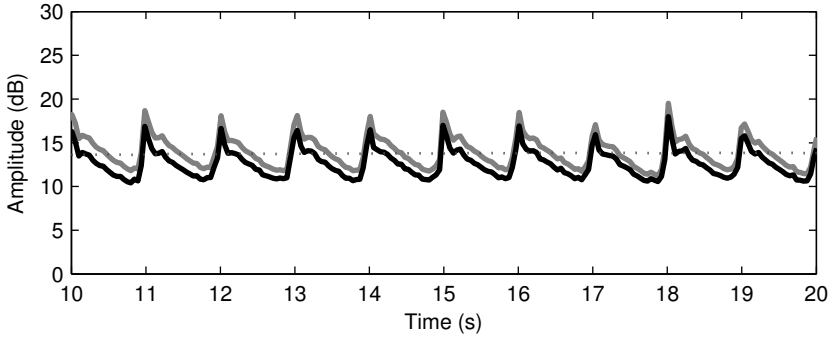


Figure 3.13: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(5) with white noise source.

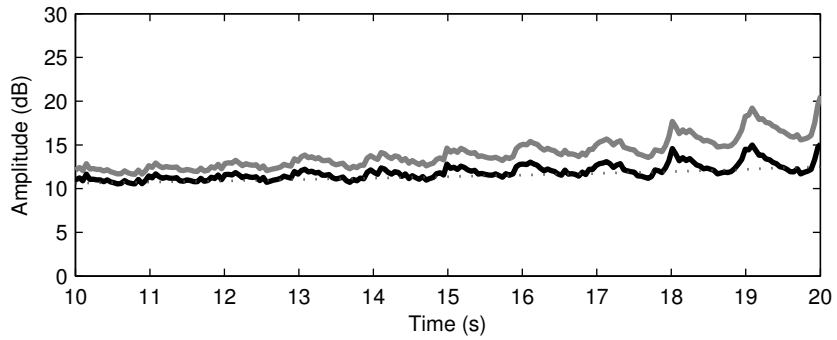


Figure 3.14: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(6) with white noise source.

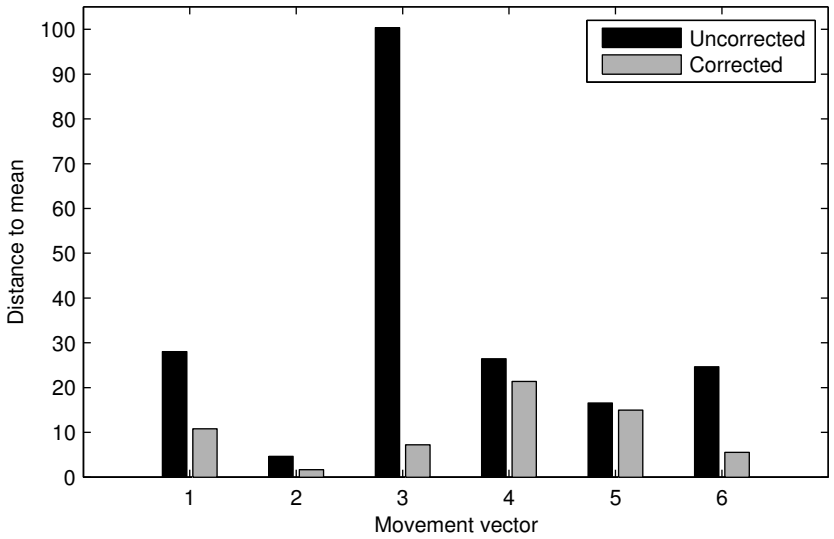


Figure 3.15: Euclidean distance to mean of the uncorrected and corrected low frequency amplitude for each movement vector from Figure 3.8 for a white noise source.

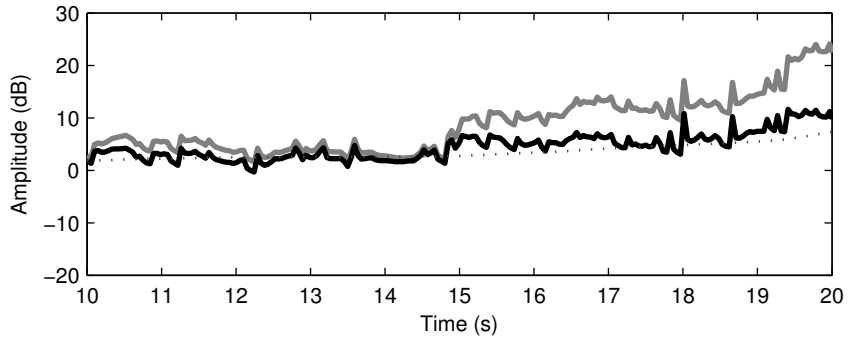


Figure 3.16: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(3) with male vocal input.

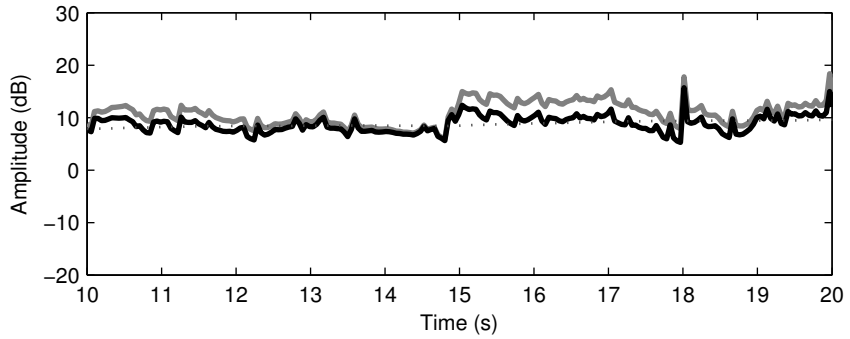


Figure 3.17: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(6) with male vocal input.

Male vocal

We now analyse the same types of movement with a male vocal source. Here we present the most interesting results. All results that are not included in the next section can be found in Appendix A, Figures A.1 to A.4.

Figure 3.16 shows the analysis for the vocal input signal with the source to microphone distance slowly increasing. As with the white noise, the reduction increased dramatically as the source to microphone distance decreased towards 0.01m. The effect of a melodic source can also be seen, since there were localised increases in low frequency amplitude due to lower notes being sung. On occasion, these rises in low frequency energy were enough to trigger the correction.

Figure 3.17 shows the sinusoidal movement gradually moving towards the source. Again, the amount of reduction increased as the source to microphone distance decreased. The results were less dramatic than the white noise case due to the changing melodic nature of the input signal.

Figure 3.18 shows the Euclidean distance between the uncorrected and cor-

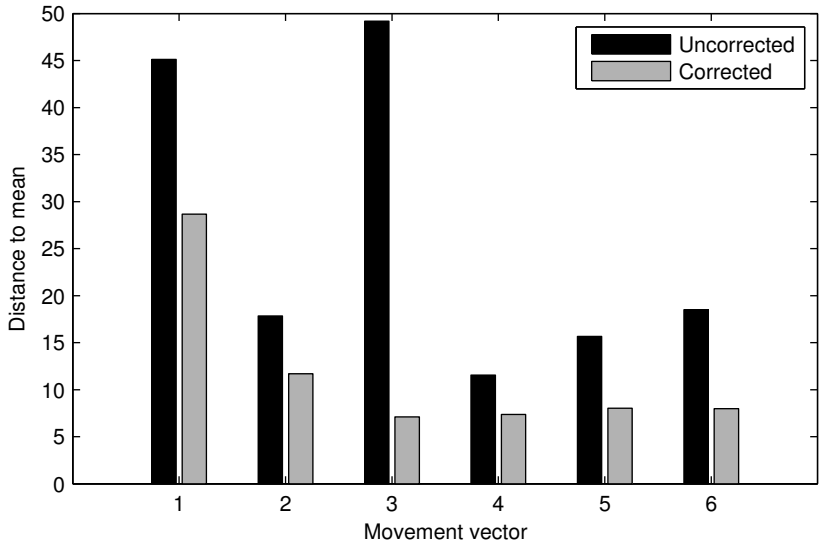


Figure 3.18: Euclidean distance to mean of the uncorrected and corrected low frequency amplitude for each movement vector from Figure 3.8 for a male vocal source.

rected low frequency amplitude and the mean for all movement vectors. As with the noise input, the results showed that in all cases under test the algorithm succeeded in correcting the low frequency amplitude towards to the mean.

We have shown that the proximity effect correction is successful at bringing down the amplitude of the low frequencies as they increased due to the proximity effect with a white noise and male vocal source. We have also shown that the method adapted to different circumstances.

3.5 Discussion and conclusions

In this chapter we have presented methods for detection and correction of the low frequency boost caused by the proximity effect in directional microphones without knowledge of the microphone or source to microphone distance. This has not been attempted in the literature.

We detect the proximity effect by employing spectral analysis to extract spectral flux. Analysis of spectral flux then determines whether the proximity effect is occurring, because spectral flux will be higher at lower frequencies as source to microphone distance decreases in directional microphones. The method was shown to accurately detect the proximity effect on recordings made with a directional microphone and unable to detect the proximity effect in recordings made with an omnidirectional microphone.

The proximity effect is then corrected by analysis of the microphone signal. The correction method is a dynamic low shelving filter with gain dependent on the analysis of the incoming audio. The filter intelligently reduces the low frequency boost to a level at the mean distance between source and microphone without prior knowledge of the microphone or initial source and microphone positions.

The correction method was shown to successfully reduce the boost in low frequency energy on a variety of movement vectors.

The work has potential to be used in live sound scenarios to retain spectral consistency when a musician naturally moves in front of the microphone while performing. It also has applications in teleconference situations to avoid erratic increases in amplitude that can cause signal distortion due a speaker suddenly moving close to the microphone. In this case previous research into speech to microphone distance estimation could be utilised to improve results.

In this chapter we have discussed an artefact that can occur when a single source is reproduced by a single microphone. In the next chapter we extend this to the case where multiple microphones reproduce a single source and investigate reducing the comb filtering that this can cause.

Chapter 4

Comb filter reduction

In the previous chapter we discussed the proximity effect, which occurs when using directional microphones and can be an unexpected problem in a configuration of a single source being reproduced by a single microphone.

Continuing with a single source, it is possible to reproduce a single source with multiple microphones. The problem with this is that often the microphones are not equidistant from the sound source. If the microphones signals are mixed then multiple, delayed versions of the same sound source are summed. This can result in comb filtering which changes the timbre of the sound source and can often lead to it sounding "thin".

In this chapter we present research into reducing comb filtering by automatically estimating the relative delay of a source to multiple microphones. We discuss how the performance of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) method of time delay estimation is dependent on the bandwidth of the input source and on the window function used.

4.1 State of the art

An introduction to the causes and effect of comb filtering in live sound has already been provided in Section 2.2.4. In this section we discuss the state of the art in comb filter reduction from the literature.

4.1.1 Reducing comb filtering

Since comb filtering is caused by a difference in time of a sound source arriving at multiple microphones, the immediate goal to reduce the comb filtering is to align the source in each microphone.

This can be achieved by physically positioning the microphones equidistant from the source but this requires accurate measurement and it may not always be the desired configuration.

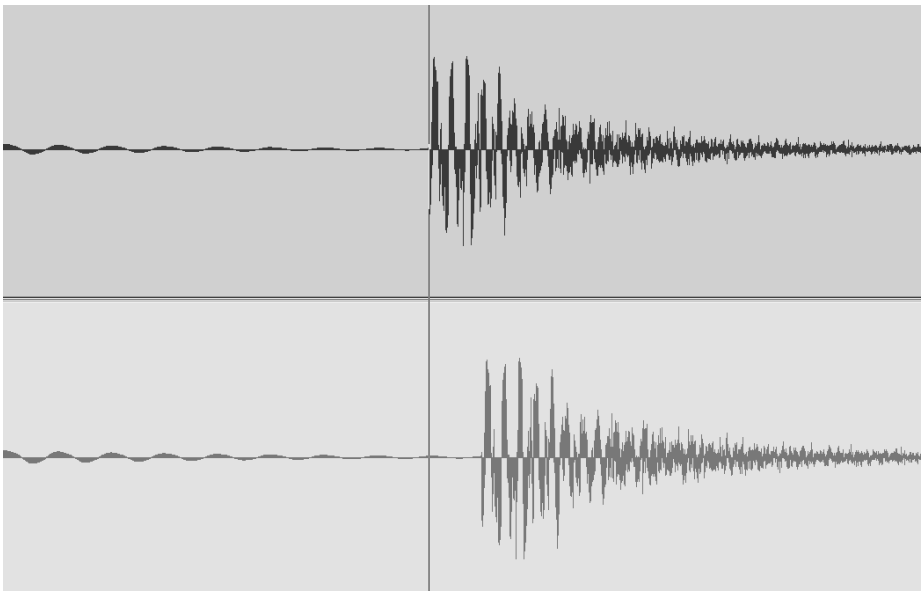


Figure 4.1: Simulated waveforms of two microphones picking up the same sound source. In live sound the top waveform would be delayed to align with the bottom. In post production the waveform regions can be shifted manually.

It is possible in live sound to apply delays to microphone signals that are reproducing the same source [Rumsey and Mccormick, 2005, chap. 13]. The correct delay to apply to each microphone can be calculated by measuring the positions of the source and microphones. It is also possible to apply delay by ear until the comb filtering has been audibly reduced, but this can become difficult when many microphones are used and is unlikely to be sample accurate. As this is a real time situation, delay is usually applied so that all the audio tracks are aligned with the microphone signal with the longest delay, but this requires knowing which microphone this is. In studio recordings it is also possible to manually move audio regions in a Digital Audio Workstation (DAW) to visually align audio tracks, as shown in Figure 4.1. Studies by Leonard [1993] and Anazawa et al. [1987] have shown that improvements to audio quality are achieved when delay compensation techniques are used.

The problem with manually estimating a compensating delay is that it is unlikely to be accurate. One sample delay is enough to result in a first order low pass filter. Assuming a sampling frequency of 44.1kHz and a speed of sound of 344 m/s at room temperature this is equivalent to a difference in source to microphone distance between two microphones of just 0.0078m. Therefore sample-accurate manual delay correction is almost impossible.

Adjusting delays by ear means that the comb filtering may appear to be reduced for the current piece of audio but if the audio changes, for example if

an instrument plays a different range of notes, the comb filtering could reappear in a different frequency range. Estimating delays by measuring distances has its own problems as the speed of sound is not constant and can easily be changed by temperature and humidity [Howard and Angus, 2000]. In both cases if the source moves, the delays will change and comb filtering will once again occur.

A number of studies exist in the literature concerned with automatically estimating the delay in microphone recordings of musical sources and applying the delay to reduce comb filtering. Work by Perez Gonzalez and Reiss [2008c] emulates the manual delay correction usually used to reduce comb filtering by estimating the delay between microphones using methods from sound system alignment [Meyer, 1992]. More recently commercial products have begun to emerge that claim to achieve automatic alignment, presumably using similar methods.

Other literature on reducing comb filtering in multiple microphone configurations includes work by Faller and Erne [2005] who propose a method aimed at live classical concerts where spot microphones are used to pick up individual instruments and a stereo pair used to reproduce the overall sound of the orchestra. Delay occurs between the sound of the instrument arriving at the spot microphone and at the stereo microphones and also there is a difference in the timbre due to reverberation that occurs on the stereo microphones but not on the spot microphones. When the spot microphones are mixed with the stereo microphones, this may result in an unnatural sound which is generally undesired in a classical recording. The impulse response between the spot microphone and the left and right stereo microphones is estimated and the spot microphone filtered with this impulse response. This method does not attempt to estimate the delay directly, but instead relies on the impulse response to introduce the delay. This method is also not used solely for comb filtering, but for the overall sound of the instrument, including attenuation and reverberation.

A study by Gnann and Spiertz [2008] proposes a method for mixing signals in the frequency domain to avoid comb filtering. This requires some estimation of the phase spectrum of the output signal, which can prove problematic, and it was not tested under noisy or reverberant conditions.

It is also possible to use decorrelation to reduce comb filtering of correlated source [Kendall, 1995] but this involves direct processing of the microphone signals that may produce artefacts.

4.1.2 Delay Estimation

We mentioned previously that it is possible to automatically estimate the delay between microphones for use in comb filter reduction. This is commonly known

as Time Delay Estimation (TDE) or Time Difference of Arrival (TDOA) and performs with no prior knowledge of the source or microphone positions. Most previous work utilises TDE for source localisation using multilateration, for use in source separation and for microphone array beamforming [Benesty et al., 2007]. In this section we present a literature survey of the common methods of delay estimation.

Huang et al. [2006] outline the challenges in the identification of MIMO (multiple input, multiple output) systems, which includes delay estimation. It states the main challenges to TDE are blind channel estimation and reverberation. This needs to be taken into account when considering methods for delay estimation. There is a wide body of literature on comparing delay estimation methods in telecommunications and a comprehensive evaluation can be found in [Chen et al., 2006].

The fundamental method of estimating the time lag between correlated signals is to perform the cross correlation between them. Recent studies still make use of this, for example work by Tamin and Ghani [2003] proposes optimising the cross correlation function to improve accuracy of TDE, suggesting that a combination of a Hilbert Transform with a pruned cross correlation function produces the greatest improvement.

The cross correlation was extended by Knapp and Carter [1976], where the Generalized Cross Correlation (GCC) was introduced. GCC performs the cross correlation in the frequency domain using the FFT. This is then transformed back to the time domain and the delay is estimated by finding the position of the maximum peak in the histogram. This is equivalent to estimating the impulse response between the microphone signals. It is sample accurate and is favoured since it is computationally cheap, straightforward to implement and allows tracking of moving sources [Benesty et al., 2008b].

Weightings can also be applied to improve the performance of the GCC in noisy and reverberant conditions. An example of this is the Phase Transform (PHAT), which has mostly been applied to speech [Benesty et al., 2008a].

Other methods of delay estimation also attempt to estimate the impulse response between the microphone signals by adaptive filtering, for example Least Mean Square (LMS) [Reed et al., 1981] and the Adaptive Eigenvalue Decomposition Algorithm (AEDA) proposed by Benesty [2000] and recently extended by Salvati and Canazza [2013]. Adaptive filtering techniques tend to require a period of convergence and the time based implementations can cause computational issues when used at high sampling rates, such as the full audio bandwidth used in music recordings as opposed to speech transmission. LMS-based methods also require knowledge of which microphone signal incites the longest delay, as adaptive filters are commonly used for echo cancellation or noise cancellation

where the configuration is known. The GCC, on the other hand, is able to manage negative delays. Adaptive filter techniques will also take time to converge to a new value in the delays changes. Therefore if the sources are moving quickly, this will not be accurately tracked by the adaptive filter.

The Degenerate Unmixing Estimation Technique (DUET) method of source separation of mixed signals [Yilmaz and Rickard, 2004] also calculates the delay parameters by estimating the phase difference for each frequency bin and performing a histogram on the result. An estimate of the amplitude of each bin is also included to produce peaks in the histogram. The position of these peaks determines the attenuation and delay of each source and the number of peaks is equal to the number of sources. Unlike most source separation methods, this does not use GCC for the delay estimation but it is able to estimate delays of multiple sources.

Work by Meyer [1992] also suggests calculating the impulse response between the microphone signals and Perez Gonzalez and Reiss [2008c] extend this by applying the Phase Transform to the impulse response and calculating the position of the maximum peak to estimate the delay. This method is used in the audio analysis and system alignment software SIM II [Meyer Sound, 1993] and is aimed at a variety of input signals, including musical instruments. The methods proposed by Meyer [1992] are equivalent to methods outlined by Knapp and Carter [1976] but different naming conventions are used. For example an undefined step in the calculation of the impulse response by Meyer [1992] is named the Roth processor (ROTH) weighting by Knapp and Carter [1976].

The review paper on delay estimation by Chen et al. [2006] compares the most popular methods of delay estimation which we have outlined: LMS, AEDA and GCC-PHAT. It concludes that the method previously proposed by the same author [Benesty, 2000], AEDA, is most robust to reverberation but at higher computational cost than the more common methods, such as the GCC-PHAT.

Other studies support this, such as work by Brutti et al. [2008] which compares the GCC-PHAT method to the AEDA specifically using the TDE to estimate source locations. It concludes that the GCC-PHAT method is more accurate under noisy conditions and that the AEDA is more computationally complex.

From this literature survey it is clear that the GCC-PHAT is the most appropriate delay estimation method for realtime comb filter reduction of musical sources, which we will use for the remainder of the chapter.

4.1.3 GCC-PHAT

In this section we provide a more in depth survey of the literature specifically concerned with the GCC-PHAT.

An accurate and stable estimation of delay is imperative to reduce errors in the subsequent usage of the estimation. This is important when used for comb filter reduction as sudden changes in the estimated delay produce audible artefacts.

It is well known that the GCC is susceptible to uncorrelated noise and reverberation which can reduce the accuracy of the estimation and how to improve the robustness of the method is an open problem [Chen et al., 2006]. Chen et al. [2005] present a method for improving the performance of the GCC technique by weighting the calculation, which is found to perform well in noisy environments. Champagne et al. [1996] present an investigation into using a maximum likelihood estimator with the GCC in reverberant environments.

There are a variety of weighting functions suggested in the literature, including Smooth Coherence Transform (SCOT) and ROTH in the original study by Knapp and Carter [1976]. The most commonly used is the Phase Transform, which has been shown to improve performance in noisy and reverberant conditions [Chen et al., 2011]. Perez-Lorenzo et al. [2012] evaluate the GCC method in real environments as opposed to simulations and concludes the PHAT weighting is most suited to these environments.

When the signal to noise ratio is reduced, the peak in the GCC function becomes more difficult to find. Rubo et al. [2011] outline work on improving the GCC-PHAT for noisy conditions by estimating the spectra of the noise component in multiple source scenarios. Hassab and Boucher [1981] specifically look at accuracy when the noise takes the form of a sinusoid and suggest a frequency dependent weighting.

Reverberation can make it difficult to discern in the GCC-PHAT output which peak corresponds to the direct sound and which peaks are early reflections and reverberation as it is correlated noise. If the room is very reverberant these early reflections can be of equal or higher amplitude to the direct sound.

Brandstein [1999] presents a method which exploits the harmonic nature of the input signals to improve results in noisy and reverberant conditions. Rui and Florenico [2004] outline a method which sets out to deal with noise and reverberation in a two stage approach but in doing so adds to the complexity of the problem. Wan and Wu [2013] propose using machine learning methods for peak picking to get a more accurate estimation of delay. Choi and Eom [2013] present a method to improve the accuracy of GCC by subsample processing.

The GCC-PHAT method is also used in source separation [Cho and Kuo,

2009]. Source separation attempts to isolate sources from a mixture by estimating the mixing parameters, usually delay and gain, of each source and using these to create unmixing filters.

Improvements to the GCC-PHAT that have been proposed are reliant on certain conditions or add additional complexity to the problem whereas the widely used GCC-PHAT has been shown to be robust in a variety of conditions. We will therefore continue to use the GCC-PHAT as it was proposed for the remainder of this chapter.

4.1.4 Delay estimation of arbitrary musical signals

A large proportion of the literature on the GCC-PHAT is aimed at human speech, often in source localisation under the name SRP-PHAT [DiBiase et al., 2001]. Therefore the input source to many experiments is a sample of human speech. More recently the GCC-PHAT has been applied to music signals. Music signals differ from speech predominantly because the type of input signal is not known beforehand and is more difficult to predict [Carey et al., 1999].

When extending any method developed for speech to be used with music inputs, the input signal is unknown and could have different characteristics e.g. spectral content, time envelope and overall energy. There is limited prior work on using the GCC-PHAT on arbitrary musical signals and what effect this might have on its performance. Work by Meyer [1992] details considerations that need to be taken when using arbitrary signals instead of traditional noise sources for transfer function calculations, such as averaging, accumulation, coherence measurement and noise reduction. Although not directly concerned with the GCC-PHAT, this work aims to estimate the impulse response between a close and a distant microphone. Therefore many of the proposals remain the same.

A study by Azaria and Hertz [1984] also suggests a link between signal bandwidth and delay estimation accuracy but focuses on narrow signal bandwidth combined with broadband noise.

Another area which has had little exposure is the effect of window shape on the GCC. The GCC requires that the Discrete Fourier Transform (DFT) of each microphone signal is calculated. When a DFT is performed a discrete frame of data is taken which can be weighted with a function such as the Kaiser or Hamming window. As each window function has its own characteristics, including the type of spectral leakage that occurs [Harris, 1978; Nuttall, 1981], this may affect the delay estimation and the window function should not be an arbitrary decision.

A theoretical study of the effect of the window function on delay estimation by Balan et al. [2000] leads to the conclusion that the error is independent

of the window shape, if the window is sufficiently wide, which is subsequently disproved by the research presented in this chapter when applied to real data.

In reality, the frame size is restrained by computation and sufficiently large frame are not necessarily practical. It also does not discuss the effect that the input signal has on delay estimation. Other work investigates the effect that window side lobes have on multifrequency signal measurement [Novotny and Sedlacek, 2010] but does not detail how this affects the phase, which is significant when discussing time delay.

A survey of the literature on implementations of the GCC-PHAT suggests no justification for the window function chosen. Research into speech source localisation [Brandstein and Silverman, 1997b] uses phase differences to calculate delay and mentions the use of a Hann window in preceding work [Brandstein and Silverman, 1997a]. An overview of delay estimation methods by Chen et al. [2006] uses the Kaiser window for the GCC-PHAT. Other works use the Hann window [Perez Gonzalez and Reiss, 2008c; Tourney and Faller, 2006] or the Hamming window [Bechler and Kroschel, 2003] without justification. Work into the differences on perception of synthesised speech using either magnitude or phase spectrum [Paliwal and Alsteris, 2005] compares two window functions, rectangular and Hamming. The GCC-PHAT relies on accurate phase measurement, but this work does not provide an explanation for how the Hamming window changes the phase and therefore alters the result compared to the rectangular window. Other examples using the GCC-PHAT in the literature do not describe the window function used.

In the remainder of this chapter we provide a novel theoretical and experimental analysis of the effect of window shape on delay estimation accuracy with real, arbitrary musical signals.

4.2 Description of the GCC-PHAT

The signal model for a single source reproduced by multiple microphones in anechoic conditions is outlined in Section 2.2.4. It is repeated here for convenience

$$x_1[n] = \alpha_1 s[n - \tau_1] \tag{4.1}$$

$$x_2[n] = \alpha_2 s[n - \tau_2] \tag{4.2}$$

where x_1 and x_2 are microphones reproducing source s, τ_1 and τ_2, and α_1 and α_2 are delays and amplitude changes associated with sound travelling from the sound source to the microphones. This is assumed to be freefield conditions. It

can also be rewritten in terms of x_1 as

$$x_2[n] = \alpha x_1[n - \tau]. \tag{4.3}$$

It is not straightforward to estimate τ_1 and τ_2 directly from (4.1) and (4.2) without any prior knowledge of s. Delay estimation methods are often referred to as Time Difference of Arrival as it is possible to estimate τ, the relative delay of a source between microphones, where $\tau = \tau_2 - \tau_1$.

The Generalized Cross Correlation, or GCC, is defined by

$$\Psi_G[k] = X_1^*[k] \cdot X_2[k] \tag{4.4}$$

in the frequency domain and

$$\psi_G[n] = \mathcal{F}^{-1}\{\Psi_G[k]\} \tag{4.5}$$

in the time domain where $\mathcal{F}^{-1}$ is the Inverse Fourier Transform, X_1 and X_2 are x_1 and x_2 in the frequency domain, $k = 0, \ldots, N-1$ where k is the frequency bin and $|*|$ denotes the complex conjugate. The delay, τ, is estimated by finding the position of the maximum of the output function, where

$$\tau = \arg\max_n \psi_G[n]. \tag{4.6}$$

The Phase Transform weighting uses only the phase of the GCC in the frequency domain to become the GCC-PHAT. This is achieved by setting the magnitude of the GCC to 1 across all frequencies, performed here by dividing (4.4) by the magnitude so (4.5) becomes

$$\Psi_P[k] = \frac{X_{1[}^*[k] \cdot X_2[k]}{|X_1^*[k] \cdot X_2[k]|} \tag{4.7}$$

in the frequency domain and

$$\psi_P[n] = \mathcal{F}^{-1}\{\Psi_P[k]\} \tag{4.8}$$

in the time domain to become the GCC-PHAT. The delay is estimated by

$$\tau = \arg\max_n \psi_P[n]. \tag{4.9}$$

An example of the output of a GCC-PHAT calculation can be seen in Figure 4.2 where the horizontal position of the peak determines the estimated delay.

Another way of expressing the GCC-PHAT is to say that it calculates the difference in phase between each microphone signal in the frequency domain

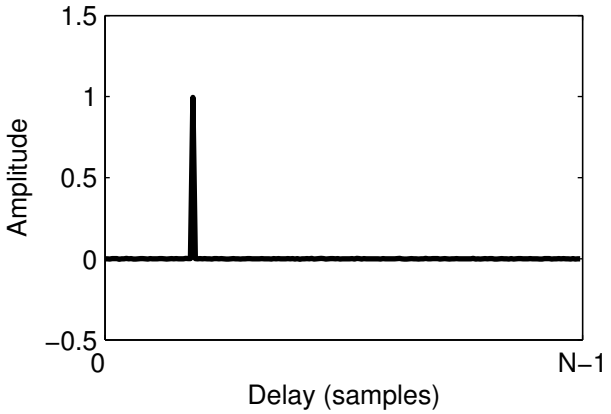


Figure 4.2: Output of the GCC-PHAT.

before being transformed back to the time domain to estimate the delay. This is because the delay between two signals is predominantly contained within the slope of the phase difference.

The shift theorem states that when a signal is delayed, a linear phase component is added. The slope of the linear phase is equal to the delay, otherwise known as group delay. The Discrete Fourier Transform X_2 of the microphone signal x_2 is defined as

$$X_2[k] = \sum_{n=0}^{N-1} w[n]x_2[n]e^{-j\omega_k n} \tag{4.10}$$

where $\omega_k = 2\pi k/N$ where N is the frame size and and w is a window function. Assuming a rectangular window function where $w[n] = 1$, using (4.3) this can be rewritten in terms of x_1 as

$$X_2[k] = \sum_{n=0}^{N-1} \alpha x_1[n-\tau]e^{-j\omega_k n} \tag{4.11}$$

$$= \alpha\Phi[k]X_1[k]. \tag{4.12}$$

where

$$\Phi[k] = e^{-j(n-\tau)\omega_k} \tag{4.13}$$

is the linear phase term applied to X_1 to become X_2. This is the desired output of the GCC-PHAT to estimate the relative delay and is therefore equivalent to

$$\Phi[k] = Arg(X_2[k]) - Arg(X_1[k]) \tag{4.14}$$

so

$$\Phi[k] = Arg(\Psi_P[k]) \tag{4.15}$$

where $Arg(\cdot)$ denotes the phase component of a complex number.

It was found by the author of this thesis that this is also equivalent to estimating the impulse response and applying the PHAT, which is the technique recommended by Perez Gonzalez and Reiss [2008c]. This is described in Appendix B.

Techniques exist to estimate the delay by calculating the gradient of the linear phase term [Brandstein and Silverman, 1997b]. This approach is highly susceptible to uncorrelated noise and requires smoothing of results. Other methods exist for using just the phase to estimate the delay [Björklund and Ljung, 2009; Assous et al., 2009] although these have been shown to exhibit poor performance. Work by Assous and Linnett [2012] outlines a method for estimating delay using a combination of frequency content and phase offset but is specific to a certain type of signal.

Studies by Donohue et al. [2007] and Salvati et al. [2011] suggest that with a harmonic input signal the Phase Transform is detrimental to the delay estimation accuracy, and outline a method for varying the degree in which the Phase Transform is applied, depending on how harmonic the signal is. We address this claim and it is discussed with analysis in Section 4.4.

4.3 Effect of windowing and signal bandwidth on delay estimation accuracy

The GCC-PHAT is still commonly used in the same form as when first introduced by Knapp and Carter [1976]. It has consistently been shown to perform adequately for speech signals in a variety of environments, and therefore no significant adaptations of the algorithm have been widely accepted.

The main variables that can be changed in the algorithm are the GCC weighting function, window shape, window size and hop size. As discussed in the previous section the research outlined in this chapter uses the Phase Transform weighting function. The window shape used with the DFTs in the GCC-PHAT has not been discussed in the literature and is an important, often overlooked stage of the calculation. This section proceeds to investigate the effect different window shapes have on delay estimation and how this relates to musical signals. The following analysis in this section was completed in collaboration with the supervisor of this research, Joshua Reiss.

As mentioned previously, the GCC-PHAT estimates the linear phase shift between X_1 and X_2 with the individual phase shift θ_k of each frequency bin k

linearly related to the sample delay τ. Taking (4.7) and assuming X_1 and X_2 are full bandwidth signals with significant data for all k, the phase difference using the GCC-PHAT then becomes

$$\Psi_P[k] = e^{j\theta_k} = e^{-j\omega\tau}. \tag{4.16}$$

The inverse DFT yields the final result

$$\psi_P[n] = \frac{1}{N}\sum_{k=0}^{N-1} e^{-j\omega\tau}e^{jn\omega_k} \tag{4.17}$$

$$= \frac{1}{N}\sum_{k=0}^{N-1} e^{j(n\omega_k-\tau\omega)} \tag{4.18}$$

$$= \begin{cases} 1 & \text{if } n = \tau \\ 0 & \text{if } n \neq \tau \end{cases} \tag{4.19}$$

which is equal to (4.8) and the delay can be accurately estimated as τ. For (4.16) to hold, θ_k has to be correct for all values of k.

A real signal, such as a musical instrument input source, will not fill the audible bandwidth. Different instruments produce notes that occupy different areas of the frequency spectrum. Percussive instruments may produce a more noise-like sound that occupies a large part of the spectrum whereas a harmonic instrument, such as a flute, will primarily produce harmonics of a fundamental frequency. There will also be a limit to the range of notes an instrument can produce and therefore the fundamental frequency.

In the extreme case of a narrow bandwidth signal, taking a single complex sinusoid $s = e^{j\omega n}$ where $\omega = 2\pi\hat{k}/N$, $\hat{k}$ is an integer $0 \leq \hat{k} < N-1$ and $s_\theta = e^{j(\omega n+\theta)}$ we know from the shift theorem that

$$S_\theta[k] = e^{j\theta}S[k] \tag{4.20}$$

where S is s in the frequency domain. S will have a single non-zero value when $k = \hat{k}$. Hence when $k \neq \hat{k}$

$$\frac{S_1^*[k]\cdot S_2[k]}{|S_1^*[k]\cdot S_2[k]|} \neq e^{j\theta} \tag{4.21}$$

as this leads to division by 0 and therefore it is undefined.

The delay cannot be estimated from the value of θ as this is only correct for when $k = \hat{k}$ so gives no context as to the slope of the phase and thus the corresponding delay in samples. The GCC-PHAT will therefore not be able to estimate a delay as the phase is only correct when $k = \hat{k}$. In reality due to

the real environment and background noise, all k will be defined. But this will manifest as noise in the GCC-PHAT, therefore the correct delay will still not be estimated.

In (4.20), s is assumed to contain an integer number of periods within N. Spectral leakage occurs when the input signal contains a non-integer number of periods within the window and can be attributed to the Gibbs phenomenon [Pan, 2001], [Gottlieb and Shu, 1997].

This is often the case with real signals. The result of this is that for a single sinusoid the frequency domain signal is no longer a delta function but resembles the frequency spectrum of the particular window function.

The spectral leakage also implies that all values of k will be defined, which is not the case in (4.21). If $s = e^{j\omega n}$ where $\omega = 2\pi\hat{k}/N$ and $\hat{k}$ is not an integer then all k will be defined and the GCC-PHAT can be calculated. Despite this, the correct delay will still not be estimated as the phase from the nearest value of k to $\hat{k}$ will spread into neighbouring bins. If $\theta_k = \theta$ for all k due to the leakage, (4.16) does not hold. As θ_k is a single value, the slope is 0. Therefore the delay is estimated as 0, which is incorrect.

The more values of θ_k that are the correct estimate of the real phase difference, the more likely the estimation of delay will be correct. The errors are caused by spectral leakage and become more apparent when considering a real signal as a sum of sinusoids at different amplitudes and frequencies. This is due to the interference between side lobes of high amplitude sinusoids and low amplitude sinusoids which is also known to affect multifrequency signal measurement [Novotny and Sedlacek, 2010]. If a sinusoid is of lower amplitude than the side lobe of a neighbouring, higher amplitude sinusoid in the frequency domain it will be distorted or completely masked in both magnitude and phase.

Therefore if the bandwidth of the signal is increased, with more higher amplitude sinusoids, more values of θ_k will be correct. Equally, if the side lobes are lower amplitude either due to the window shape producing lower maximum amplitude side lobes or having a steeper side lobe roll off rate, then less lower amplitude side lobes will be masked and accuracy will be improved.

From this we hypothesise that delay estimation accuracy is dependent on the incoming signal bandwidth and the characteristics of the window shape chosen.

4.4 Experimental analysis

This section outlines an experiment analysis of how the bandwidth of the input signal and the window used affect the accuracy of the subsequent delay estimation when performing the GCC-PHAT on simulated and real musical signals.

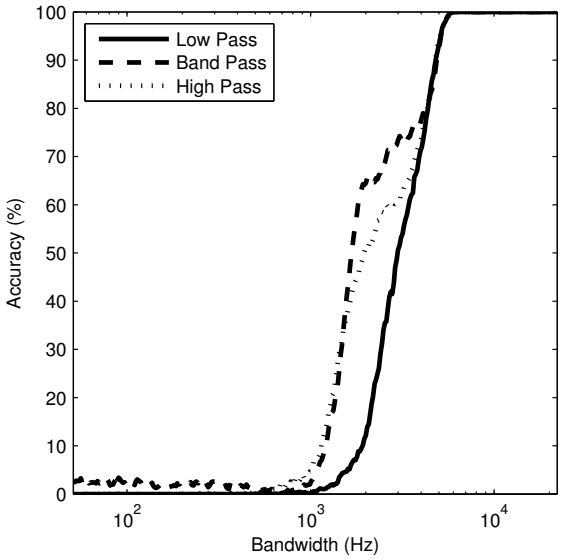


Figure 4.3: Accuracy of delay estimation as a percentage of correct frames with an error of ± 2 samples using a rectangular window with increasing bandwidth using low pass, high pass and band pass filter centred at 11.25kHz.

4.4.1 Bandwidth limited white noise

The variation between musical signals in the frequency domain can be simplified as stating that different instruments will produce sounds which occupy different areas of the frequency spectrum with different bandwidths [Katz, 2007]. The effect that this has on the GCC-PHAT can be observed under controlled conditions, not taking into account amplitude or temporal changes, by using filtered white noise as an input signal. This was used in the analysis as an input to simulate microphone signals by duplicating the filtered input signal and delaying the duplicate by 10 samples at 44.1kHz sampling rate. The audio excerpts were 10 seconds in duration.

The white noise was filtered using low pass, high pass and band pass 4th order Butterworth filters centred at 11.25kHz to investigate whether the centroid of the spectrum altered the accuracy. For each execution of the simulation the bandwidth of the three filters was changed. In the case of the low and high pass filters the cut-off frequency was changed to achieve the desired bandwidth. The bandwidth of each filter was then varied between 50Hz and $\frac{F_s}{2}$ where F_s is the sampling frequency. The delay was estimated at each execution with the GCC-PHAT using seven window shapes: Blackman, Blackman-Harris, Flat Top, Gaussian, Hamming, Hann and rectangular, with a frame size of 2048 samples. The accuracy is determined as a percentage of frames over the 10

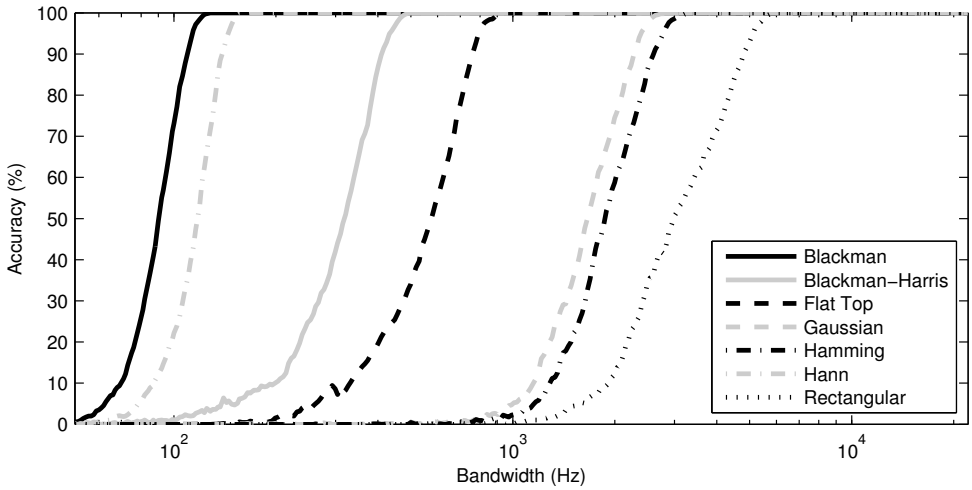


Figure 4.4: Accuracy of delay estimation as a percentage of correct frames with an error of ±2 samples using a selection of windows with increasing bandwidth using a low pass filter.

second sample in which the delay was estimated correctly with an error of ±2 samples.

Figure 4.3 shows the results using the rectangular window. It can be seen that for all filters at the same bandwidth the results were similar and the point at which 100% accuracy is achieved was the same for all filters. This leads to the conclusion that the centroid of the spectrum has only a minor effect on the accuracy of delay estimation. Therefore the low pass filter results are used for the analysis in the rest of this section.

Figure 4.4 shows the results for all windows tested for the low pass filter with increasing bandwidth of input signal. This shows that the performance of the delay estimation was different for each window and therefore the choice of window should not be trivial. The rectangular window reached 100% accuracy at a bandwidth of 5937Hz, whereas the Blackman window reached 100% accuracy at a bandwidth of 128Hz. The accuracy increased as bandwidth increased for all window shapes.

Table 4.1 shows the mean accuracy for each window shape over all input source bandwidths ranked in descending order from most accurate to least accurate. The side lobe height, side lobe roll-off and start and end values are also shown. The window shapes with a 60dB/decade side lobe slope outperformed the windows with 20dB/decade slope. The Blackman window also appeared more accurate than the Hann window by 4% since it has a lower side lobe maximum height. The accuracy of the windows that do not taper to 0 then decreased

Window	Mean accuracy (%)	Maximum side lobe height (dB)	Side lobe roll-off (dB/decade)	Start/end value
Blackman	90.74	-58.1	60	0
Hann	86.67	-31.5	60	0
Blackman-Harris	71.00	-71.5	20	6.0×10^{-5}
Flat Top	61.34	-93.6	20	-4.2×10^{-4}
Gaussian	43.00	-43.3	20	4.3×10^{-2}
Hamming	40.82	-42.7	20	0.08
Rectangular	32.85	-13.3	20	1

Table 4.1: Mean accuracy over all filter bandwidths for low pass filtered noise for each window shape showing window features.

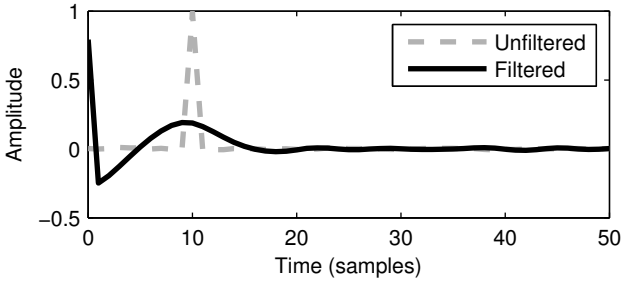
according to the start value. This confirms the hypothesis that windows with a steeper side lobe roll off slope or lower side lobe maximum height result in higher accuracy.

To explain this further, Figure 4.5 shows the GCC-PHAT output using a rectangular window and equivalent phase spectrum for white noise low pass filtered with a cut off frequency of 1000Hz using a 4th order Butterworth filter and unfiltered white noise delayed by 10 samples. Figure 4.5a shows the GCC-PHAT output of the low pass filtered and unfiltered white noise. The unfiltered GCC-PHAT shows a very clear peak at the delay value of 10 samples. The filtered GCC-PHAT has a peak at the correct delay value but also a peak at 0, which is the maximum and therefore the estimated delay.

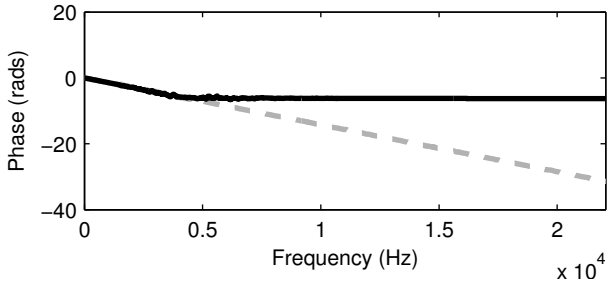
One should not ignore the values at $\tau = 0$ when performing the GCC-PHAT as it is possible that no delay occurs and these needs to be estimated. This is explained by examining the corresponding phase spectrum in Figure 4.5b. The unfiltered example shows a distinct linear phase whereas the filtered example shows the sloped linear phase for the pass band of the filter, up to 1000Hz, but in the cut band of the filter the phase is constant, corresponding to the significant 0 peak in the GCC-PHAT output. This is a result of the higher amplitude spectral leakage of the rectangular window. With the Blackman or Hann windows, this does not occur and hence the GCC-PHAT output is the same for both filtered and unfiltered signals.

4.4.2 Real recordings

The window shapes being evaluated were tested on real recordings. The recordings were made using two omnidirectional AKG C414 microphones. They were placed at arbitrary distances from a Genelec 8040 loudspeaker to incite a delay between the microphone signals and were recorded in the listening room at the



(a) GCC-PHAT output of white noise.



(b) Phase spectrum of white noise.

Figure 4.5: The GCC-PHAT output and corresponding unwrapped phase spectrum of unfiltered and low pass filtered white noise.

Centre for Digital Music in Queen Mary, University of London. The microphone signals were analysed using the GCC-PHAT with various window shapes. 20 different musical audio samples were tested, with each audio sample 30 seconds in duration. The audio samples were a selection of instrument recordings that occupy different frequency ranges.

The bandwidth of each audio sample was measured by calculating spectral spread, or standard deviation in the frequency domain, defined by

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1}(|X[k]| - \bar{X})^2} \tag{4.22}$$

where

$$\bar{X} = \frac{1}{N}\sum_{k=0}^{N-1}|X[k]|. \tag{4.23}$$

and X is the input signal x in the frequency domain. The spectral spread was estimated over the whole duration of the audio sample. Therefore N is the duration of the audio clip measured in samples.

Figures 4.6 and 4.7 show the accuracy of delay estimation for each audio sample plotted against the spectral spread. Figure 4.6 shows the results of

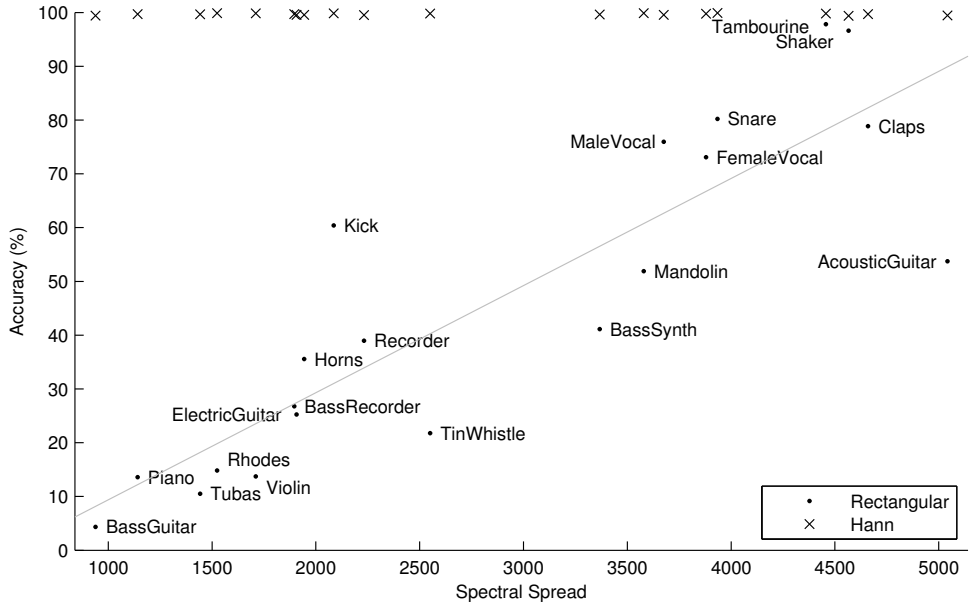


Figure 4.6: Delay estimation accuracy for 20 audio excerpts using a rectangular window plotted against spectral spread. The accuracy is also shown for the Hann window unlabelled for comparison and enlarged in Figure 4.7

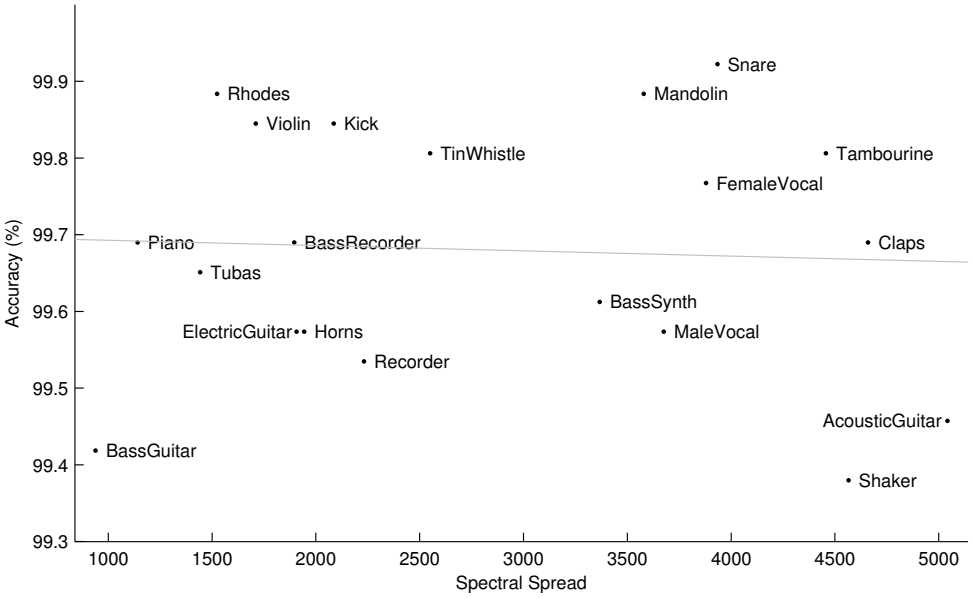
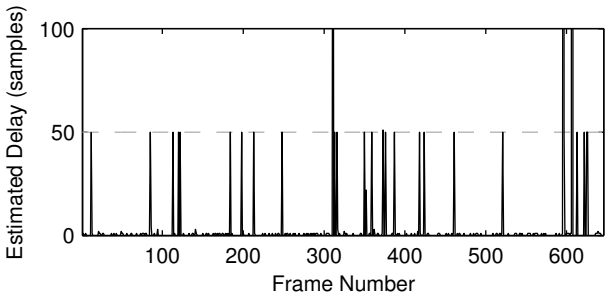
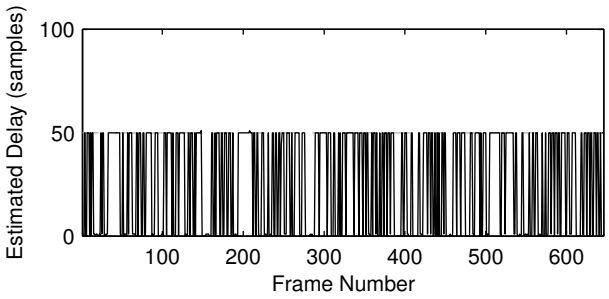


Figure 4.7: Delay estimation accuracy for 20 audio excerpts using a Hann window plotted against spectral spread.



(a) Bass guitar



(b) Acoustic guitar

Figure 4.8: Output of the GCC-PHAT using the rectangular window shown as estimated delay for each frame of data. The dashed horizontal line indicates the correct delay.

delay estimation using the rectangular window and Figure 4.7 the results using the Hann window. The Hann window is used because in the literature survey the Blackman was not found to have been used with the GCC-PHAT previously and there was only a 4% difference in accuracy between the Hann and Blackman windows in the previous section. In Figure 4.6 it is apparent that the accuracy of the delay estimation increased as the spectral spread (and thus the bandwidth of the signal) increased. As expected, this is not the case for the Hann window, which gave better performance for all test audio sample, although 100% accuracy was not achieved due to the recording environment.

This can be further explained by analysing the estimation data over time for different inputs. Figures 4.8a and 4.8b shows the delay estimation using a rectangular window for each frame of data of two example audio samples, a bass guitar and an acoustic guitar. The estimation for the bass guitar was inaccurate with the correct delay rarely being estimated and an estimate of 0 being more likely. In comparison, the acoustic guitar resulted in an estimated delay of either 0 or the correct delay per frame.

Figure 4.9 shows the mean accuracy of all 20 test recordings for frame sizes from 128 samples to 8192 samples for each window shape. There was a general

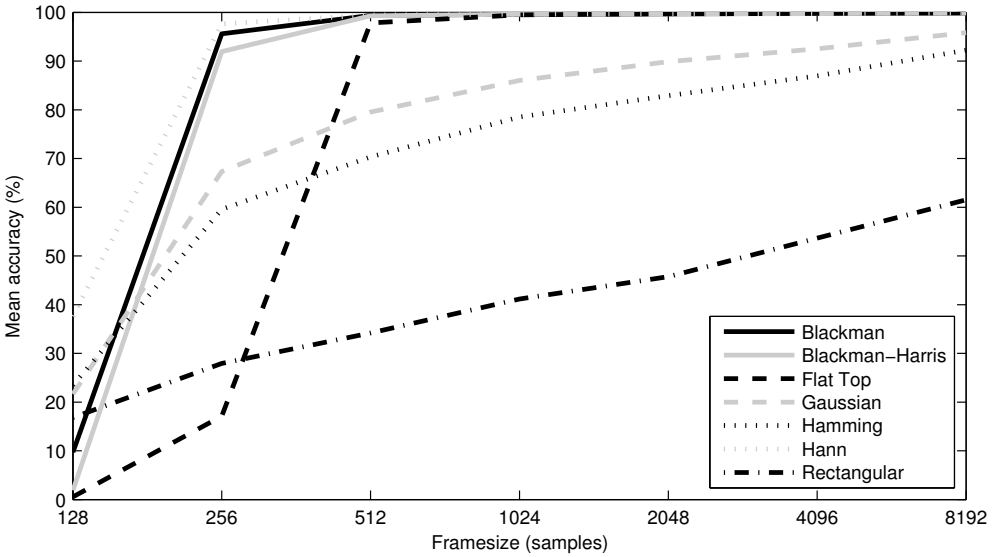


Figure 4.9: Mean accuracy of delay estimation over all audio excerpts using a selection of common frame sizes and windows.

trend of increasing accuracy as frame size increased. This is expected as it is known that increasing the window size increases the accuracy of the GCC-PHAT [Jillings et al., 2013]. But the differences in performance from each window remained even at large frame sizes. Although a large frame size achieved the greatest accuracy, larger frame sizes reduces the ability of the GCC-PHAT to track changing delays at fine accuracy.

Table 4.2 shows the mean of all frame sizes for each window. The results followed a similar trend as that for the filtered white noise. The Hann and Blackman windows provided the greatest accuracy with a side lobe roll of 60dB/decade followed by windows with low amplitude side lobes. The rectangular window continued to perform the worst.

In this section we have shown that accuracy of delay estimation for comb filter reduction is dependent on the incoming signal bandwidth and the DFT window used.

4.5 Discussion and conclusions

In this chapter we have discussed using delay estimation to reduce comb filtering in single source, multiple microphone configurations.

We have provided a novel analysis of the GCC-PHAT method of delay estimation regarding the bandwidth of the incoming signal and the DFT window

Window	Mean accuracy (%)	Maximum side lobe height (dB)	Side lobe roll-off (dB/decade)	Start/end value
Hann	90.52	-31.5	60	0
Blackman	86.24	-58.1	60	0
Blackman-Harris	84.58	-71.5	20	6.0×10^{-5}
Gaussian	76.11	-43.3	20	-4.2×10^{-4}
Flat Top	73.49	-93.6	20	4.3×10^{-2}
Hamming	70.57	-42.7	20	0.08
Rectangular	40.14	-13.3	20	1

Table 4.2: Mean accuracy over all audio excerpts and frame sizes for each window shape showing window features.

shape used. This is important when applying the GCC-PHAT to musical instrument sound sources in live sound because the bandwidth of different sources can vary.

The literature review into the GCC-PHAT for a variety of applications shows no consideration for the window shape used. Therefore the results of this research have implications for all uses of the GCC-PHAT for delay estimation.

We found that delay estimation of low bandwidth signals can be improved by using an appropriate window function prior to the GCC-PHAT calculation. We showed that windows which taper to 0 at the extremities are most appropriate, for example the Hann or Blackman windows, as they produce lower side lobes in the frequency domain which means less lower amplitude frequencies in the incoming signal are masked and therefore contribute to an accurate estimation of delay.

Within a ± 2 sample error, a 58% mean increase in accuracy was achieved when using a Blackman window over a rectangular window in simulated recordings. On real recordings an improvement in mean accuracy of 50% was achieved. The improvement was shown over a range of window sizes, with the Hann window offering the best performance at a 128 sample window size, the smallest size tested, with a mean accuracy of 37% compared to a mean accuracy of 17% for the rectangular window.

The results also showed that the instrument recordings with low bandwidth, measured by spectral spread, such as a Bass guitar achieved the greatest increase in accuracy when using a Hann window over a rectangular window. Percussive sounds which have a high bandwidth were less affected by the difference in window shape.

In the next chapter we further extend the single source, multiple microphone case to the multiple source multiple microphone case. This scenario can cause

bleed between the microphones and we discuss a method for reducing this. We also discuss multiple source delay estimation, which extends the GCC-PHAT to the multiple source case.

Chapter 5

Determined microphone bleed reduction

This chapter is the first of two chapters concerned with reducing microphone bleed in multiple source, multiple microphone configurations. Microphone bleed occurs when a microphone is picking up other sources in an acoustic space that are not the target source. This is common in an ensemble performance where each instrument has its own microphone but they are in close proximity to each other.

We present the state of the art in approaches to microphone bleed reduction, and outline Crosstalk Resistant Adaptive Noise Cancellation (CTRANC), on which the bleed reduction methods proposed in this thesis are based. The two source, two microphone CTRANC is extended by combining it with centred adaptive filters. Centring the filters is achieved using a multiple source extension of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) method of delay estimation, as presented in the previous chapter for use in comb filter reduction.

The proposed centred CRANC method is compared to a method of source separation and a method of noise cancellation, as well as the original CTRANC. It is shown to perform well in anechoic conditions but begins to break down in reverberant conditions.

5.1 State of the art

In Section 2.2.5 we described the cause and effect of microphone bleed in multiple source, multiple microphone configurations. In this section we present the state of the art in reducing microphone bleed.

5.1.1 Physical methods

Microphone bleed is caused by multiple microphones reproducing multiple sources in the same acoustic space. Microphone bleed can be reduced by physically sep-

arating the sources, either complete separation by placing the instruments in separate spaces, or maximising the separation of sources in the same space. In a studio situation, for example, instruments can be isolated either in separate live rooms or by erecting baffles to provide some sound isolation. However, in a live sound situation this is often not aesthetically appropriate.

Microphone bleed can also be reduced by using appropriate microphone placement, for example by using directional microphones directed towards the source of interest and placing interfering sources in the rejection areas of the microphone pick up area. The problem with this is that complete rejection of interfering sources is challenging and using directional microphones introduces other issues, such as the proximity effect, which is addressed in Chapter 3.

Sound engineers may apply equalisation (EQ) to the microphone signals to try and reduce the effect of an interfering source. However, if the interfering and target source overlap in frequency then EQ will also affect the target source, which is undesirable. It is possible to apply a noise gate to a particular microphone to only allow the target source to be heard when it is played [Izhaki, 2007, chap. 18]. This is particularly effective in drum recordings where the target drum is very high amplitude. It is an effective technique if the target source is not played often, such as tom-toms in a drum kit, but the gate is not selective so all sounds will be heard, including the bleed, once the gate is triggered.

5.1.2 Blind source separation

As the amount of manual correction that a sound engineer can achieve with the tools they have available is limited, we have to turn to signal processing techniques to reduce the microphone bleed effectively.

This can be approached from a Blind Source Separation (BSS) perspective. BSS methods aim to separate multiple sources in underdetermined, overdetermined or determined configurations with little to no information about the sources or the mixing process. It is a wide and active area of research with many approaches offered for different configurations. Makino et al. [2007] describe the early research into blind source separation methods, which initially assumed instantaneous mixtures of sources, i.e. where the only mixing parameter is amplitude. The signal model we outlined in Section 2.2.5 includes delay and gain, as is often seen in a real acoustic environment, therefore this can be considered a convolutive mixture.

BSS of convolutive mixtures involves estimating the unmixing filters of a particular mixing matrix [Araki et al., 2003; Pedersen et al., 2007; Mitianoudis and Davis, 2003]. There are a wide variety of methods to achieve this in the time and frequency domain. An overview of convolutive BSS techniques is provided

by Pedersen et al. [2007] and outlines assumptions and definitions relative to convolutive BSS.

The straightforward method is to invert the FIR mixing filters with IIR filters. This requires that the IIR filters are stable [Uhle and Reiss, 2010]. Once applied to real scenarios any errors in the filter estimation cause audible artefacts in the target signal. They are also inherently time invariant. Therefore, if the position of sources or microphones is changed, this causes errors in the filters. For convolutive mixtures the mixing and inverting filters can be long, causing computation and stability issues.

A commonly used technique in the frequency domain is Independent Component Analysis (ICA) [Comon, 1994; Cardoso, 1998], although this assumes statistical independence between sources, which cannot always be guaranteed in a real situation, for example if different instruments perform the same piece of music. ICA of convolutive mixtures is performed in the frequency domain by assuming each frequency bin is an instantaneous mixture and processed as such. A full overview of ICA methods is provided by Hyvärinen et al. [2001]. The performance of ICA methods begins to break down on filters with a large number of weights [Vincent, 2006; Vincent et al., 2007, 2010], such as with long reverberation times. Other frequency domain methods also perform BSS on each bin separately [Araki et al., 2003], which can cause frequency artefacts of the separated signals.

Many BSS methods are developed for separation of speech signals [Makino et al., 2007] and can fail when they are applied to a real world environment [Westner, 1999]. It is noted by Benesty et al. [2008a], Parra and Spence [2000] and Pedersen et al. [2007] that many BSS methods are shown to perform in simulations but fail when applied to sources in real world conditions.

This work is aimed at live sound, therefore it is important that a method is able to run in real time in real world conditions. A number of real-time BSS methods exist [Barry et al., 2004; Rickard et al., 2001; Baeck and Zölzer, 2003; Aichner et al., 2006]. The method by Baeck and Zölzer [2003] is taken from the Degenerate Unmixing Estimation Technique (DUET) method of source separation, first presented by Jourjine et al. [2000] and extended by Yilmaz and Rickard [2004]. Although stated to run in realtime, this method of source separation is aimed at the unmixing of L sources from 2 mixtures, i.e. from a stereo mix of panned sources. It is possible to use this method in the two source, two microphone configuration, but it is limited to configurations with a small distance between the microphones, which are dependent on sampling frequency. For example at a 16kHz sampling frequency the maximum distance allowed between the microphones for the method to run is when $d \leq 2.15$cm [Yilmaz and Rickard, 2004] where d is the distance between the microphones.

The method by Barry et al. [2004] is also used for stereo mixtures, assuming there is phase coherence between the mixtures and only intensity differences. This cannot be assumed in the multiple microphone case.

A selection of BSS methods are also aimed at Music Information Retrieval applications [Nesbit et al., 2007] where distortion of the target signal is acceptable in favour of complete separation. This is also echoed by Pedersen et al. [2007] who state that the separated signals from BSS can be considered interference-free and scaled or filtered versions of the original signals. In the live sound context we are investigating, large distortions in the target signal are not acceptable. Kokkinis et al. [2011] also suggest that BSS methods can rescale or reorder the separated signals, which may cause problems in live sound with gain structure and feedback.

5.1.3 Noise cancellation

Many of the problems that affect live sound are also present in telecommunications, for example noise and reverberation. Techniques exist in telecommunications for echo and noise cancellation, which share the same principles, and also run in real-time [Benesty et al., 2008a]. The drawback is that most techniques are optimised for speech signals with lower bandwidths, for example a sampling rate between 4kHz and 8kHz is common [Mirchandani et al., 1992; Hetherington et al., 2010] whereas in live sound we require a bandwidth to represent all the audible frequencies from 20Hz to 20kHz. For this reason, when an algorithm optimised for speech application is extended to incorporate wider bandwidth signals, the computational cost inherently increases.

In telecommunications, it is common that an external noise source will interfere with the direct source. For example, there may be an interfering noise, such as air conditioning, in the same room as a person speaking into a telephone. If an adequate estimation of the noise source is possible, this can be removed from the direct signal. This is where noise and echo cancellation can be used.

Common techniques for noise cancellation make use of an adaptive filter to estimate the impulse response of the noise signal to the target microphone, first proposed by Widrow et al. [1975]. These methods rely on a clean reference of the noise signal. In reality, this is not always achievable. In a live sound scenario, a clean reference signal may not be available as microphone bleed is assumed to be occurring on all microphone signals. The scenario we are concerned with in this chapter assumes that any interfering source is also a target source for a different microphone which also contains interference. It cannot be assumed that a clean reference of each interfering source is available.

Common adaptive filters for audio applications are Least Mean Squares

(LMS) or Recursive Least Squares (RLS) [Haykin, 2001]. The RLS filter is considered to have the faster adaption rate but at a higher computational cost than the LMS filter [Hadei and Iotfizad, 2010]. In this research the LMS filter is used since the computational cost is already increased due to the wideband music signals that are being processed.

The performance of adaptive filters can be improved by using an estimate of delay to centre the updated coefficients and improve convergence and computational cost [Margo et al., 1993; Lu et al., 2005; Gross et al., 1992].

Adaptive filters are sometimes favoured over BSS techniques due to the reduced target signal distortion [Ramadan, 2008]. Adaptive filters also do not require assumptions about the layout of the sources and microphones.

Work by Kokkinis and Mourjopoulos [2010] and Kokkinis et al. [2011] addresses the same problem and assumes close microphones. This is achieved by finding the Wiener filter solution in multiple channel configurations (MCWF) by Power Spectral Density estimation. This method has been shown to outperform a common BSS technique in real world conditions.

CTRANC

Noise cancellation has been extended in the literature to tackle scenarios where crosstalk occurs, known as CTRANC. First proposed by Zinser et al. [1985] and extended by Mirchandani et al. [1992], this approaches a similar problem to that of microphone bleed. CTRANC has been extended more recently, but only applied in telecommunications and to speech signals in the determined case.

Lepaulox et al. [2009] propose a method to reduce the complexity of the algorithm through changes to the filter update equations. Lepauloux et al. [2010] also suggest frequency domain implementation, applied to beamformers. Moir and Harris [2012] outline an extension to CTRANC for non-stationary sources using multivariate LMS, but it is still applied to speech signals. Madhavan and de Bruin [1990] outline another extension but is only tested on toy examples. Ramadan [2008] proposes a method for the two source case where three microphones are used and exploits the extra microphone for increased crosstalk reduction. Zeng and Abdulla [2006] combine CTRANC with improved spectral subtraction. CTRANC has also been referred to in the literature as symmetric adaptive decorrelation [Van Gerven and Van Compernolle, 1995; Geravanchizadeh and Rezaii, 2009]. All the publications mentioned only apply CTRANC to speech signals.

CTRANC tackles the same problem as unmixing filters in BSS in the determined case, but uses adaptive filters instead. The advantage of using adaptive filters is that stationarity is not assumed and they can adapt to changing condi-

tions. They are also built for real time application and are well established and will introduce fewer artefacts. By using a noise cancellation based approach, we assume that each microphone contains a single target source and this target source is a contaminated noise source of another microphone. CTRANC has not previously been applied to the live sound configuration outlined in this thesis and it is has not been applied to musical signals.

5.2 Description of Crosstalk Resistant Adaptive Noise Cancellation

This section presents a description of Adaptive Noise Cancellation (ANC) and the extension of this to CTRANC.

An example of of an application noise cancellation in telecommunications is a situation where the voice of a person speaking into a telephone may be contaminated by external noise, such as air conditioning or other background noise, that is in the same space. The most straightforward method of removing this noise is to convolve a clean reference of the noise signal with the impulse response between the noise microphone and the target microphone and subtract it from the target microphone.

The source microphone, x_s, can be described by

$$x_s[n] = h_x[n] * s[n] + h_v[n] * v[n] \tag{5.1}$$

where h_x is the impulse response from the source to the microphone x_s, h_v is the impulse response of the interfering source v to x_s and $*$ denotes convolution. Our clean reference is

$$x_v[n] = h_d * v[n] \tag{5.2}$$

where h_d is the impulse between the interfering source and x_v. To achieve noise cancellation we have to perform

$$\hat{x}_s[n] = x_s[n] - h_{x,v}[n] * x_v[n]. \tag{5.3}$$

This relies on v being the only source in x_v, and an accurate estimate of $h_{x,v}$.

Assuming (5.2), $h_{x,v}$ is often estimated using an adaptive filter since it is able to adapt to changing conditions, such as movement of sources and microphones or amplitude changes in either the source or the noise.

Adaptive filtering can be achieved with an LMS approach. We can rewrite (5.3) as

$$\hat{x}_s[n] = x_s[n] - \mathbf{w}^T[n]\mathbf{x_v}[n] \tag{5.4}$$

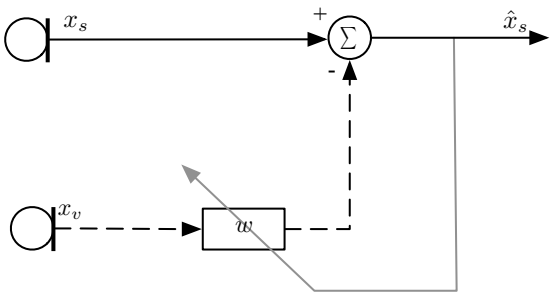


Figure 5.1: Block diagram of an adaptive filter.

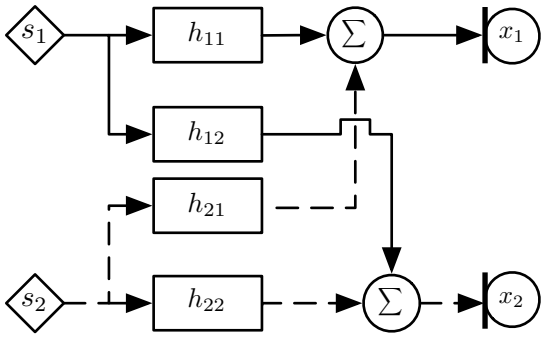


Figure 5.2: Block diagram of sources s_1 and s_2 processed by RIRs to become microphones x_1 and x_2.

where $\mathbf{w}$ are the current estimate filter weights, $\mathbf{w}[n] = [w[0], \ldots, w[N-1]]$, $\mathbf{x_v}[n] = [x_v[n], \ldots, x_v[n-N+1]]$ and N is the filter length. This is shown in Figure 5.1.

In the literature $\hat{x}_s$ is the error signal which we want to minimise by way of our cost function $\mathcal{E}\{|\hat{x}_s[n]^2|\}$ by optimising $\mathbf{w}$.

The filter weights are then updated by

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu \mathbf{x_v}[n]\hat{x}_s[n] \tag{5.5}$$

where μ is the adaptation step, which is generally a small value that affects convergence speed and accuracy.

In the multiple source, multiple microphone scenario outlined in Section 2.2.5, we cannot assume that a clean reference of the interfering noise sources is available, due to all sources being in the same acoustic space.

Figure 5.2 shows how (5.1) can be extended to the two source, two microphone case. In Section 2.2.5 this was written in the anechoic case, repeated

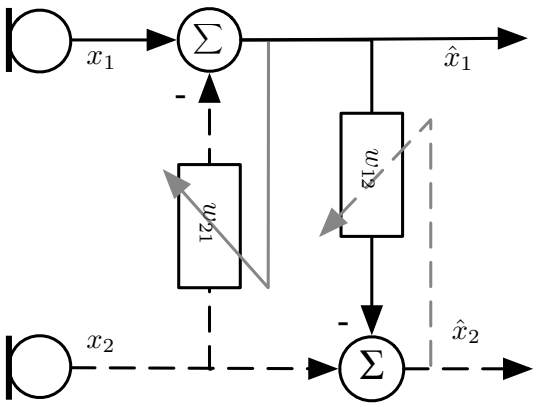


Figure 5.3: Block diagram of the two source, two microphone CTRANC method of noise cancellation.

here

$$x_1[n] = \alpha_{11} s_1[n - \tau_{11}] + \alpha_{21} s_2[n - \tau_{21}] \tag{5.6}$$

$$x_2[n] = \alpha_{22} s_2[n - \tau_{22}] + \alpha_{12} s_1[n - \tau_{12}] \tag{5.7}$$

where x_1 and x_2 are microphone signals at timestep n, s_1 and s_2 are the sound sources, τ_{lm} is the delay of source l to microphone m and α_{lm} is the amplitude due to distance of source l to microphone m.

It is apparent that both microphones are reproducing both sources. Therefore the single microphone ANC cannot be applied here.

In CTRANC, adaptive filters are cascaded so the output of one becomes the input of the other [Parsa et al., 1996], as shown in Figure 5.3. In this way, once one signal has the interference cancelled out it can be used as the reference for the interference cancellation of another source, and vice versa.

This relies on the assumption that the source with the highest amplitude in each microphone signal is the target source, i.e. $\alpha_{11} > \alpha_{21}$ and $\alpha_{12} > \alpha_{22}$. If this is the case, then each microphone can be considered an approximation of a interfering source.

In the live sound case this usually equates to each microphone being positioned closest to a single sound source. Placing spot microphones is a technique used in ensemble performances, where a single microphone is positioned to reproduce a single instrument source and therefore this is not a difficult assumption to hold in real conditions. Thus, each microphone is an approximation of an interfering noise source for a microphone other than itself.

CTRANC is described by the block diagram in Figure 5.3. The processed

microphone signals are estimated by

$$\hat{x}_1 = x_1[n] - \mathbf{w}_{21}^T[n]\mathbf{x}_2[n] \tag{5.8}$$

$$\hat{x}_2 = x_2[n] - \mathbf{w}_{12}^T[n]\hat{\mathbf{x}}_1[n] \tag{5.9}$$

and the filter weights updated by

$$\mathbf{w}_{21}[n+1] = \mathbf{w}_{21}[n] + \mu\mathbf{x}_2[n]\hat{x}_1[n] \tag{5.10}$$

$$\mathbf{w}_{12}[n+1] = \mathbf{w}_{12}[n] + \mu\hat{\mathbf{x}}_1[n]\hat{x}_2[n]. \tag{5.11}$$

5.3 Centred adaptive filters

In the previous section, (5.4) and (5.5) outline the standard adaptive filter architecture. In the purely anechoic case, the ideal output of the adaptive filter in (5.5) will be an impulse response with a single value at a position representing delay and an amplitude representing gain and all other values are assumed to be 0.

In reality, with the addition of reverberation and noise it is unlikely that the any of the values in the impulse response will be equal to 0, but there will still be a peak at the delay position. If the delay value is known, it is then possible to update only the values around the delay value. Updating fewer coefficients means faster and more accurate convergence and less computational cost. Errors in the adaptive filters will also be reduced, which will reduce the artefacts in the processed signal. Only a rough estimation of delay is required as a range of coefficients around the estimated delay value are updated. If the delay estimation is inaccurate by less than the number of coefficients in the range being updated, then the method will still converge to the solution [Margo et al., 1993; Lu et al., 2005; Gross et al., 1992].

If we centre the adaptive filters in (5.4) and (5.5) then the following variables are defined by

$$\mathbf{w}[n] = [w_{\tau-D}[n], \ldots, w_{\tau+D}[n]] \tag{5.12}$$

$$\mathbf{x}_2[n] = [x_2[n-\tau-D], \ldots, x_2[n-\tau+D]\,, \tag{5.13}$$

where τ is the estimation of the delay and D is a user-defined error distance around the delay to update the coefficients. A higher value of D will yield slower convergence but will encompass additional echoes or reverberation.

5.4 Centred CTRANC

We propose combining CTRANC with the centred adaptive filters which we will refer to as 'centred CTRANC'. In this way we can improve performance and convergence of the CTRANC method.

As with the CTRANC method, the error signals are defined as

$$\hat{x}_1[n] = x_1[n] - \mathbf{w}_{21}^T[n]\mathbf{x_2}[n] \tag{5.14}$$

$$\hat{x}_2[n] = x_2[n] - \mathbf{w}_{12}^T[n]\hat{\mathbf{x}}_\mathbf{1}[n], \tag{5.15}$$

but now

$$\mathbf{w}_{21}[n] = \left[w^{21}_{\tau_2-D}[n], \ldots, w^{21}_{\tau_2+D}[n]\right] \tag{5.16}$$

$$\mathbf{w}_{12}[n] = \left[w^{12}_{\tau_1-D}[n], \ldots, w^{12}_{\tau_1+D}[n]\right] \tag{5.17}$$

and

$$\mathbf{x_1}[n] = [x_1[n-\tau_1-D], \ldots, x_1[n-\tau_1+D] \tag{5.18}$$

$$\mathbf{x_2}[n] = [x_2[n-\tau_2-D], \ldots, x_2[n-\tau_2+D] \tag{5.19}$$

and the filter coefficients are updated using

$$\mathbf{w}_{21}[n+1] = \mathbf{w}_{21}[n] + \mu\hat{\mathbf{x}}_\mathbf{1}x_2[n] \tag{5.20}$$

$$\mathbf{w}_{12}[n+1] = \mathbf{w}_{12}[n] + \mu\mathbf{x_2}x_1[n], \tag{5.21}$$

which requires estimation of both τ_1 and τ_2.

5.5 Multiple source delay estimation

To successfully implement the centred adaptive filters, an accurate estimation of the delay is required. In the previous chapter we outlined a number of methods for delay estimation and investigated the GCC-PHAT method, which is fully described in Section 4.2.

For the centred CTRANC applied to (5.6) and (5.7) we need to estimate both τ_1 and τ_2 where

$$\begin{aligned} \tau_1 &= \tau_{21} - \tau_{11} \\ \tau_2 &= \tau_{12} - \tau_{22} \end{aligned} \tag{5.22}$$

and it is assumed that

$$\begin{aligned} \tau_{11} &< \tau_{21} \\ \tau_{22} &< \tau_{12}. \end{aligned} \tag{5.23}$$

The GCC-PHAT is aimed at estimating the delay of a single source to multiple microphones since the delay is estimated by finding the maximum peak in the time domain function. If the GCC-PHAT is applied to the microphones described in (5.6) and (5.7) the relative delay would only be estimated for the source with the highest amplitude in both microphones as this will have the greatest correlation.

We can use the GCC-PHAT to estimate τ_1 and τ_2 separately by interchanging x_1 and x_2 in the calculation, using the constraint that the estimated delay must be less than $N/2$. The GCC-PHAT has been described previously in (4.7). In the two source case the GCC-PHAT is rewritten for each delay as

$$\begin{aligned} \Psi_{P12}[n] &= \mathcal{F}^{-1}\left\{\frac{X_1^*[k]\cdot X_2[k]}{|X_1^*[k]\cdot X_2[k]|}\right\} \\ \Psi_{P21}[n] &= \mathcal{F}^{-1}\left\{\frac{X_1[k]\cdot X_2^*[k]}{|X_1[k]\cdot X_2^*[k]|}\right\} \end{aligned} \tag{5.24}$$

where $\mathcal{F}^{-1}$ is the Inverse Fourier Transform, X_1 and X_2 are microphones x_1 and x_2 in the frequency domain, k is the frequency bin number and $|*|$ denotes the complex conjugate. Delay estimation is then achieved by

$$\tau_1 = \arg\max_n \Psi_{P12}[n] \tag{5.25}$$

$$\tau_2 = \arg\max_n \Psi_{P21}[n]. \tag{5.26}$$

This will be accurate but is only correct for the two source, two microphone case and is performing the same calculation twice.

Other methods for multiple delay estimation exist. The DUET method of BSS is able to calculate multiple delays, but it relies on the input sources having W-disjoint orthogonality, meaning they do not overlap in frequency at any given time. It is also very sensitive to noise and reverberation, which affects the quality of delay estimation. The DUET method also requires that the microphones be close together and it is only useful for multiple sources and two microphones. This is because the distance between the microphones is determined by the highest frequency in the audio sample. If the highest frequency is assumed to be 16kHz there can be a maximum distance of 2.15cm [Yilmaz and Rickard, 2004], which is a significant constraint, especially if estimating delays of spot microphones, as instruments will be placed much further apart.

There is also research in [Kwon et al., 2010] that suggests a method for multiple source delay estimation with the GCC-PHAT, and looks at extracting data from the GCC-PHAT to ascertain characteristics of the sources.

5.5.1 Multiple source GCC-PHAT

We propose a method where multiple delays can be estimated from a single GCC-PHAT calculation making use of redundant information that is usually ignored. The proposed multiple source GCC-PHAT is able to calculate relative delays for cases where $L >= M$, where L is the number of sources and M is the number of microphones, whereas the single source method is not. The proposed multiple source method also does not require W-disjoint orthogonality; both sources can be active. Therefore they can be highly correlated and the delays can still be calculated.

If we take into account that when using the GCC-PHAT, only delays of $\pm N/2$ can be estimated and that the GCC-PHAT can estimate negative delays, we can use the position of the L maximum peaks to estimate multiple delays. This can be achieved by either knowing the number of sources or by peak picking. If it is known that $L = M$ then the number of sources will be known.

Figure 5.4 shows the output of a GCC-PHAT calculation in the two source, two microphone case with the delays labelled. In this case the estimation of the delays using multiple peaks is described as

$$\tau_1 = \arg\max_n \left[\Psi_P[0], \ldots, \Psi_P[N/2]\right] \tag{5.27}$$

$$\tau_2 = \frac{N}{2} - \arg\max_n \left[\Psi_P[N/2+1], ..., \Psi_P[N-1]\right]. \tag{5.28}$$

We will use this for performing the delay estimation in the centred CTRANC for the two source, two microphone case in the remainder of this chapter.

If the configuration is extended to the M microphone and L source case the technique is the same as in (5.27) and (5.28) but for L peaks. If a peak occurs in the first half of the function, the delay is calculated by (5.27). If it occurs in the second half, it is calculated by (5.28). This is repeated up to the number of sources.

The multiple source GCC-PHAT provides other information about the sources. A peak that occurs at the 0 or $N-1$ position is caused by a source that is equidistant from both microphones. Figure 5.5a shows the position of simulated sources and microphones and Figure 5.5b shows the corresponding GCC-PHAT between the microphones. A peak that occurs in the first half of the output function is caused by a source positioned to the left of the centre line between the micro-

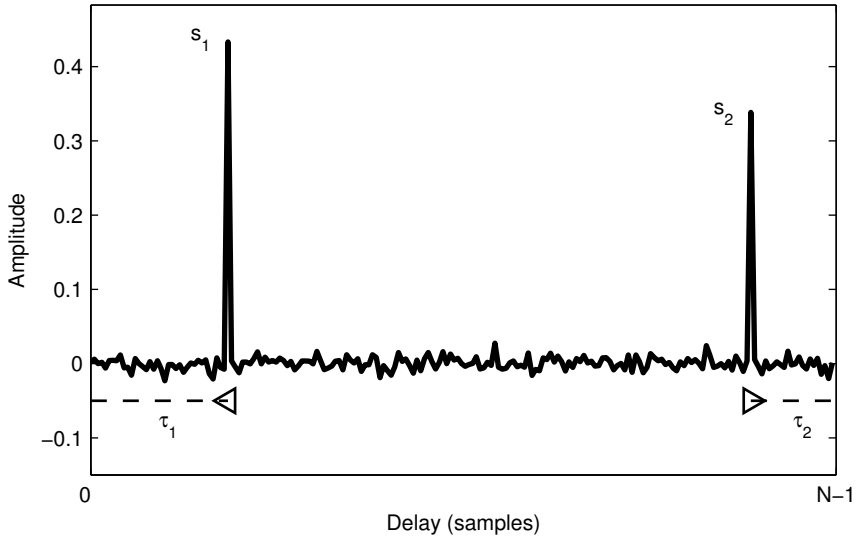


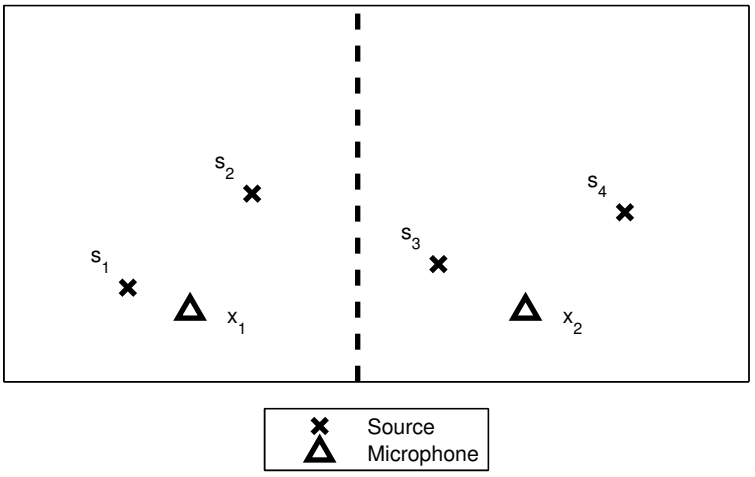
Figure 5.4: Output of the GCC-PHAT where two sources are present with the delays labelled.

phones and a peak that occurs in the second half of the output will be caused by a source to the right.

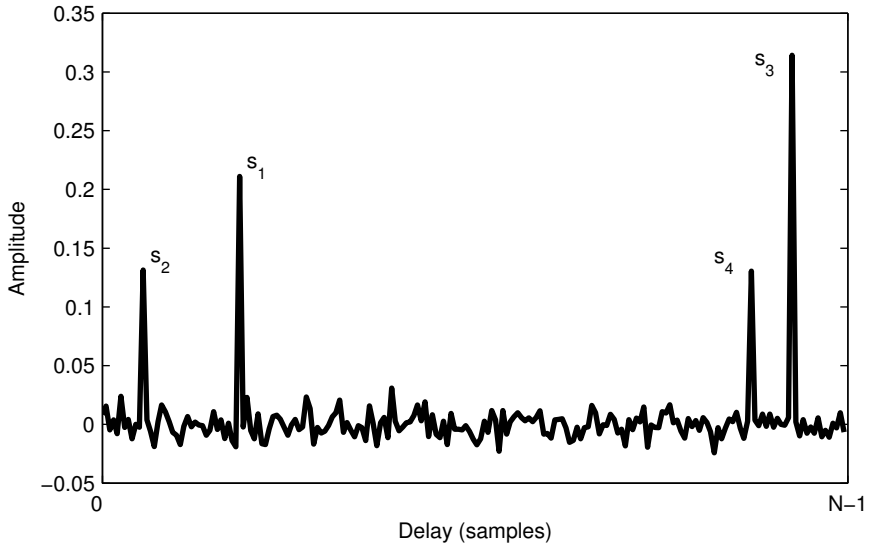
The amplitude of the peak also determines the relative distance of each source to the microphones. The peak with the highest amplitude will be caused by a source placed closest to the microphones, and the smallest caused by a source placed furthest away.

For example in Figure 5.5a, s_1 is closest to x_1, positioned to the left of the centre (dashed) line. This is demonstrated in the GCC-PHAT function in Figure 5.5b where s_1 in positioned in the first half of the function with a large amplitude.

After the multiple delays have been calculated, it is desirable to know which delays correspond to which sources. For this, a simple estimation of the relative placement of sources and/or distance from microphones is required to assign each estimated delay to the correct source.



(a) Simulated multiple microphone, multiple source configuration.



(b) Corresponding GCC-PHAT output for 5.5a.

Figure 5.5: Sample layout of sources and microphones (5.5a) and the resulting GCC-PHAT function (5.5b) showing how the amplitude and position of the peaks is related to the position of the sources.

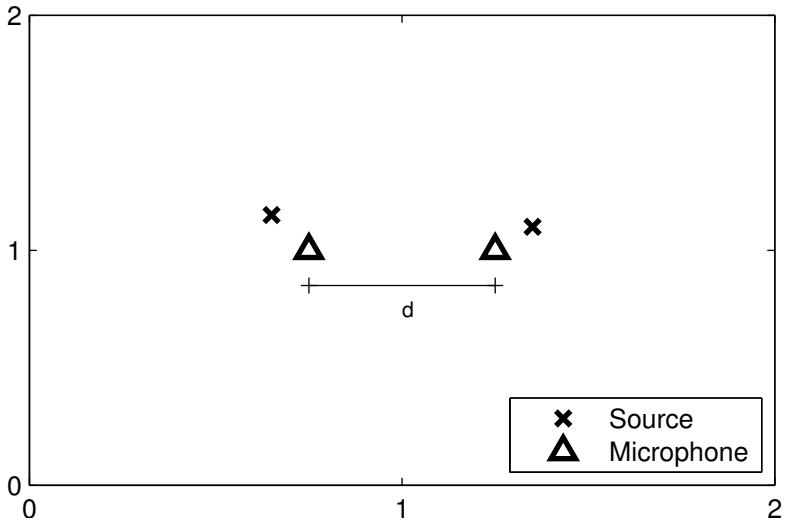


Figure 5.6: Simulation microphone and source layout where $d = 0.5$m.

5.6 Evaluation

We performed an evaluation of the proposed centred CTRANC method to determine how it compared to similar methods in the two source, two microphone case. The methods under test were CTRANC, centred CTRANC, DUET method of source separation [Jourjine et al., 2000] and the Multichannel Wiener Filter (MCWF) method [Kokkinis et al., 2011].

CTRANC and centred CTRANC methods were optimised to produce the best results by selecting a suitable value for the adaption step, μ and the error distance D. A framesize of 2048 samples was used for the CTRANC methods. The DUET and MCWF methods were used with parameters suggested by the creators of each method.

5.6.1 Simulation experimentation

The methods were first compared using microphone signals in simulated anechoic conditions. The source and microphones were positioned as in Figure 5.6. The sources were placed between 10cm and 12cm from the microphones. Delay and gain was applied according to the positions using the inverse square law and the delay estimated from the speed of sound in air at 20^{o}C.

The input sources were a clean guitar and male vocal, which is a common two source, two microphone configuration. The distance d was increased from 10cm to 5m, producing different values for delay and gain. The relative position of each source to each microphone remained the same.

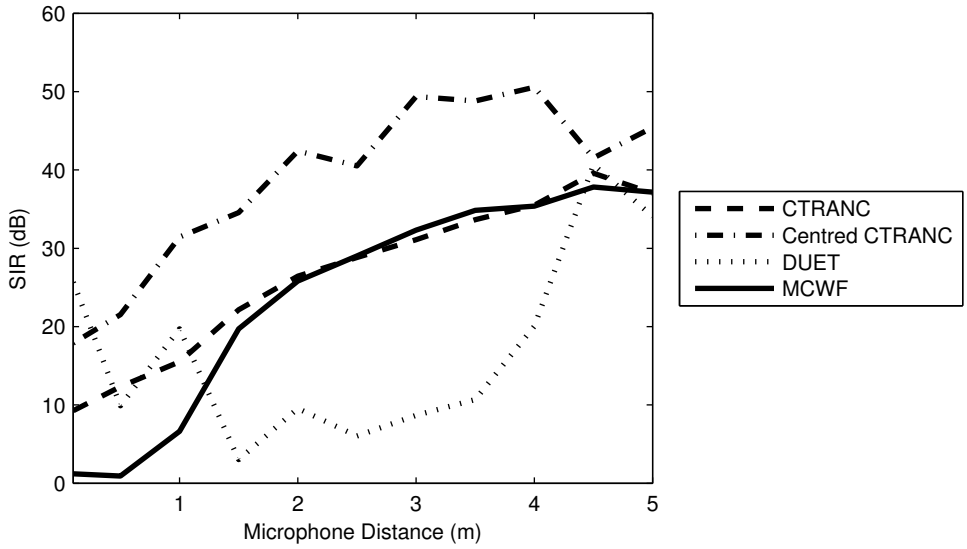


Figure 5.7: Signal-to-interference ratio of each method at each iteration of microphone distance for the simulated case.

5.6.2 Results

Each microphone was then processed using the BSS-EVAL Matlab toolbox [Vincent et al., 2006; Févotte et al., 2005] which is used to objectively evaluate BSS methods and is applicable in this case. The toolbox returns signal-to-interference (SIR), signal-to-artefact (SAR) and signal-to-distortion (SDR) ratios. SIR is used to evaluate how well the interfering source has been suppressed. SAR shows the level of artefacts that have been introduced to the target signal. SDR is used as a global measure which incorporates both [Vincent et al., 2006]. We show the results for microphone x_1 where s_1 is the target signal and s_2 is the interfering signal.

Figure 5.7 shows the SIR for each method at each microphone distance of d. The centred CTRANC resulted in the greatest values of SIR for all values of d over 0.1m, offering a maximum improvement over the CTRANC of 18.2dB. In the $d = 0.1$m case DUET produced the greatest SIR at 25.6dB while the centred CTRANC produced an SIR of 17.9dB. It was expected that the DUET method may perform well for small values of d since it can perform source separation at small distances.

The MCWF method assumes each microphone is an approximation of the ideal impulse response of the direct sound path. If the interference is of a high enough amplitude, this assumption will no longer hold. The CTRANC resulted in greater SIR at low distances compared to the MCWF because of this. The maximum difference in SIR between the CTRANC and MCWF was 11.4dB at

0.5m but over $d = 2$m the results were very similar with a mean difference in SIR of just 0.75dB.

Figure 5.8 shows the SAR for each method. Although DUET performed best when tested for SIR at $d = 0.1$m, Figure 5.8 shows the centred CTRANC had a higher value of SAR at the same distance with an SAR of 12.3dB compared to 7.3dB for DUET. For all distances above $d = 0.5$m the DUET performed consistently worse out of all methods tested. This shows that the DUET method introduced artefacts to the processed signal. The other methods were ranked with centred CTRANC performing the highest followed by the MCWF method and then the CTRANC. The maximum improvement in SAR by using the centred CTRANC compared to the CTRANC is 13.4dB.

Methods based on adaptive filters will generally not add a high level of additional artefacts to the target source since they attempt to subtract the interfering source in the time domain. In live sound, this is desired as it would be preferable to remove some of the interference but leave the target signal intact rather than completely remove the interference but heavily distort the target signal.

The centred CTRANC introduced the least amount of artefacts because the estimated filter will have only a few coefficients. This shows that for the CTRANC method the artefacts come from errors in the filter.

Figure 5.9 shows the SDR for each method and reflects the results seen in both Figure 5.7 and 5.8. The centred CTRANC resulted in the greatest SDR for all distances. At the shortest distance, DUET outperformed all other methods apart from centred CTRANC but then dropped in performance as distance increased. The CTRANC also performed more highly than the MCWF method at up to $d = 1.0$m but then the MCWF increased in performance above this.

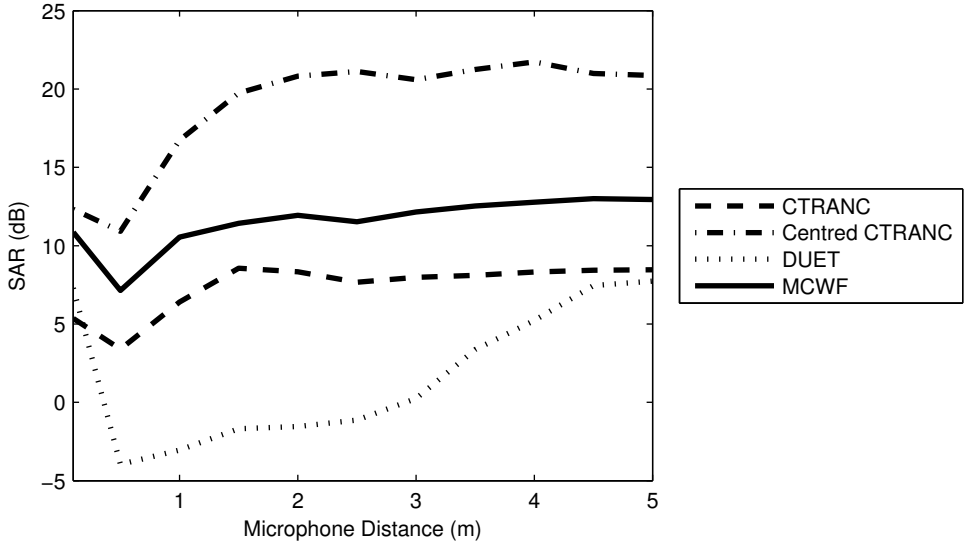


Figure 5.8: Signal-to-artefact ratio of each method at each iteration of microphone distance for the simulated case.

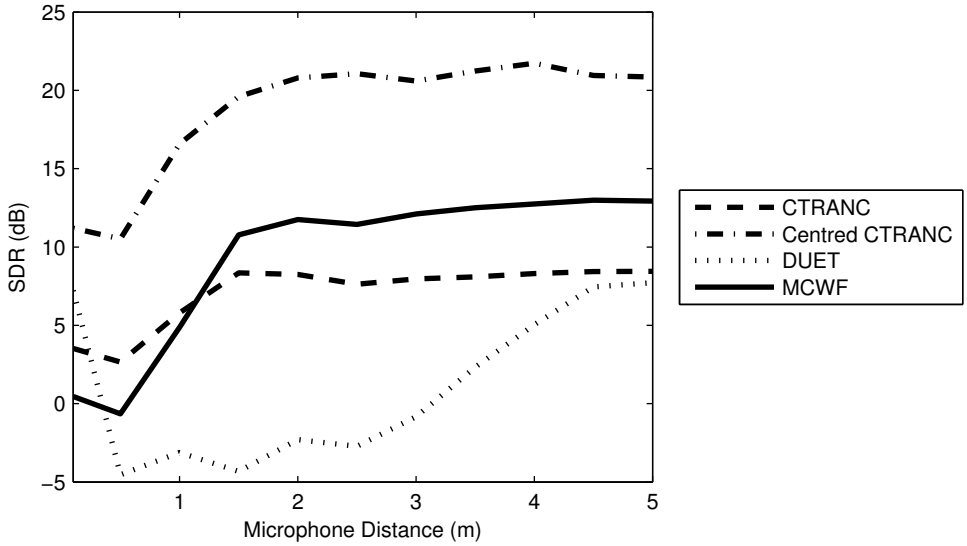


Figure 5.9: Signal-to-distortion ratio of each method at each iteration of microphone distance for the simulated case.

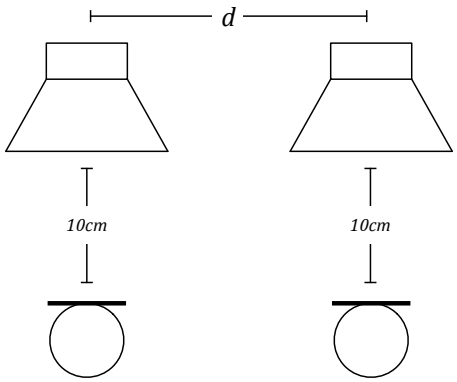


Figure 5.10: Layout of speakers and microphones in the test recordings.

5.6.3 Real recordings

The methods were also tested on real recordings to establish each method's effectiveness in a convolutive environment. Recordings were made using two Genelec 8040 loudspeakers and two AKG C414 microphones in the Listening Room at Queen Mary, University of London with an approximate RT30 of 0.2s, where RT30 is the time taken for the amplitude of the reverberation to drop below 30dB. The loudspeakers were spaced from 10cm to 100cm at 10cm intervals while the microphones were always placed 10cm from each speaker, with an error of $\pm$ 1cm as in Figure 5.10. This distance was chosen to simulate a close microphone configuration. It is not assumed the layout is symmetric.

5.6.4 Results

As with the simulation, the SIR, SAR and SDR for each method and value of d was calculated.

Figure 5.11 shows the SIR for each method. In this case, CTRANC performed greater than the centred CTRANC at all distances with a maximum difference in SIR of 20.6dB. The DUET method also performed more highly than the centred CTRANC at distances above 35cm. The MCWF performed similarly to the CTRANC method, slightly outperforming it for distances between 40 and 50cm, with an overall mean difference in SIR between the MCWF and CTRANC of 2.8dB and a maximum of 5.3dB.

The reason for this is that the centred CTRANC only updates a small range of coefficients around the direct bleed source. This will cause some improvement in SIR compared to the unprocessed microphone signal since some of the direct bleed is reduced in amplitude, but it will not update coefficients related to the reverberation of the microphone bleed. Therefore the reverberation from the

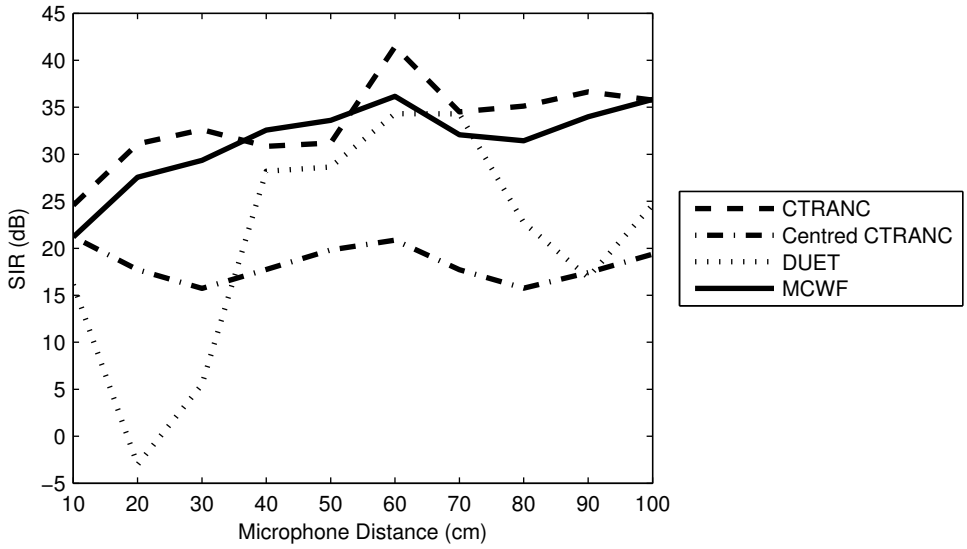


Figure 5.11: Signal-to-interference ratio of each method at each iteration of microphone distance for the real microphone recording.

microphone bleed is retained.

Using a higher value of D may improve this, but by increasing D the computational cost increases and the advantages over CTRANC diminish. The MCWF method performed only slightly lower than the traditional CTRANC method. In contrast to the simulation experiments, the MCWF method performed more consistently with real recordings. The DUET method proved to be more successful at some lengths of d but was not consistent over all the distances tested.

Figures 5.12 and 5.13 show the SAR and SDR for the real recording audio. The results shown in each figure were very similar. As seen in the simulations, the DUET method added additional artefacts and performed consistently the lowest over all distances. The centred CTRANC performed greatest overall when measuring SAR and SDR with a maximum difference to the CTRANC of 8.1dB for SAR and 6.6dB for SDR. This was consistent with the results seen in the simulation tests. Although CTRANC was shown to perform well by the SIR measure, but performed worse than the centred CTRANC in measures of SAR and SDR.

The MCWF method performed worse than the centred CTRANC method in terms of SAR and SDR with a mean difference in SAR of 3.8dB and mean difference in SDR of 2.3dB but with slightly higher values of SAR and SDR than the traditional CTRANC for all distances with a mean difference of 2.5dB for both SAR and SDR.

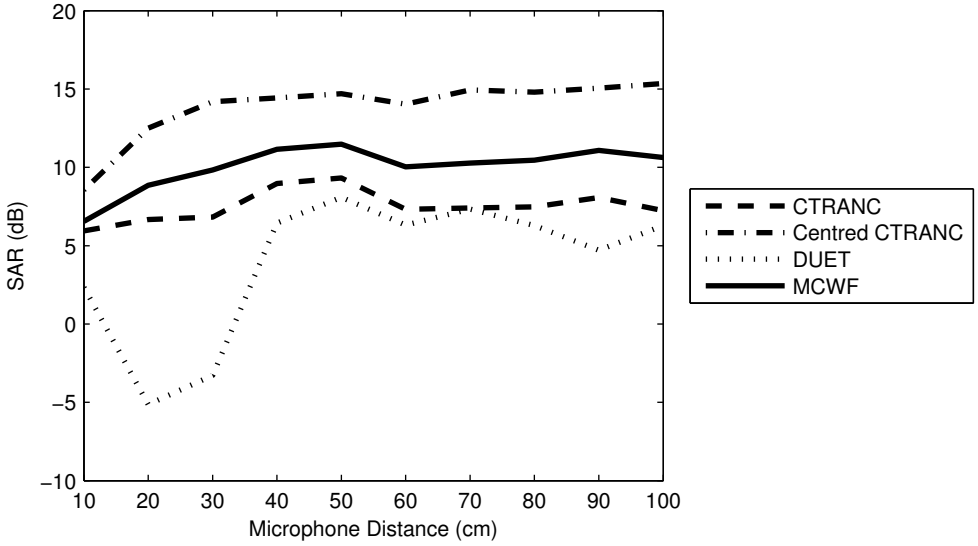


Figure 5.12: Signal-to-artefact ratio of each method at each iteration of microphone distance for the real microphone recording.

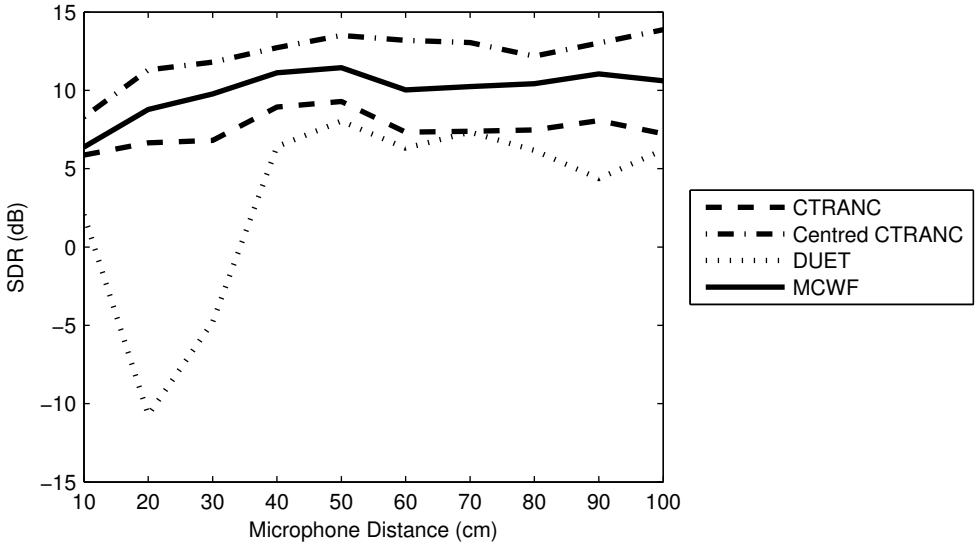


Figure 5.13: Signal-to-distortion ratio of each method at each iteration of microphone distance for the real microphone recording.

5.7 Discussion and conclusions

In this chapter a method for microphone bleed reduction in the two source, two microphone case has been proposed that combines centred adaptive filters with the CTRANC method of noise cancellation. The CTRANC has not been applied to music signals in the literature and it has not previously been combined with centred adaptive filters.

The adaptive filters are centred using a novel adaptation of the GCC-PHAT method of delay estimation, described in the previous chapter, taking into account multiple peaks in the output function.

The proposed method, centred CTRANC, outperformed other methods for interference reduction in the simulated anechoic case and improved the SIR over the CTRANC by 18.2dB with fewer additional artefacts than the other methods tested. In simulated conditions the centred CTRANC offered an increase in SDR of 7.7dB at the smallest distance tested up to 12.4dB at the largest distance tested compared to the original CTRANC method. The centred CTRANC also resulted in a maximum improvement in SIR of 24.8dB compared to the Multichannel Wiener Filter method.

Centred CTRANC was shown to be outperformed with regards to SIR by the original CTRANC system in real recordings by a mean SIR of 15dB over all distances tested. But the centred CTRANC was shown to introduce fewer artefacts with a mean improvement in SAR of 6.3dB compared to the original CTRANC. The centred CTRANC is therefore suited to low reverberation environments.

The efficacy of the centred CTRANC was not affected by the level of the interference but by the environment which the sources and microphones are placed, and the reverberation and noise present. Therefore it is currently best suited to close microphone applications.

We have shown that the centred CTRANC struggles in reverberant conditions as the centring restricts the amount of reverberation that can be estimated in the impulse response as it is truncated. In the next chapter we apply CTRANC in the frequency domain to improve results and efficiency. We also extend CTRANC to the L source, M microphone case in both determined and overdetermined configurations with a variety of sources.

Chapter 6

Overdetermined microphone bleed reduction using selective FDCTRANC

In the previous chapter we examined a two source, two microphone method of microphone bleed reduction using CTRANC combined with centred adaptive filters. The main problem with this method is the computational cost and the performance in reverberant conditions. In this chapter we propose performing CTRANC in the frequency domain, FDCTRANC, to improve computation and performance. We show that performing CTRANC in the frequency domain uncovers additional problems, which were not at first apparent. We propose performing FDCTRANC iteratively to reduce the effect of these problems. The proposed method is then compared to similar methods, including the centred CTRANC presented in the previous chapter, in a subjective listening test. FDCTRANC was shown to perform well at target preservation while reducing microphone bleed amplitude.

We also extend FDCTRANC to the overdetermined case, where there are more microphones than sources. We still assume each microphone is positioned to reproduce a single source and is therefore closest to one source. Applying FDCTRANC to the overdetermined case requires establishing whether any of the microphones are reproducing the same target source and performing intelligent bleed reduction dependent on this. This is achieved by comparing similarity between microphone signals using correlation in the frequency domain. In the overdetermined case the selective FDCTRANC was shown to outperform the standard FDCTRANC in all overdetermined cases under test.

6.1 Determined CTRANC

In the previous chapter we outlined using CTRANC for microphone bleed reduction in the two source, two microphone case. This can be extended to the multiple source, multiple microphone case with L sources and M microphones.

In the determined case, where $L = M$, (5.8) and (5.9) are rewritten as

$$\hat{x}_m[n] = x_m[n] - \sum_{\substack{l=1 \\ l \neq m}}^{L} \mathbf{w}_{lm}^T[n]\mathbf{x}_l[n] \tag{6.1}$$

with the filters updated by

$$\mathbf{w}_{lm}[n+1] = \mathbf{w}_{lm}[n] + \mu\hat{\mathbf{x}}_l[n]\hat{x}_m[n] \tag{6.2}$$

where

$$\mathbf{x}_l = [x_l[n], \ldots, x_l[n-N+1]] \tag{6.3}$$

where x_l is the current interference microphone, $l = 1, \ldots, L$, x_m is the current target microphone, $m = 1, \ldots, M$ and $\mathbf{w}_{lm}$ contains the N filter coefficients.

Traditionally the adaptive filter weights are updated every time step n. For efficiency the scheme can be altered to only update the filter weights every block k of N time steps, replacing the timestep n with a reference to block i to become $n = iN + n_i$ where $n_i = 1, \ldots, N$. This is known as block LMS (BLMS) [Haykin, 2001]. For CTRANC using BLMS the processed microphone signals are updated by

$$\hat{x_m}[iN+n_i] = x_m[iN+n_i] - \sum_{\substack{l=1 \\ l \neq m}}^{L} \mathbf{w}_{lm}^T[iN+n_i]\mathbf{x_l}[iN+n_i] \tag{6.4}$$

which is equivalent to (6.1). The filter weights are updated by

$$\mathbf{w}_{lm}[k+1] = \mathbf{w}_{lm}[k] + \mu\sum_{l=1}^{L} \hat{\mathbf{x}}_l[kN+n_i]\hat{x}_m[kN+n_i]. \tag{6.5}$$

But $\hat{x}_m$ is still updated as (6.1), which can cause computation issues.

It should also be noted that by scaling the CTRANC method to the determined case, the number of adaptive filters in the scheme is $A = M(M-1)$. Thus increasing the number of microphones significantly increases computational cost.

6.2 FDCTRANC

6.2.1 Derivation

The computational cost of CTRANC can be furthered improved by implementing the adaptive filters in the frequency domain, which we will refer to as FDCTRANC [Haykin, 2001; Shynk, 1992]. The convolution of the filter with the

incoming signal in (6.4) and the correlation of the filter and incoming signal in (6.5) are computed using the Fast Fourier Transform and are only estimated every N time steps.

Performing CTRANC in the frequency domain is briefly mentioned in [Lepaulox et al., 2009] but the chosen scheme is not stated. Here we present the overlap-add scheme for adaptive filters as described by Shynk [1992], adapted for CTRANC.

Each interfering microphone signal is defined as

$$\mathbf{X}'_l[k] = diag(\mathcal{F}[x_l[kN], \ldots, x_l[kN+N-1], 0, \ldots, 0]^T) \tag{6.6}$$

where $\mathcal{F}$ denotes the Discrete Fourier Transform. Due to the overlap add constraints, this is then processed as

$$\mathbf{X}_l[k] = \mathbf{X}'_l[k] + \mathbf{J}\mathbf{X}'_l[k-1] \tag{6.7}$$

where $\mathbf{J} = diag(1, -1, \ldots, -1)$.

The current filter weights are applied to each interfering source as

$$\mathbf{\Phi}[k] = \sum_{\substack{l=1 \\ l \neq m}}^{L} \mathbf{X}_l[k]\mathbf{W}_{lm}[k] \tag{6.8}$$

and the processed target microphone signal from (6.4) are updated in FDCTRANC by

$$\hat{\mathbf{x}}_m[k] = \mathbf{x}_m[k] - \mathbf{k}\mathcal{F}^{-1}\mathbf{\Phi}[k] \tag{6.9}$$

where $\mathcal{F}^{-1}$ denotes the Inverse Discrete Fourier Transform and the sectioning constraints are $\mathbf{k} = [\mathbf{0}_N \mathbf{1}_N]$. The interfering microphone filters from (6.5) are updated in the frequency domain by

$$\mathbf{W}_{lm}[k+1] = \mathbf{W}_{lm}[k] + \mathcal{F}\mathbf{g}\mathcal{F}^{-1}\boldsymbol{\mu}[k]\mathbf{X}_l[k]^H\hat{\mathbf{X}}_m[k] \tag{6.10}$$

where

$$\mathbf{g} = \begin{bmatrix} \mathbf{1}_N & \mathbf{0}_N \\ \mathbf{0}_N & \mathbf{0}_N \end{bmatrix} \tag{6.11}$$

and

$$\hat{\mathbf{X}}_m[k] = \mathcal{F}\mathbf{k}^T\hat{\mathbf{x}}_m[k] \tag{6.12}$$

and the frequency dependent step size $\boldsymbol{\mu}$ is calculated by

$$\boldsymbol{\mu}[k] = \mu \cdot diag(P^{-1}[k]) \tag{6.13}$$

where

$$P[k] = \gamma P[k-1] + (1-\gamma)|X_l[k]|^2 \tag{6.14}$$

and where γ is a forgetting factor. In all equations $X_l = \hat{X}_l$ when it exists.

Apart from improving computational cost, frequency domain implementation of the adaptive filters also allows the addition of a frequency dependent step size, calculated by (6.13). This allows the step size of each separate frequency bin to adjust so that the filters will update at a much slower rate in periods of silence, reducing the amount of errors. It also allows only significant spectral information to be used in the update of the filter weights, resulting in more accurate and faster convergence [Soo and Pang, 1990].

6.2.2 Issues

Implementing CTRANC with a block-based approach highlights problems which have not previously been addressed in the literature. This is best explained by using the two source and two microphone model in anechoic conditions, outlined in Section 2.2.5 and repeated here

$$x_1[n] = \alpha_{11} s_1[n-\tau_{11}] + \alpha_{21} s_2[n-\tau_{21}] \tag{6.15}$$

$$x_2[n] = \alpha_{22} s_2[n-\tau_{22}] + \alpha_{12} s_1[n-\tau_{12}] \tag{6.16}$$

where x_1 and x_2 are microphone signals at timestep n, s_1 and s_2 are the sound sources, τ_{lm} is the delay of source l to microphone m and α_{lm} is the amplitude change due to distance of source l to microphone m.

Applying the time domain CTRANC from (5.8), $\hat{x}_1$ is estimated by

$$\hat{x}_1[n] = x_1[n] - \mathbf{w}_{21}^T[n]\mathbf{x}_2[n] \tag{6.17}$$

where $\mathbf{w}_{21}$ is a delayed Dirac delta to align s_2 in x_1 and x_2. Therefore $\mathbf{x}_2$ will be delayed by $\tau_{21} - \tau_{22}$ and the gain reduced by $\alpha_{22} - \alpha_{21}$. In terms of s_1 and s_2 this is

$$\begin{aligned}
&\mathbf{w}_{21}^T[n]\mathbf{x}_2[n] \\
&\quad = (\alpha_{22} - (\alpha_{22} - \alpha_{21}))s_2[n - (\tau_{22} + (\tau_{21} - \tau_{22}))] \\
&\qquad + (\alpha_{12} - (\alpha_{22} - \alpha_{21}))s_1[n - (\tau_{21} + (\tau_{21} - \tau_{22}))] \qquad (6.18) \\
&\quad = \alpha_{21} s_2[n - \tau_{21}] + \alpha'_{12} s_1[n - \tau'_{12}] \qquad (6.19)
\end{aligned}$$

where

$$\alpha'_{12} = \alpha_{12} - (\alpha_{22} - \alpha_{21}) \tag{6.20}$$

$$\tau'_{12} = \tau_{12} + (\tau_{21} - \tau_{22}) \tag{6.21}$$

therefore (6.17) in terms of s_1 and s_2 is

$$\begin{aligned}\hat{x}_1[n] =& \alpha_{11}s_1[n-\tau_{11}] + \alpha_{21}s_2[n-\tau_{21}] - \alpha_{21}s_2[n-\tau_{21}] \\ & - \alpha'_{12}s_1[n-\tau'_{12}] \qquad (6.22)\\ =& \alpha_{11}s_1[n-\tau_{11}] - \alpha'_{12}s_1[n-\tau'_{12}]. \qquad (6.23)\end{aligned}$$

So the interfering source s_2 has been cancelled out but $\hat{x}_1$ contains s_1 summed with a delayed version of itself, which will cause comb filtering.

Continuing with CTRANC, $\hat{x}_2$ from (5.9) is then estimated by

$$\hat{x}_2[n] = x_2[n] - \mathbf{w}_{12}^T[n]\hat{\mathbf{x}}_\mathbf{1}[n] \tag{6.24}$$

and $\mathbf{w}_{12}$ delays $\hat{\mathbf{x}}_\mathbf{1}$ by $\tau_{12} - \tau_{11}$ and reduces the amplitude by $\alpha_{11} - \alpha_{12}$ to align s_1, so in terms of s_1 and s_2 this eventually becomes

$$\mathbf{w}_{12}^T[n]\hat{\mathbf{x}}_\mathbf{1}[n] = \alpha_{12}s_1[n-\tau_{12}] - \alpha''_{12}s_1[n-\tau''_{12}] \tag{6.25}$$

where

$$\alpha''_{12} = \alpha'_{12} - (\alpha_{11} - \alpha_{12}) \tag{6.26}$$

$$\tau''_{12} = \tau'_{12} + (\tau_{12} - \tau_{11}) \tag{6.27}$$

therefore

$$\begin{aligned}\hat{x}_2[n] =& \alpha_{22}s_2[n-\tau_{22}] + \alpha_{12}s_1[n-\tau_{12}] - \alpha_{12}s_1[n-\tau_{12}] \\ & + \alpha''_{12}s_1[n-\tau''_{12}] \qquad (6.28)\\ =& \alpha_{22}s_2[n-\tau_{22}] + \alpha''_{12}s_1[n-\tau''_{12}]. \qquad (6.29)\end{aligned}$$

This leaves $\hat{x}_1$ as a comb filtered version of s_1 and $\hat{x}_2$ as a summation of s_2 and s_1 where s_1 has reduced in amplitude.

So with this scheme, only x_2 has had the amplitude of the microphone bleed reduced while minimising the distortion to the target source, in this case s_2. It would be possible to run the scheme again from the beginning interchanging x_1 and x_2 to reduce the amplitude of the bleed in x_1 but as mentioned previously the number of adaptive filters is related to the number of microphones and this would further increase computational cost.

6.2.3 Iterative FDCTRANC

We propose performing subsequent iterations of the algorithm to use $\hat{x}_2$ as a more accurate representation of the interfering source s_2 in order to reduce the amplitude of the bleed from x_1. By doing this, the same scheme can be applied to both microphones at the same time, instead of applying the same algorithm to each separately. The ultimate goal is to reduce the comb filtering that occurs on $\hat{x}_1$, thus improving the bleed reduction in both x_1 and x_2.

In the proposed scheme, the next stage of the algorithm after (6.24) is to repeat (6.17) but where x_2 is replaced by $\hat{x}_2$, as follows

$$\hat{x}'_1[n] = x_1[n] - \mathbf{w}'^T_{21}[n]\hat{\mathbf{x}}_2[n] \tag{6.30}$$

so

$$\mathbf{w}'^T_{21}[n]\hat{\mathbf{x}}_2[n] = \alpha_{21}s_2[n-\tau_{21}] + \alpha'''_{12}s_1[n-\tau'''_{12}] \tag{6.31}$$

where

$$\alpha'''_{12} = \alpha''_{12} - (\alpha_{22} - \alpha_{21}) \tag{6.32}$$

$$\tau'''_{12} = \tau''_{12} + (\tau_{21} - \tau_{22}) \tag{6.33}$$

therefore in terms of s_1 this becomes

$$\begin{aligned}\hat{x}'_1[n] =& \alpha_{11}s_1[n-\tau_{11}] + \alpha_{21}s_2[n-\tau_{21}] - \alpha_{21}s_2[n-\tau_{21}] \\ & - \alpha'''_{12}s_1[n-\tau'''_{12}] \qquad (6.34)\\ =& \alpha_{11}s_1[n-\tau_{11}] - \alpha'''_{12}s_1[n-\tau'''_{12}] \qquad (6.35)\end{aligned}$$

where $\alpha'''_{12} < \alpha''_{12}$ and $\tau'''_{12} > \tau''_{12}$. Thus the comb filtering effects are reduced as the gain of the delayed source is reduced. As this is now a more accurate representation of s_1, (6.24) can be recalculated as

$$\hat{x}'_2[n] = x_2[n] - \mathbf{w}^T_{12}[n]\hat{\mathbf{x}}'_1[n] \tag{6.36}$$

and therefore the amplitude reduction of the bleed is greater in $\hat{x}'_2$ than in $\hat{x}_2$. (6.17) and (6.24) can be repeated subsequent times to further improve the reduction, although each iteration increases the number of adaptive filters.

It is important to note that each iteration of (6.17) and (6.24) requires different filter coefficients for the best performance since, for example, it is possible that $\mathbf{w}_{21} \neq \mathbf{w}'_{21}$. This is because as the scheme progresses more reduction in the bleed amplitude is achieved and the amplitude of the filter for each iteration may be different.

Figure 6.1 shows a block diagram of the proposed method showing two iter-

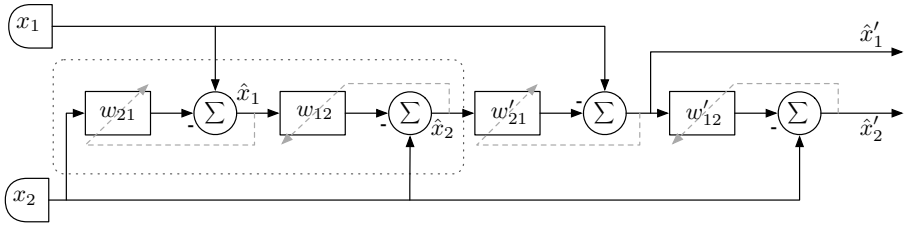


Figure 6.1: A block diagram of the proposed FDCTRANC method of interference reduction. The repeated iteration step is highlighted.

ations. In this case w_{12} and w_{12} are adaptive filters in the first iteration while w'_{12} and w'_{12} are adaptive filters in the second.

We have shown the proposed scheme in the two source, two microphone case but this can also be scaled to when the number of sources is greater than two. In the general case, (6.1) becomes

$$\hat{x}_m[n] = x_m[n] - \sum_{\substack{l=1 \\ l \neq m}}^{L} \mathbf{w}_{lmi}^T[n]\mathbf{x}_l[n] \tag{6.37}$$

and (6.2) becomes

$$\mathbf{w}_{lmi}[n+1] = \mathbf{w}_{lmi}[n] + \mu\hat{\mathbf{x}}_l[n]\hat{x}_m[n] \tag{6.38}$$

where $i = 1, \ldots, I$ and I is the number of iterations and $\mathbf{x}_l = \hat{\mathbf{x}}_l$ when it has been calculated. Running the scheme in this way, assuming $L = M$, the number of adaptive filters per iteration is $A = M(M-1)I$. Therefore the results will be improved but the number of adaptive filters increases.

6.2.4 Number of iterations

We analysed the proposed iterative FDCTRANC algorithm on simulated microphone signals to ascertain the optimal number of iterations for the scheme. The experiment was performed in simulated anechoic conditions using an image source toolbox by Lehmann and Johansson [2008] to generate room impulse responses (RIRs). The room was 5m x 5m x 2.5m in size. The sources were placed at approximately 0.5m intervals and a single microphone was positioned between 0.15m and 0.25m in front of each source to simulate the layout of a real configuration where equally spaced sources and microphones are unlikely. The configuration was tested in the determined case from two to four sources. The maximum source layout can be seen in Figure 6.2. The sources used were a male vocal, an acoustic guitar, a piano and a fiddle, respectively.

The simulated audio was analysed using the BSS-EVAL toolbox, as men-

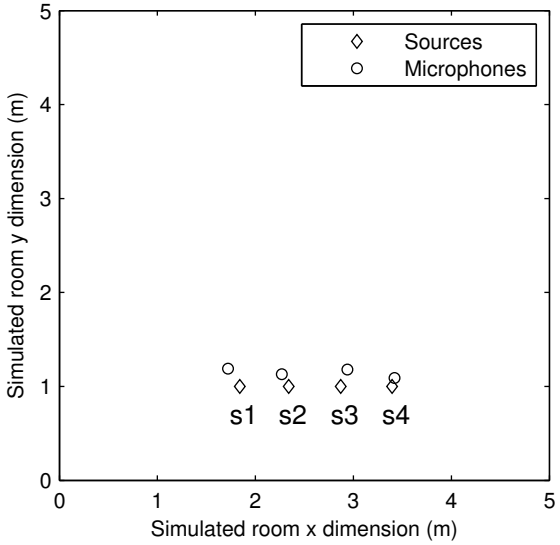


Figure 6.2: Virtual source and microphone layout for analysis of the number of iterations of the FDCTRANC method of bleed reduction.

tioned in the previous chapter. The results are shown in terms of the improvement of Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR) in decibels compared to the original microphone signal. Signal-to-Artefact Ratio (SAR) is not included as the results for SDR and SAR in the previous chapter were similar therefore we chose not to show both in this chapter.

Figure 6.3 shows the results as SDR and SIR improvement compared to the SDR and SIR of the original audio samples in decibels against the number of iterations. There was a clear increase in improvement as the number of iterations increases, with the greatest improvement occurring for four sources. The increase in SDR between one and two iterations can be attributed to the reduced comb filtering of x_1 but the effect begins to diminish after four iterations, which may be due to increased artefacts in the signal. The SIR improved with a steady increase up to three iterations for two and three sources but further increased for four sources.

6.3 Evaluation

The proposed method, iterative FDCTRANC, was evaluated against the basic CTRANC, the centred CTRANC from Chapter 5 and the MCWF, which was also used in the previous chapter. The CTRANC has already been compared to the DUET method of Blind Source Separation in the previous chapter, therefore we chose to focus on noise cancellation methods for this evaluation.

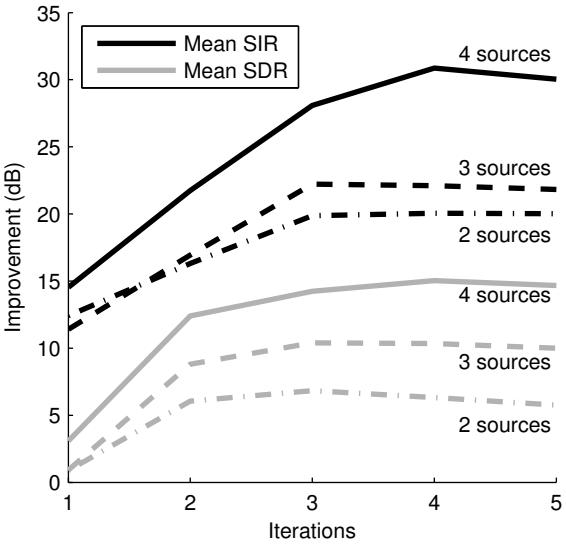


Figure 6.3: Comparison between different number of iterations for different determined microphone configurations showing mean SIR and SDR improvement from the unprocessed microphone signal.

The algorithms were tested in both anechoic and simulated reverberant conditions using RIRs generated using the Lehmann image source MATLAB toolbox. The RT60 was 0.4s in the reverberant case, where RT60 is the time taken for the amplitude of the reverberation to drop below 60dB [Howard and Angus, 2000]. It is defined by the absorption coefficients of the simulated space, generated by the toolbox. The layout was equal for both RIR cases using two sources and two microphones. The room dimensions were 5m x 5m x 2.5m and the sources were positioned at (2.9,1.0,1.3) and (3.4,1.0,1.3), 0.5m apart to simulate a real configuration. The microphones were spaced the same width apart as the sources but positioned at a distance of 0.12m to simulate a close microphone configuration.

The audio was sampled at 44.1kHz. FDCTRANC used a window size of 2048 samples. The basic and centred CTRANC used a window size of 512 samples to reduce the computation time. The MCWF used a window size of 4096, recommended by Kokkinis et al. [2011]. The audio samples were scaled so the RMS of each sample matched the RMS of the original microphone signal. This was done to reduce the perceptual effects of amplitude changes between comparative audio samples.

6.3.1 Subjective evaluation

A subjective listening test was performed to evaluate each method.

Setup

A Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test was conducted, adhering to the ITU standard [International Telecommunication Union, 2003]. This type of test was chosen because it is commonly used to assess blind source separation algorithms [Emiya et al., 2011] and is suitable for assessment of intermediate audio quality [Bech and Zacharov, 2006]. It is also very time efficient and allows a large amount of data to be collected in a shorter period of time than a pairwise comparison test, for example.

The participants were presented with the interface shown in Figure 6.4. In each trial the reference was a simulated microphone signal consisting of a target source combined with an interfering source in either anechoic or reverberant conditions. The target audio sources were a rhythm acoustic guitar, fiddle, bass guitar, slide guitar, male vocal and tin whistle. The interfering sources were a male vocal, tin whistle, electric guitar, slide guitar, piano and organ for each target source respectively. There were 12 trials for each repetition of the test.

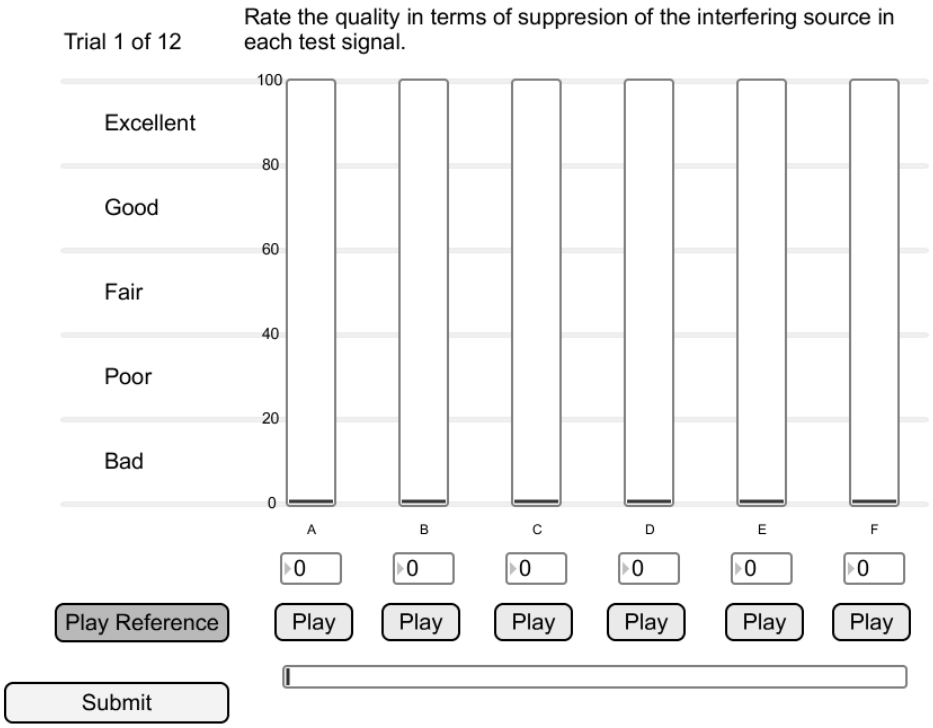


Figure 6.4: User interface for the MUSHRA listening test.

The participants were presented with a bank of six audio samples to rate from 0 (bad) to 100 (excellent) using corresponding sliders. Four of the audio samples

were the reference processed with each of the four bleed reduction methods under test. The reference was also included as one of the audio samples along with an anchor audio sample, which is deliberately processed in such a way that it should be rated the lowest out of all the samples. The audio samples were assigned randomly to the sliders by the software.

The test was conducted twice as the participants were asked to rate each audio sample for two different quality criteria, similar to criteria used by Emiya et al. [2011], compared to the reference.

The participants were firstly asked to rate the sounds in terms of the quality of the suppression of the interference. After all trials were complete, the test was repeated but they were asked to rate the samples in terms of the quality of the preservation of the target source, referring to any additional artefacts they may hear. The listening test was performed in this way to get a perceptual overview of how each algorithm performed and whether bleed reduction can be performed without additional artefacts.

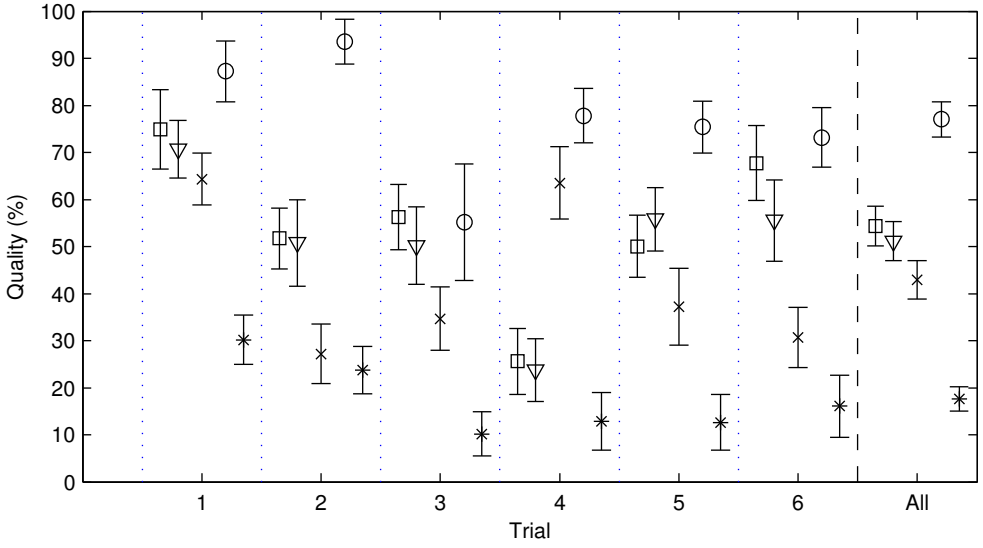
For the different quality assessments two different anchors were used. For the interference rating stage the original simulated microphone signal was used as an anchor. For the artefact rating stage a version of the clean target signal, low pass filtered at 3.5kHz, was used. This is the same as proposed in the original MUSHRA standard [International Telecommunication Union, 2003]. The clean target source was the hidden reference in both cases. Participants were asked to rate at least one audio sample at 100 (excellent), as per the ITU standard. Each stage of the test was approximately 25 minutes in duration with a 5 minute break in between.

There were 15 participants between the ages of 20 and 41. 12 were male and all had critical listening experience. Post screening of the results rejected a participant for the artefact criteria because the anchor had been rated at 100% in a number of trials. All participants' results were used for the interference criteria.

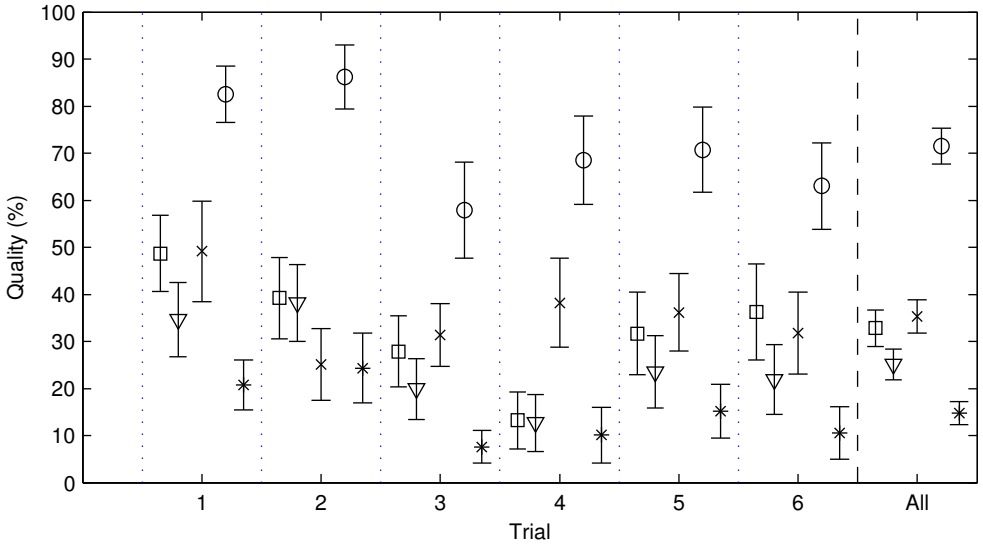
Results

As per the ITU standard, the results are reported as the mean rating of each condition for each trial for both anechoic and reverberant condition. We also show the mean rating of all responses in each trial for each audio sample. The reference was rated as 100 in 96.26% of all trials in all cases and conditions with a minimum rating of 82 in all others therefore the results for the reference are not shown. We show the 95% confidence intervals based on Student's t-distribution [Sporer et al., 2009].

For each trial the data for each algorithm was compared using the Wilcoxon



(a) Anechoic



(b) Reverberant

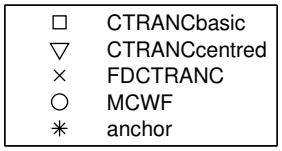


Figure 6.5: Results of the subjective listening test for the interference criteria showing means of all participants for each trial for FDCTRANC, MCWF and anchor with 95% confidence intervals.

rank sum test, as recommended by Nagel et al. [2010], due to the non-parametric nature of the data. In all cases the significance is accurately demonstrated by the confidence intervals, where an overlap of confidence intervals over neighbouring means does not reject the null hypothesis of equal distributions.

Figure 6.5 shows the results for the interference criteria for each trial in both anechoic and reverberant conditions. Overall, it can be seen that the results were trial, and therefore source, dependent.

In the anechoic case, centred CTRANC and CTRANC did not reject the null hypothesis of equal distributions in any trials, therefore they can be considered as having similar performance. The anchor was consistently rated lowest quality in all trials, which was expected. In trial 2 FDCTRANC did not reject the null hypothesis of equal distribution to the anchor. The MCWF was rated highest apart from in trial 3 where the difference between the centred CTRANC, CTRANC and MCWF was not considered significantly different and in trial 6 where the means on the CTRANC and MCWF were also considered statistically similar. The results of FDCTRANC were inconclusive as the performance ranged from being similar to the anchor in trial 2 to being rated just below the MCWF in trial 3.

Across all trials in the anechoic case, the MCWF was rated highest with a mean rating of 77%, followed by the basic and centred CTRANC methods with mean ratings of 54% and 51% respectively, although the difference between the CTRANC methods was not statistically significant. FDCTRANC performed slightly worse than CTRANC based methods at 43%.

In the reverberant case, the centred CTRANC performed significantly worse than the basic CTRANC in trial 1 and 6. This was expected from the results in the previous chapter. The anchor was rated low in all cases. FDCTRANC was rated higher than the basic CTRANC in more trials than in the anechoic conditions.

Across all trials the MCWF method was also rated highest in the reverberant case with a mean rating of 71%. FDCTRANC is then rated similarly to the basic CTRANC with mean ratings of 35% and 33% respectively with the centred CTRANC performing worse with a mean rating of 25%.

Figure 6.6 shows the same data for the artefact criteria. In both the anechoic and reverberant cases FDCTRANC, centred CTRANC and basic CTRANC did not reject the null hypothesis of equal distributions therefore they can be considered to have performed similarly in terms of artefact reduction.

In trial 3 in both cases, the MCWF performed worse than the anchor by 52% in the anechoic case and 54% in the reverberant case. This was because the MCWF is performed in the frequency domain and therefore changes in frequency content are expected. In this particular trial the output of the MCWF audio

was as if it had been processed with a low pass filter, similar to that of the anchor. In all other trials MCWF performed the worst out of the four methods under test.

In trial 2 and trial 6 in both cases the differences between all methods under test were not statistically significant. The FDCTRANC had a mean rating of 84% in the anechoic case and 89% in the reverberant case. The MCWF had a mean rating of 60% in both the anechoic and reverberant cases. Overall CTRANC based and FDCTRANC methods are rated highest with no significant difference between them.

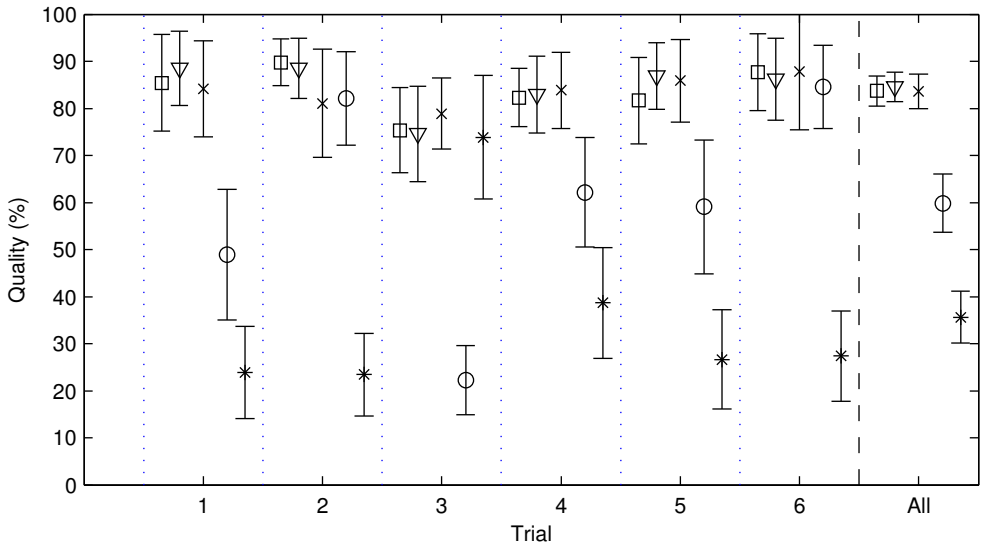
Tables 6.1 and 6.2 report the p-value of the Wilcoxon rank sum test between the anechoic and reverberant criteria of each method under test for each trial and overall for both the interference and artefact criteria. The mean of the anechoic and reverberant condition is shown under each p-value. When $p > 0.05$ it is indicated in bold, which signifies the results reject the null hypothesis of equal distributions and the difference in means can be considered statistically significant.

Table 6.1 shows that on a trial-by-trial basis, the performance of the basic CTRANC and centred CTRANC were dependent on room conditions for the interference criteria, with anechoic conditions achieving better performance. FDCTRANC was less dependent on room conditions in this test, with the performance of 4 of the 6 trials being independent of room conditions. The MCWF was not affected by room conditions in all cases for the interference criteria. In the overall comparison, the FDCTRANC and MCWF methods were considered to be affected by reverberation.

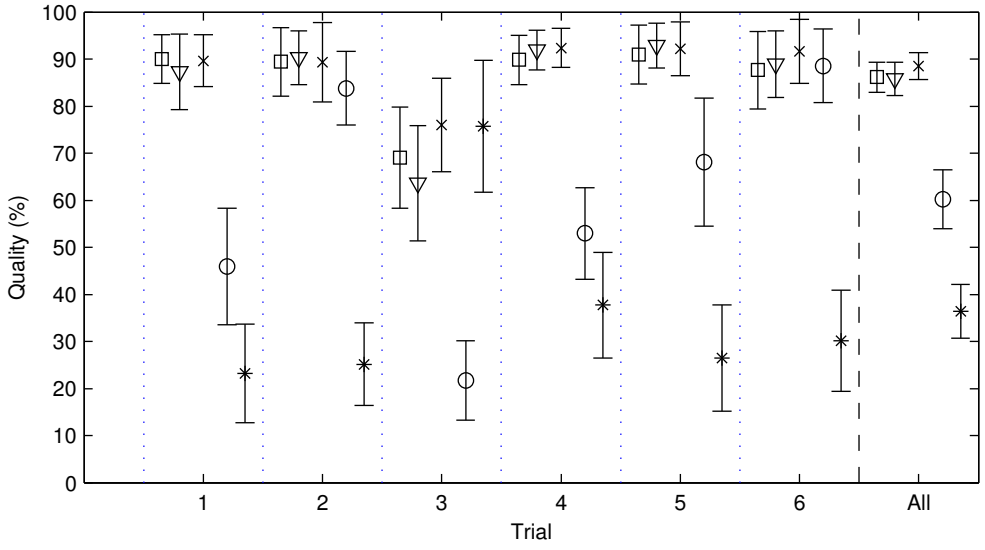
Table 6.2 shows that all methods performed independent of room criteria apart from one trial for the centred CTRANC. Each method exhibited examples where the performance in reverberant conditions is better than that in anechoic conditions. The difference in means of FDCTRANC in anechoic and reverberant conditions are considered statistically significant with $p = 0.048$, although this is very close to the test p-level of 0.05.

Overall, the MCWF achieved the greatest quality of interference reduction but at a cost of increased artefacts in the target signal. This is an expected outcome of Wiener filter implementations [Lim and Oppenheim, 1979]. Although not shown in the results, MCWF also introduced time varying artefacts and in trial 3 in both room conditions the MCWF was confused with the anchor. The MCWF also altered the gain of the target signal, whereas FDCTRANC does not. In this evaluation, all audio samples were normalised for amplitude with the same RMS to test only interference and artefacts, although it can be argued that altering the gain is introducing an artefact.

The proposed FDCTRANC method performed higher than the MCWF in



(a) Anechoic



(b) Reverberant

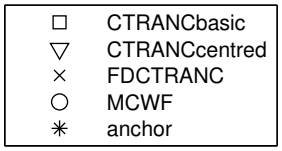


Figure 6.6: Results of the subjective listening test for the artefact criteria showing means of all participants for each trial for FDCTRANC, MCWF and anchor with 95% confidence intervals.

	Trial						
Algorithm	1	2	3	4	5	6	All
CTRANC basic	0 (75,49)	0.025 (52,39)	0 (56,28)	0.008 (26,13)	0.001 (50,32)	0 (68,36)	0 (54,33)
CTRANC centred	0 (71,35)	0.036 (51,38)	0 (50,20)	0.021 (24,13)	0 (56,24)	0 (56,22)	0 (51,25)
FDCTRANC	0.029 (64,49)	**0.618** (27,25)	**0.442** (35,31)	0.001 (64,38)	**0.934** (37,36)	**0.835** (31,32)	0.011 (43,35)
MCWF	**0.22** (87,82)	**0.057** (94,86)	**0.819** (55,58)	**0.081** (78,69)	**0.589** (75,71)	**0.078** (73,63)	0.028 (77,71)

Table 6.1: Interference - showing p-level for each trial and each algorithm between RIR conditions using Wilcoxon rank sum. Mean for anechoic and reverb are shown below. Those that are not different with statistical significance are highlighted in bold.

	Trial						
Algorithm	1	2	3	4	5	6	All
CTRANC basic	**0.798** (85,90)	**0.743** (90,89)	**0.37** (75,69)	**0.065** (82,90)	**0.109** (82,91)	**0.926** (88,88)	**0.153** (84,86)
CTRANC centred	**0.726** (89,87)	**0.889** (89,90)	**0.113** (75,64)	0.036 (83,92)	**0.204** (87,93)	**0.579** (86,89)	**0.35** (85,86)
FDCTRANC	**0.645** (84,90)	**0.245** (81,89)	**0.695** (79,72)	**0.14** (84,92)	**0.235** (86,92)	**0.886** (88,92)	0.048 (84,89)
MCWF	**0.8** (49,46)	**0.963** (82,84)	**0.782** (22,22)	**0.214** (62,53)	**0.259** (59,68)	**0.446** (85,89)	**0.952** (60,60)

Table 6.2: Artefacts - showing p-level for each trial and each algorithm between RIR conditions using Wilcoxon rank sum. Mean for anechoic and reverb are shown below.

artefact criteria along with the centred CTRANC and CTRANC and performed similarly to the time domain CTRANC methods in the interference criteria, compared to the MCWF method. We can see that FDCTRANC reduced the level of the microphone bleed, since in 5 out of 6 trials it was rated higher than the anchor. This does not give an indication as to how much reduction has taken place as the scale is a percentage quality.

We can conclude that perceptually, each method is highly source dependent. It is also apparent that different features of input sources affected each method differently. More analysis is required to isolate which of these features affects the FDCTRANC method, for example spectral bandwidth, percussiveness or temporal changes.

6.3.2 Objective evaluation

The same audio samples used in the listening test outlined in the previous section were analysed using the BSS-EVAL toolbox to gain an objective view of the performance of each method. The unprocessed audio was also tested for comparison.

Figures 6.7 and 6.8 show the results of the analysis for the anechoic and reverberant audio samples respectively showing the SDR, SIR and SAR of each method for each room condition. The SIR and SDR of the unprocessed signal are also shown for comparison although the SAR is assumed to be ∞.

Figure 6.7 shows the centred CTRANC from the previous chapter resulted in the greatest mean SIR at 55.31dB, compared to the MCWF at 38.56dB, in anechoic conditions. This was expected as the centred CTRANC is particularly suited to anechoic conditions. FDCTRANC still performed higher than the MCWF with a mean SIR of 43.25dB. In terms of SAR and SDR there was little difference between each algorithm, with the centred CTRANC still performing the highest.

Figure 6.8 shows FDCTRANC produced the greatest SIR at 40.58dB in the reverberant case followed by the MCWF and basic CTRANC which resulted in mean SIRs of 32.69dB and 31.87dB respectively with little difference between them. The centred CTRANC performed worse, which was expected due to the reverberation and was the same result as the previous chapter. In the reverberant case the MCWF performed highest given in terms of SAD and SDR with FDCTRANC performing second.

6.3.3 Computational efficiency

Another factor to consider when evaluating a method is the computational cost. We processed 10 seconds of the test audio 100 times with each method using

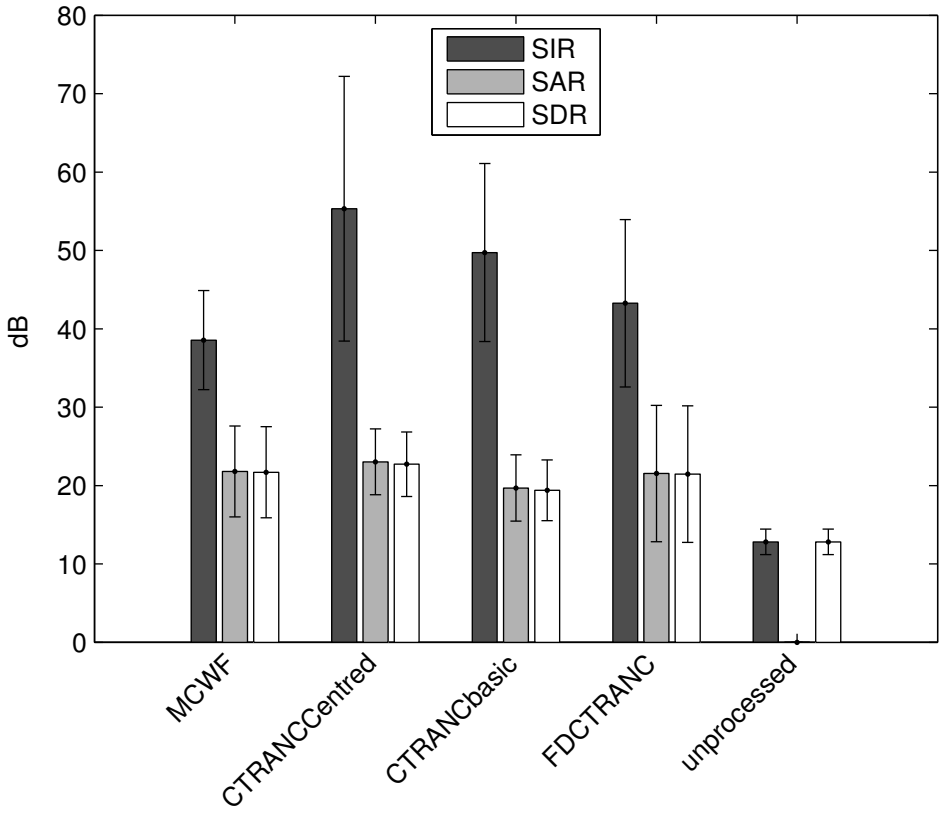


Figure 6.7: Objective measures of listening test audio data in anechoic conditions showing mean SDR, SAR and SIR for all trials for each algorithm under test. Standard deviation is shown.

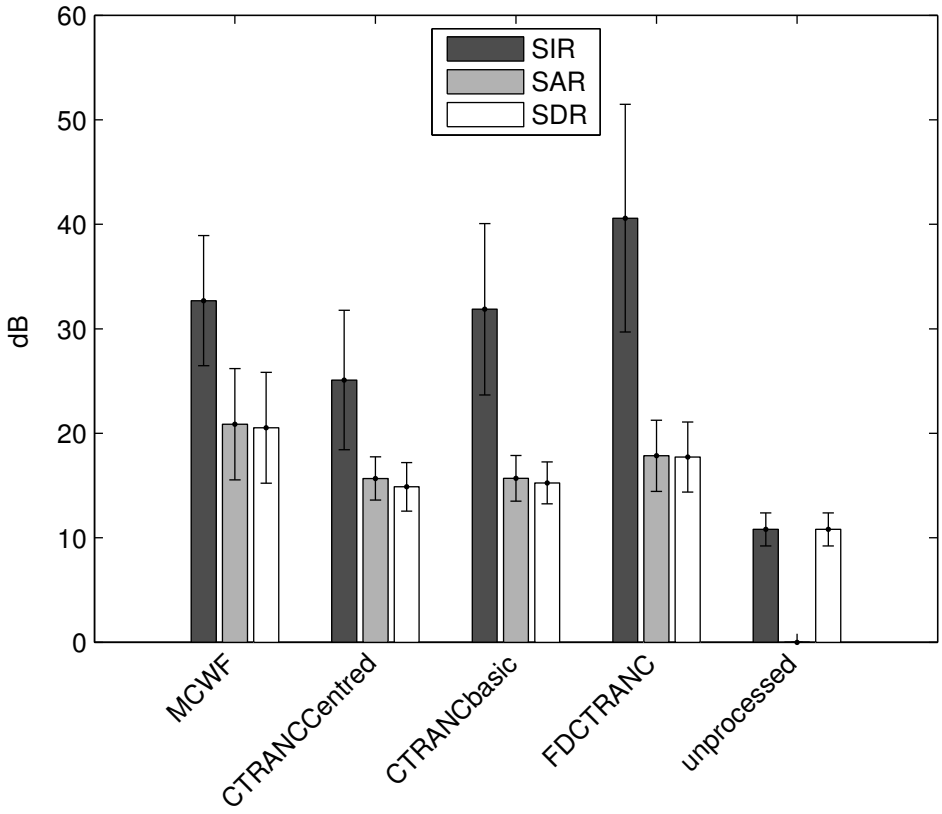


Figure 6.8: Objective measures of listening test audio data in reverberant conditions showing mean SDR, SAR and SIR for all trials for each algorithm under test. Standard deviation is shown.

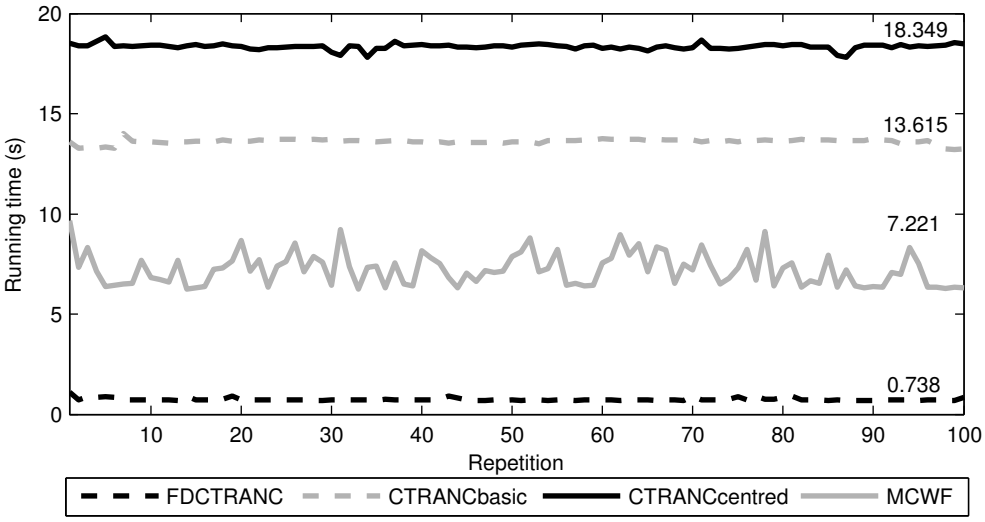


Figure 6.9: Running time of each algorithm in seconds for 100 repetitions of processing on 10 second audio samples at 44.1kHz sampling rate. The mean running time is indicated.

MATLAB on the 12 CPU core processing server at the Centre for Digital Music. Figure 6.9 shows the time taken for each method to complete each repetition. The mean running time is also shown.

Although these results are dependent to an extent on the implementation and potential savings in time could be made, the amount of optimisation that could be achieved is unlikely to offer a large decrease in time.

FDCTRANC completed the processing in the fastest time with a mean running time of 0.738 seconds. The centred CTRANC performed the slowest with a mean running time of 18.349 seconds. This was predominantly due to the delay estimation involved. This results may improve relative to the basic CTRANC with a larger frame size.

As FDCTRANC took less than 1 second to complete 10 seconds of processing and it is a frame-based method, it is likely real-time implementation can be achieved.

6.3.4 Discussion

Overall we can see that the results of the objective evaluation were different to the subjective listening test results. This may be because the subjective listening test results show how the algorithms are input dependent, which was shown by the difference in results for each trial, whereas in the objective case the results were similar for each different trial, as shown by the indicated standard deviation.

Regardless, the objective measures give an indication as to how well a method may work with real data but they suggest there is still some research to be done in developing a usable, perceptual objective measurement system. There is literature in this area [Emiya et al., 2011] which we have not used due to the computation time of the accompanying toolbox.

We can say that in the subjective and objective measurements all CTRANC based methods performed similarly and reduced the level of the microphone bleed while adding very little artefacts to the target source. Implementing CTRANC in the frequency domain provided a much lower computational cost with similar results.

These results also show that SDR and SAR are again very similar. Therefore we will continue to only use SDR in the next section.

6.4 Overdetermined FDCTRANC

The iterated FDCTRANC we have proposed is only relevant in the determined case. It is possible that the configuration of microphones and sources in a live sound production can be overdetermined. This may happen if single sources are being reproduced by multiple microphones in the same acoustic space, still assuming that each microphone is closest to one source, the target source. This is common if the sound engineer requires recordings of different aspects of the same instrument to be mixed together. If the microphones are not equidistant from the sound source comb filtering can occur. The comb filtering can be reduced by delay estimation, as described in Chapter 4. But if there is bleed on the microphone signals this will also result in comb filtering of the bleed.

Taking (6.15) and (6.15), if we assume that s_1 is the target source in both microphones and we apply a compensating delay to align s_1 in both microphones, this then becomes

$$\begin{aligned} x_1[n] &= \alpha_{11} s_1[n - \tau_{12}] + \alpha_{21} s_2[n - \tau_{21} + (\tau_{12} - \tau_{11})] \\ x_2[n] &= \alpha_{12} s_1[n - \tau_{12}] + \alpha_{22} s_2[n - \tau_{22}]. \end{aligned} \tag{6.39}$$

When x_1 and x_2 are summed, s_2 will still be comb filtered. Therefore bleed reduction has to be performed prior to comb filter reduction.

The problem with applying FDCTRANC in the determined case to this scenario is that it will attempt to remove bleed from multiple microphones that reproduce the same target source.

For example, extending the two source, two microphone case to three micro-

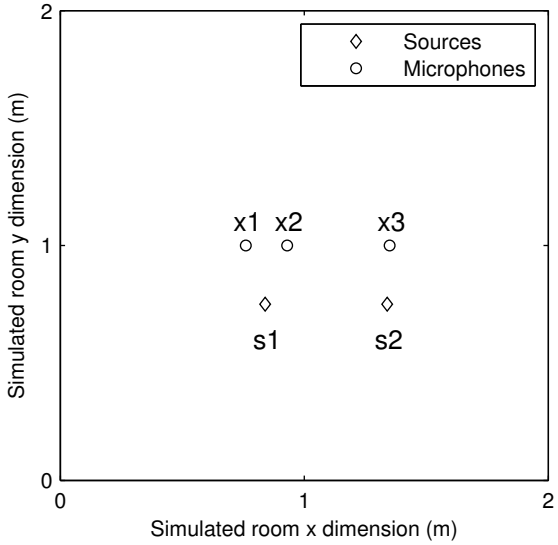


Figure 6.10: Example layout of sources and microphones as defined in (6.40),(6.41) and (6.42).

phones

$$x_1 = \alpha_{11} s_1[n - \tau_{11}] + \alpha_{21} s_2[n - \tau_{21}] \tag{6.40}$$

$$x_2 = \alpha_{12} s_1[n - \tau_{12}] + \alpha_{22} s_2[n - \tau_{22}] \tag{6.41}$$

$$x_3 = \alpha_{13} s_1[n - \tau_{13}] + \alpha_{23} s_2[n - \tau_{23}] \tag{6.42}$$

where s_1 is closest to x_1 and x_2 and s_2 is closest to x_3 and assuming $\tau_{11} > \tau_{12}$ and $\tau_{21} > \tau_{23}$ and therefore $\alpha_{11} < \alpha_{21}$ and $\alpha_{21} < \alpha_{23}$, as in the example shown in Figure 6.10.

If the iterated FDCTRANC algorithm is applied to this configuration, the first step will be

$$\hat{x}_1[n] = x_1[n] - (\mathbf{w}_{21}^T[n]\mathbf{x}_2[n] + \mathbf{w}_{31}^T[n]\mathbf{x}_3[n]) \tag{6.43}$$

as the algorithm assumes x_2 and x_3 are representations of interfering sources in x_1. In terms of s_1 and s_2 this then becomes

$$\begin{aligned} \hat{x}_1[n] =& \alpha_{11} s_1[n - \tau_{11}] + \alpha_{21} s_2[n - \tau_{21}] \\ & - (\alpha_{11} s_1[n - \tau_{11}] + \alpha'_{22} s_2[n - \tau'_{22}] \\ & + \alpha'_{13} s_1[n - \tau'_{13}] + \alpha_{21} s_2[n - \tau_{21}]) \qquad (6.44) \\ =& - \alpha'_{13} s_1[n - \tau'_{13}] - \alpha'_{22} s_2[n - \tau'_{22}] \qquad (6.45) \end{aligned}$$

where

$$\alpha'_{22} = \alpha_{22} - (\alpha_{12} - \alpha_{11}) \tag{6.46}$$

$$\tau'_{22} = \tau_{22} + (\tau_{11} - \tau_{12}) \tag{6.47}$$

$$\alpha'_{13} = \alpha_{13} - (\alpha_{23} - \alpha_{21}) \tag{6.48}$$

$$\tau'_{13} = \tau_{13} + (\tau_{21} - \tau_{23}) \tag{6.49}$$

where $\alpha'_{22} < \alpha_{22}$ and $\alpha'_{13} < \alpha_{13}$. As x_1 and x_2 have the same target source, the algorithm attempts to remove the same interfering source from each microphone. In this scenario, the result is that $\hat{x}_1$ contains both s_1 and s_2 reduced in amplitude and bleed reduction has not been achieved.

$\hat{x}_2$ is then calculated by

$$\hat{x}_2[n] = x_1[n] - (\mathbf{w}_{12}^T[n]\hat{\mathbf{x}}_1[n] + \mathbf{w}_{32}^T[n]\mathbf{x}_3[n]) \tag{6.50}$$

which will have a similar output to $\hat{x}_2$. $\hat{x}_3$ is then calculated by

$$\hat{x}_3[n] = x_1[n] - (\mathbf{w}_{13}^T[n]\hat{\mathbf{x}}_1[n] + \mathbf{w}_{23}^T[n]\hat{\mathbf{x}}_2[n]). \tag{6.51}$$

In this case the amplitude of the interfering source will be reduced and the target source retained.

6.5 Selective FDCTRANC

We have shown that the iterated FDCTRANC will fail when applied to the overdetermined case. In this section we propose a modification to FDCTRANC to include a selection process to avoid performing bleed reduction between microphones which have the same target source.

So in (6.43), the outcome will be that x_2 would not be considered a microphone reproducing an interfering source of x_1 and therefore would not have to be removed from x_2. So (6.43) would become

$$\hat{x}_1[n] = x_1[n] - \mathbf{w}_{31}^T[n]\mathbf{x}_3[n]. \tag{6.52}$$

The selection can be achieved by measuring the similarity between microphone signals. As we know that we are attempting to decide if two microphones are reproducing exactly the same source, traditional methods of similarly can be used, such as cross correlation. Therefore the GCC-PHAT outlined in Chapter 4 can also be used for this purpose by analysing the peak value of the output function. The problem with this method is that it relies on an accurate estimate of the delay for an accurate estimate of the degree of similarity.

Another approach is to measure the Pearson's correlation coefficient (PCC) between the frequency spectrum of each microphone. This is appropriate because in anechoic conditions with a single source and two microphones, in the frequency domain the difference between two microphones will be linear amplitude. Microphones positioned closest to the same target source will have a high correlation in the frequency domain as the same target source is the highest amplitude in each microphone. This has advantages because it is delay independent since delay only affects the phase. This also gives a single value to the amount of correlation between microphones. The correlation between frequency spectra is calculated by

$$\rho = \frac{\sum_{k=0}^{N-1}(|X_l[k]| - \overline{X}_l)(|X_m[k]| - \overline{X}_m)}{\sqrt{\sum_{k=0}^{N-1}(|X_l[k]| - \overline{X}_l)^2(|X_m[k]| - \overline{X}_m)^2}} \tag{6.53}$$

where X_l and X_m are x_l and x_m in the frequency domain and

$$\overline{X}_l = \frac{1}{N}\sum_{k=0}^{N-1}|X_l[k]| \tag{6.54}$$

$$\overline{X}_m = \frac{1}{N}\sum_{k=0}^{N-1}|X_m[k]| \tag{6.55}$$

are the mean magnitudes of X_l and X_m.

A high value of ρ indicates the microphones are reproducing the same target source but further analysis is required to establish a threshold at which to make this decision.

6.5.1 Correlation Threshold

To establish a suitable threshold of correlation, we analysed the correlation measure ρ with simulated microphones in anechoic conditions using two pink noise sources. The same image source toolbox as mentioned in Section 6.2 was used. The sources were placed 0.5m apart and two microphones, x_1 and x_2 were positioned 0.1m in front of each source. Another microphone, x_2 was moved in 0.025m increments between the two microphones from the position of x_1 across to the position of x_3. Figure 6.11 shows an example of this layout. The correlation using (6.53) was calculated for every frame of N samples of each microphone signal.

Figure 6.12 shows the mean ρ over all frames between each microphone at each position of x_2. The correlation as a function of distance between x_1 to x_2 and between x_2 to x_3 intersected at a point where x_2 was equidistant from x_1 and x_3. The mean correlation at this point was 0.83. We can consider this

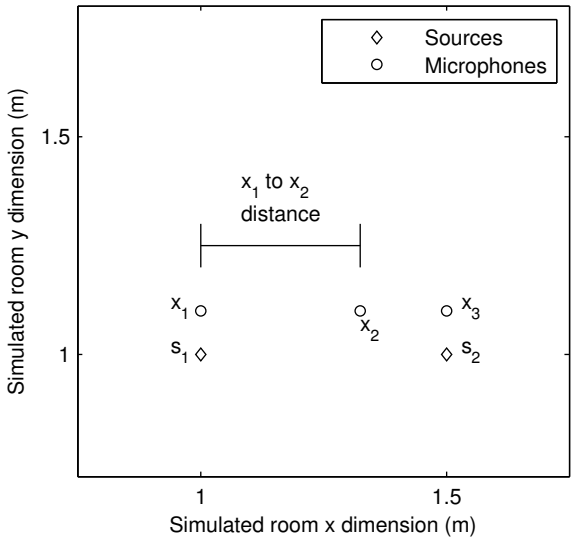


Figure 6.11: Layout of correlation test zoomed to show configuration.

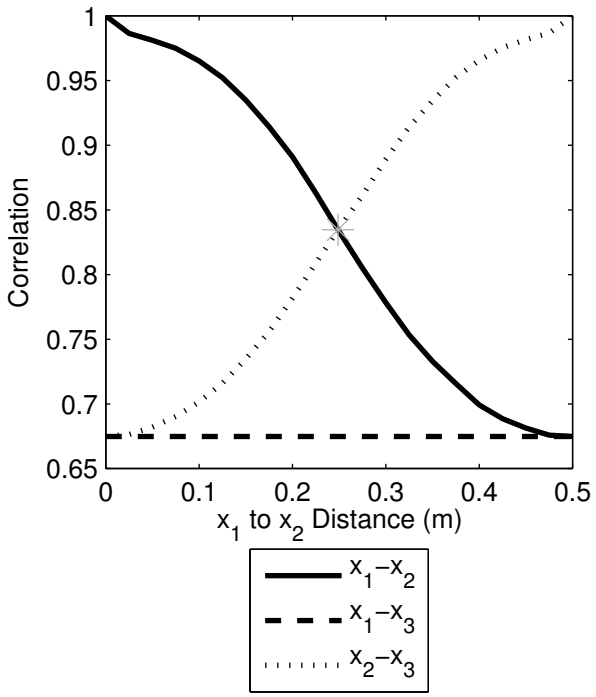


Figure 6.12: Mean correlation between microphones x_1 and x_2, x_2 and x_3 and x_1 an x_3 as the x_1 to x_2 distance is changed. The point of intersection is shown.

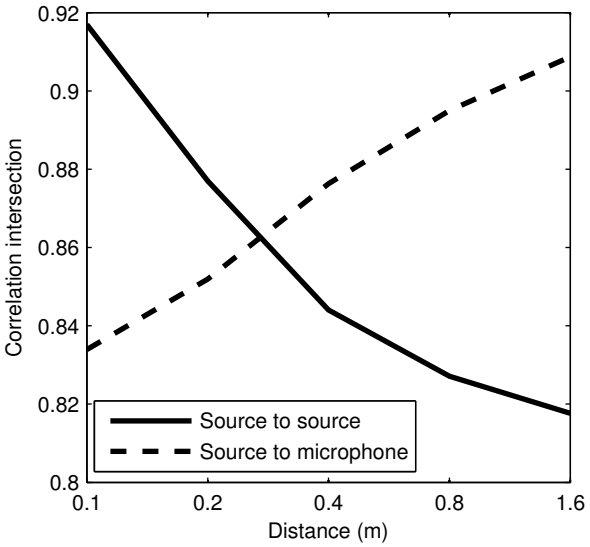


Figure 6.13: Plot showing ρ at the point of intersection when the source to source and source to microphone distance is altered.

the point where ρ between x_1 and x_2 and between x_2 and x_3 intersect, which we will refer to as ρ_I. The minimum correlation between x_1 and x_2, which are static, was 0.67. This suggests that a correlation coefficient above 0.83 would indicate that two microphones were highly correlated and therefore reproducing the same target source.

The same experiment was repeated under different conditions to establish how ρ_I changes with changing configurations and also to ascertain an acceptable threshold to indicate when two microphones are reproducing the same source.

In the first case the source to microphone distance was altered from 0.1m to 1.6m and the source to source distance retained at 0.5m. The source to source distance was then altered from 0.1m to 1.6m and the source to microphone distance retained at 0.1m.

Figure 6.13 shows ρ_I for different source to source and source to microphone distances. For the source to source distance the correlation ranged from 0.92 at a distance of 0.1m and 0.82 at a distance of 1.6m. This decrease in correlation was due to the increased distance between microphones at ρ_I. For the source to microphone distance the correlation ranged from 0.84 at a distance of 0.1m to 0.91 at a distance of 1.6m.

We then used the same configuration as in Figure 6.12 but altered the RT60 of the simulated environment.

Figure 6.14 shows the results for the change in RT60. In this case ρ_I ranged from 0.82 at 0s RT60 (anechoic) to 0.71 when the RT60 was 0.8s. This was due

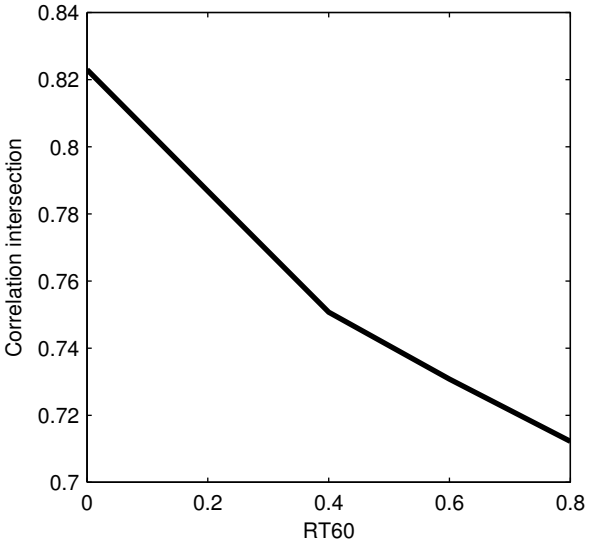


Figure 6.14: Plot showing ρ at the point of intersection when the RT60 of the virtual environment is altered.

to increasing amplitude of early reflections as RT60 increases, which will reduce correlation between a microphone close to a source and one further away due to timbral changes the reverberation will have on the source.

From this it was decided that a correlation coefficient of 0.9 was sufficient to indicate that two microphones are correlated and that they are reproducing the same target source. This value was chosen because at ρ_I, shown in the previous figures, x_2 is equidistant from x_1 and x_3. If this configuration occurs, the assumption that each microphone is closest to a single microphone does not hold and x_2 will no longer be a sufficient estimate of a single source. At this point the bleed reduction will fail, and therefore there is no need to run the selection process.

Figures 6.13 and 6.14 show that ρ_I changes with the configuration. Therefore the chosen value allows for a margin of error if a particular microphone is in a position equidistant to two sources.

Including this measure into the FDCTRANC framework, Figure 6.15 shows the proposed method as a block diagram. ρ is measured between each microphone prior to the subtraction of the filtered bleed signals in (6.43). If $\rho < 0.9$ then x_l is considered to be estimating an interfering source of x_m, else the two microphones are considered correlated and FDCTRANC should not be performed between them.

Adding the correlation measure will not increase computation significantly since the spectrum of each microphone signal is estimated by performing an

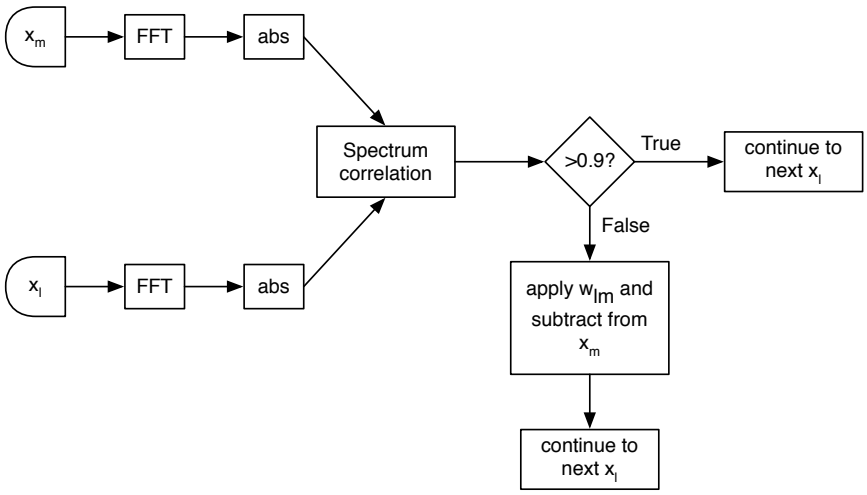


Figure 6.15: Block diagram of selective FDCTRANC.

FFT, which is already calculated for FDCTRANC. In some cases computation will decrease if some microphones are found to be highly correlated therefore bleed reduction will not be performed and less adaptive filters will be utilised. The correlation measure can also be utilised in the determined case to establish whether bleed reduction will be successful between two particular microphones

6.6 Evaluation

In this section we compare the proposed selective FDCTRANC against the basic iterative FDCTRANC in a variety of configurations.

We ran an experiment in simulated conditions. The room was 5m x 5m x 2.5m in size. The sources were placed at 0.5m intervals with a random error of ±0.05m to simulate real world situations. There were between two and six sources. For each number of sources, between one and three microphones were positioned in front of each source. The initial position for the first microphone was directly in front of the each source with a distance randomly selected between ±0.1m and 0.2m. Subsequent microphones were then placed ±0.1m either side of the initial microphone. The maximum layout with six sources and three microphones per source can be seen in Figure 6.16. The RT60 of each configuration was changed from 0s (anechoic) to 0.8s in 0.2s increments.

The sources were a selection of recordings of solo musical instruments; an acoustic guitar, a male vocal, a piano, a fiddle, an electric guitar and an organ. All audio samples were taken from multitrack recordings available under Creative Commons. The samples used in the test were 20 second excerpts taken from 60 second samples to allow for the adaptive filters to stabilise as the source

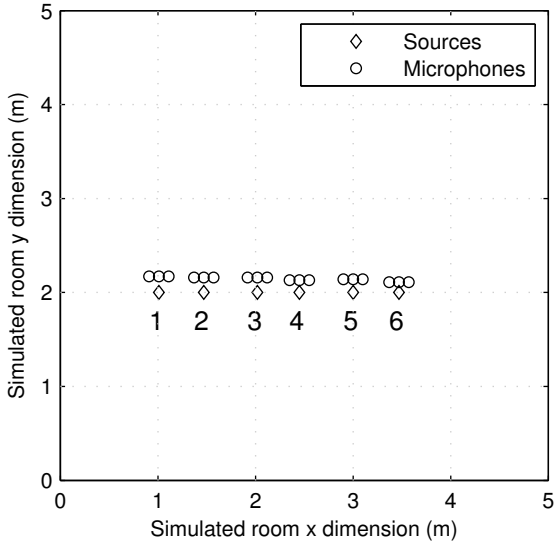


Figure 6.16: Virtual layout of sources and microphones in the maximum configuration for the results in Figures 6.17 and 6.18.

and microphone positions were static. The iterative FDCTRANC and selective FDCTRANC were compared.

The resulting audio was analysed using the BSS-EVAL toolbox. Figures 6.17 and 6.18 show the mean improvement in SDR and SIR from the SDR and SIR of the original microphone signals for each configuration and RT60.

Figure 6.17 shows that SDR improvement decreased as the number of microphones per source increased for all number of sources. This was expected, as explained in the previous section. In all cases there was a decrease in SDR when more than one microphone was used, due to attempted bleed reduction between microphones with the same target source. The SDR improvement for the Selective FDCTRANC also decreased as the number of microphones increased but in most cases there is improvement. There is also a trend of decreasing improvement in both methods as the number of sources increased, tailing off as the number of sources reaches six.

The results shown in Figure 6.18 were less consistent. The SIR improvement shows a similar trend for all sources. In the standard FDCTRANC case the SIR improvement decreased as the number of microphones increased. This was due to the bleed being removed for the same target microphones. For the selective FDCTRANC the SIR improvement remained consistent as the number of microphones increased, especially for four to six sources, which showed similar results. There was also a decrease in performance as RT60 increased. This was expected as the estimation of the target source at a close microphone

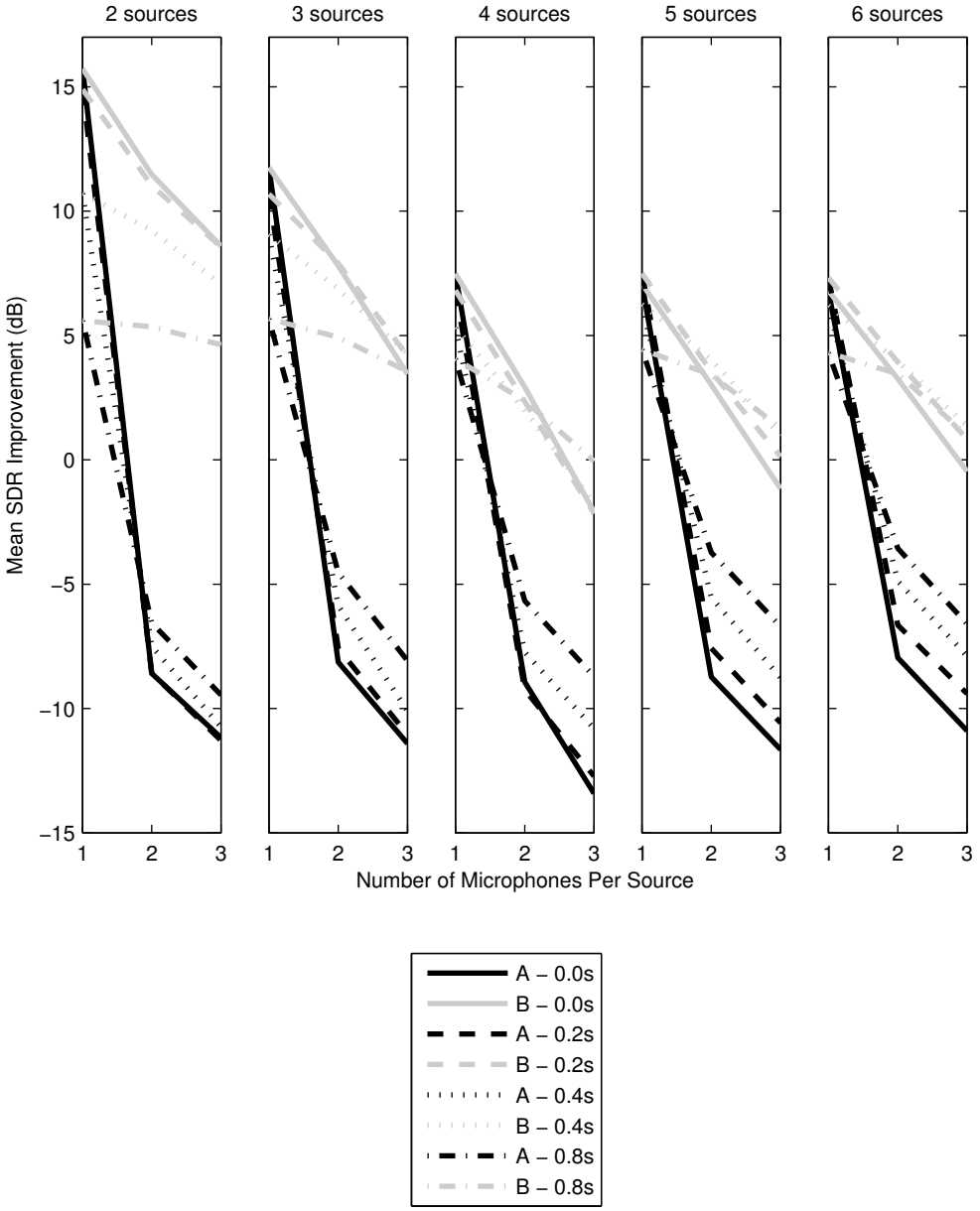


Figure 6.17: Mean SDR Improvement comparing FDCTRANC (A) and Selective FDCTRANC (B) with varying number of sources and number of microphones per source for different RT60 values.

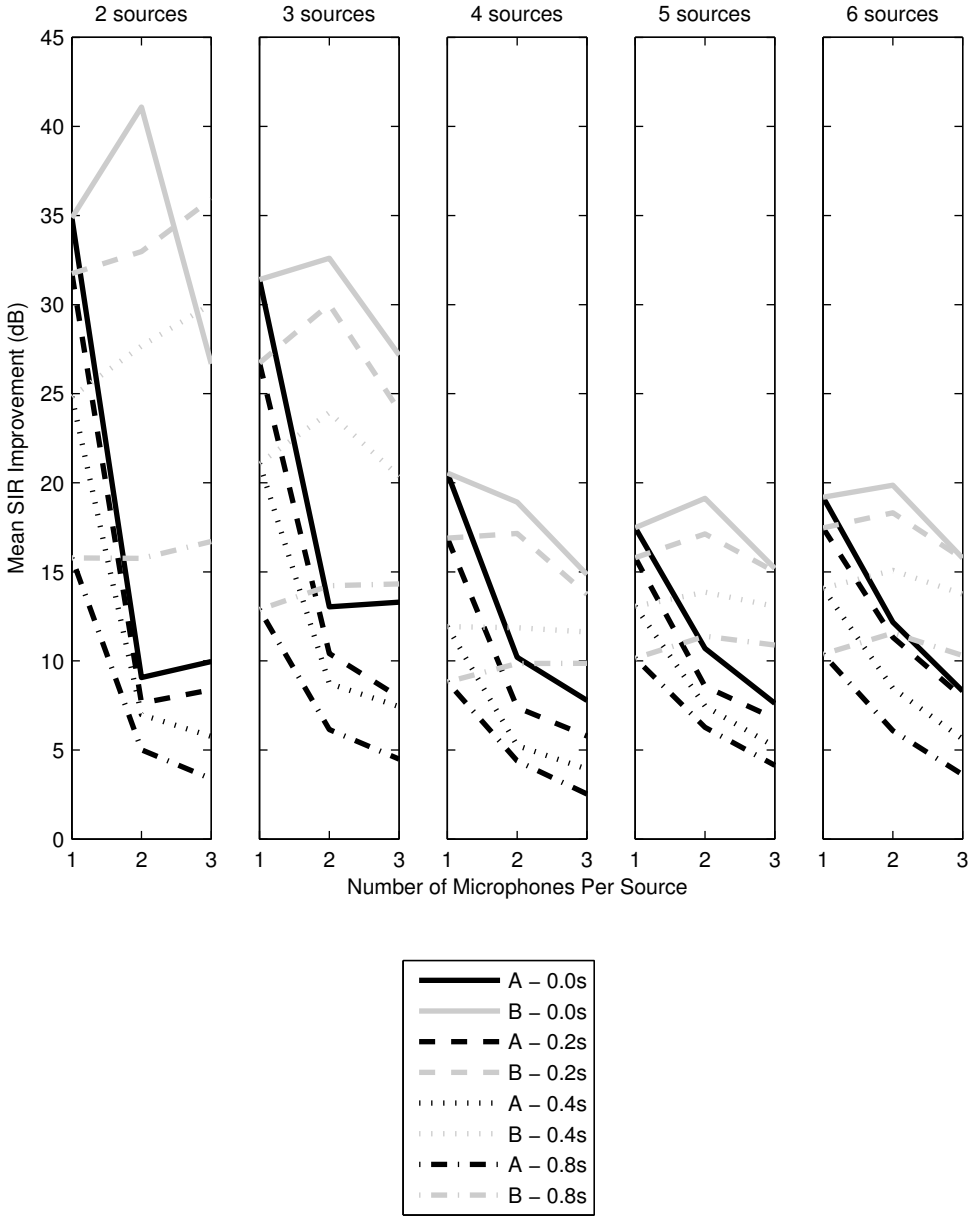


Figure 6.18: Mean SIR Improvement comparing FDCTRANC (A) and Selective FDCTRANC (B) with varying number of sources and number of microphones per source for different RT60 values.

will increasingly differ from the same source as an interfering source in a far microphone as RT60 increases.

By including the selection process in FDCTRANC, up to 20dB more SDR improvement is achieved compared to the standard FDCTRANC and as much as 32dB more SIR improvement in the two source case with two microphones per source in anechoic conditions.

As with the SDR improvement, the SIR improvement decreased as the number of sources increased due to the more complex interfering source mix, but after four sources the results remain similar.

6.7 Discussion and conclusions

In this chapter we extended CTRANC method in the previous chapter to the frequency domain to improve computation and performance. By doing this we uncovered problems with comb filtering that can occur with a straightforward implementation and proposed performing the method iteratively to reduce this effect.

The proposed iterative FDCTRANC method has been shown to be more computationally efficient, taking a mean time of 0.74s to process a 10 second audio sample whereas the time domain CTRANC was shown to take 13.6s.

We conducted a listening test to compare the proposed method to the Multichannel Wiener Filter method and on the methods outlined in the previous chapter. The proposed method was shown to perform similarly to the time domain CTRANC in terms of introducing artefacts. In terms of interference reduction the time domain CTRANC performed significantly better in 5 out of 6 trials in anechoic conditions but only one trial in reverberant conditions. This was echoed in objective metrics taken from the audio.

We then extended FDCTRANC to the overdetermined case. We showed that applying FDCTRANC to an overdetermined example will not result in bleed reduction of microphones reproducing the same source. We proposed a selection stage to counteract this by measuring the correlation in the frequency domain between microphones as microphones reproducing the same sound source will be highly correlated.

The selection process was shown to provide an improvement in the Signal-to-Interference Ratio to the original microphone signal by up to 40dB, which was as much as a 32dB increase compared to FDCTRANC. The proposed method was shown to outperform FDCTRANC in all overdetermined cases under test.

This chapter has shown a method for overdetermined microphone bleed reduction. The next chapter takes the knowledge we have gained in microphone bleed and uses it to investigate a different perspective, where microphone bleed is added to a signal to improve results rather than taken away.

Chapter 7

Microphone bleed simulation in multisampled drum workstations

The previous two chapters discuss research into the causes of microphone bleed and potential methods for removal. This chapter presents a preliminary investigation into a particular scenario where microphone bleed may be desired and how to artificially simulate this from audio data. In doing this we present a deeper understanding into microphone bleed and the positive aesthetic qualities it can provide in some circumstances.

Microphone bleed is inherent to all microphones recording multiple sources in the same acoustic space. A drum kit is an example of this as it can be thought of as a group of separate instruments always positioned in close proximity. This close proximity means microphone bleed is expected of a live drum kit recording. Bleed is considered the sound from a drum that is not the target arriving in the target microphone. In certain cases the absence of bleed reveals the artificial nature of the source material, such as in artificial drum recordings generated using multisampled drum workstations (MDWs). MDWs offer a user interface that triggers precisely recorded drum samples with the intention of producing realistic sounding drum loops. These drum samples are recorded in isolation and the software allows the user a large amount of control to load any drum piece into a certain location. Due to this, lack of ambience and microphone bleed can reduce the credibility of a realistic sounding drum kit. In such cases it is desirable to provide an approximation of the microphone bleed.

In this chapter we present a novel method of simulating tom-tom drum microphone bleed and resonance in MDWs while contributing to deeper understanding of microphone bleed and its applications. We first describe MDWs and explain why bleed is often required. We then present a method that only makes use of the audio samples generally available and evaluate the method using a listening test of expert participants. The results of the listening test showed that listeners are not able to discern the real recording with statistical

significance.

The research presented in this chapter was undertaken as part of a research project in collaboration with FXpansion Audio, an industry partner.

7.1 Multisampled drum workstations

Recording a full drum kit comes with many challenges, from simply finding a space big enough to adequately record a drum kit to dealing with issues that occur with the large amount of separate instruments in close proximity. MDWs allow amateur and professional engineers to recreate the sound of a full kit recorded in a professional studio simply from a laptop, for example FXpansion's BFD2 [FXpansion, 2013].

The premise of an MDW is to go one step further than a sampler or synthesiser. A drum kit is laid out in a studio with a standard microphone setup and each drum, or kit piece, is recorded in isolation and struck at many different velocities and positions. An interface is then developed to access these samples and allow the user to program their own drum beats and render all of the individual recordings together to create a studio quality emulation of a real drummer.

Ideally every microphone would be recorded for every drum hit to reproduce the bleed between the microphones. Then if the user sequences a drum loop and listens to one microphone in isolation, much like a real recording all of the drums would still be heard due to the bleed.

The problem with recording every microphone for every drum is that this ends up being a lot of data that needs to be recorded, distributed and stored. For this reason it is often the case that only the bleed into the kick or snare drum microphones is included with an MDW, as these are considered the most important kit pieces.

Another problem is that many MDWs allow users to construct their own complete drum kit, choosing from many different drum pieces. If the drum pieces were not recorded as part of the same drum kit, the microphone bleed will not be accurate.

It would be advantageous to be able to reproduce microphone bleed without having to provide the actual audio data. It may be possible to synthesise this missing data but this is at odds with the philosophy of creating an MDW from recorded samples. Techniques also exist for modelling drum kits [Laird, 2001; Bilbao, 2012] but are computationally complex and therefore simplified models of real drums.

This chapter outlines a method to simulate the bleed of a kick or snare drum into the tom-tom drum microphones using the bare minimum of data that

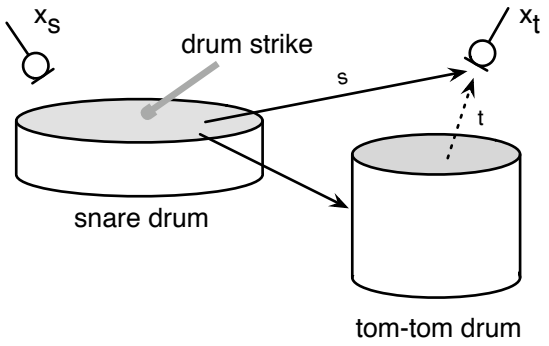


Figure 7.1: Drum microphone bleed and resonance.

would be available in an MDW. We evaluate how effective these simulations are compared to real data through listening tests.

7.2 Microphone bleed in drum kits

Generally while recording a drum kit the direct sound from each drum is recorded by a dedicated microphone. Therefore each microphone will have a single target drum. The bleed that occurs in a drum kit is more specialised than that described in Section 2.2.5 as the close proximity of drum pieces in a drum kit means the microphone bleed also contains the distinctive sympathetic resonances of the drum pieces.

The bleed that occurs on a microphone positioned to record a tom-tom drum is primarily from two sources; the direct sound of the interfering drum arriving at the microphone and the tom-tom resonating due to this direct sound. As we are particularly looking at the case where the snare or kick drum are the interfering drums, for the case of the snare drum as the interfering source this can be described as

$$x_t[n] = h_s[n] * s[n] + h_t[n] * \hat{t}[n] + w[n] \tag{7.1}$$

where x_t is the tom-tom microphone signal, s is the sound of the snare drum being struck, $\hat{t}$ is the tom-tom resonance excited by the snare drum, w is uncorrelated noise and h_s and h_t are room impulse responses between the snare drum and the microphone and the tom-tom resonance at the microphone when the snare drum is struck. This is demonstrated in Figure 7.1.

Drums can be generalised as a circular membrane stretched over an air space [Fletcher and Rossing, 1998]. When the membrane, or drum skin, is struck this causes the membrane to vibrate at different modes. This also causes the air within the drum to resonate as well as the drum body itself, producing a

characteristic sound. Drums can also resonate due to excitation from vibrations in the air due to other drums in the kit being struck, known as sympathetic resonance.

Tom-tom drums are tuned to resonate at different, complementary fundamental frequencies when struck. They are also notorious for resonating or "ringing" when other drums are played and may be tuned up or down to change the resonant frequency to avoid this. Although the ringing can be avoided it is an integral part of a real drum kit. In addition to this there are many different factors which will determine how the resonance of a tom-tom will sound in the microphone, including microphone type, the positions of the microphones, tom-toms, other drums, listening position, room characteristics and mechanical connections to other instruments.

These factors can be used to inform a method for simulating the drum bleed. For example if the exact position of drums and microphones was known then it would be possible to estimate the amplitude and delay changes and also use known equations for estimating the spectral change of a sound source over distance [Moorer, 1979]. Unfortunately it is unlikely that the details of all these factors are noted during a recording session. MDWs also allow users to place drums in almost any configuration and position, regardless of the original recording position. Assumptions therefore need to be made and the same algorithm needs to be able to simulate drums in a variety of configurations with a general approach.

For our purposes we assume the direct recording of the kick, snare and tom-tom microphones are available. In terms of an MDW, this is the bare minimum required for a convincing, configurable drum kit. Real recordings of kick and snare drum hits in tom-tom microphones were also available for analysis and comparison to our proposed method and simulations.

7.3 Microphone bleed simulation

In this section we outline the proposed method for simulating kick and snare drum bleed into tom-tom microphones. The bleed consists of the direct sound of the kick or snare drum in the tom-tom microphones, and also the sympathetic resonance that occurs on the tom-toms due to the direct sound.

7.3.1 Direct bleed

The direct kick or snare drum sound in the tom-tom microphone can be simulated from the direct recording of each drum. The direct recording has to be processed to simulate the changes that occur to the direct sound as it travels through air from the drum itself to the tom-tom microphone [Kuttruff, 2006].

It is unlikely the bleed will be heard in isolation, and therefore an approximate simulation will suffice.

The processing that the sound travelling through air undergoes can be generalised as a reduction in high frequency amplitude. Equations are well established for modelling air absorption dependent on distance [Moorer, 1979] but it is assumed the relative distances between drums are unknown. A high shelving filter taken from [Zölzer, 2002, pg. 51] was used to simulate air absorption on the direct recordings. The gain of the filter was then calculated from informal listening tests of previously recorded microphone bleed recordings, leading to a filter specification of -8dB gain at a 5kHz cutoff. In addition to this the source instrument was attenuated so that there would not be noticeable positive reinforcement when the bleed signals were mixed together.

7.3.2 Extracting tom-tom resonance

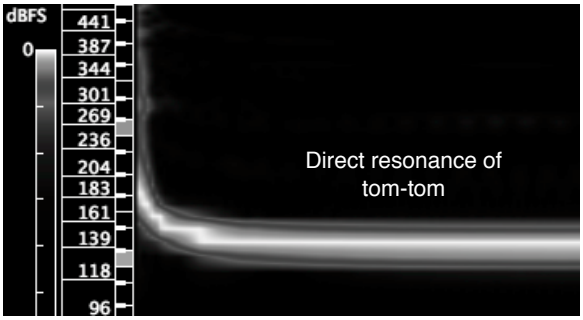
The next stage is to simulate the sympathetic resonance of the tom-tom drum to the external excitation of the kick or snare drum. The modes excited this way are also excited when the drum is struck directly. Therefore the modes can be extracted from the direct tom-tom recording.

The modes of an ideal circular membrane can be predicted [Fletcher and Rossing, 1998], although real tom-toms appear to diverge from the ideal case. It is known that the modes of a tom-tom will rise if struck with a large force as the strike displaces the membrane and changes the tension. Figure 7.2a shows a spectrogram of a tom-tom hit recorded at the tom-tom microphone, showing the fundamental mode of 138Hz. At the beginning of the hit the mode is at a higher frequency due to the force of the drum stick against the membrane. Figure 7.2b shows a spectrogram of a snare hit in the tom-tom microphone. The resonance of the fundamental mode of the tom-tom can clearly be seen at the same frequency but it is delayed due to the delay of the sound of the snare arriving at the tom-tom. The frequency of the mode is the same throughout the spectrogram.

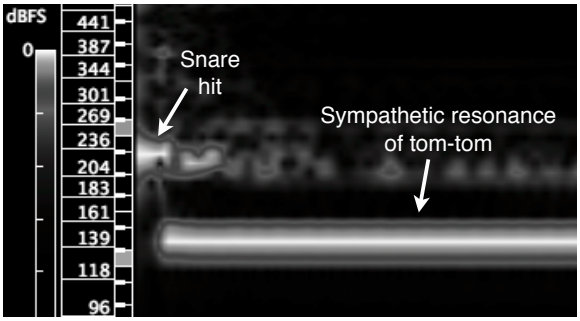
It is therefore not appropriate to use the unprocessed direct recording of the tom-tom to reproduce the tom-tom resonance due to the initial rise in frequency.

We can extract the stable resonance by measuring the spectral flux of the tom-tom signal [Lartillot and Toiviainen, 2007]. Spectral flux is a measure of the change of spectral content over time and can be used for transient and steady state detection [Zölzer, 2002, chap. 8], [Duxbury, 2001]. It is calculated by taking the Euclidean distance of the magnitude of subsequent frames of data,

[1] `http://www.sonicvisualiser.org/`



(a) Normalised spectrogram of direct tom-tom microphone while tom-tom is struck showing frequency over time.



(b) Normalised spectrogram of direct tom-tom microphone while snare is struck showing frequency over time.

Figure 7.2: Spectrograms taken from Sonic Visualiser [Cannam et al., 2010][1].

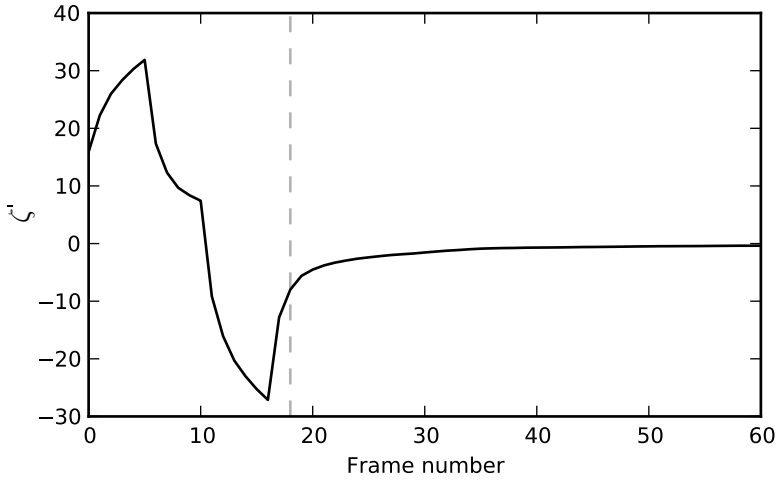


Figure 7.3: The first derivative of spectral flux ζ plotted against time. The beginning of the resonance is indicated by a dashed vertical line.

described by

$$\zeta[i] = \sqrt{\sum_{k=0}^{N-1} [|X[i,k]| - |X[i-1,k]|]^2} \tag{7.2}$$

where X is the microphone signal x in the frequency domain, k is the bin number from $0, \ldots, N-1$, N is the window size and i is the current frame. Once the fundamental mode of the tom-tom stabilises to a single value the spectral flux will also converge.

Figure 7.3 shows the first derivative of the spectral flux of a direct tom-tom signal, ζ'. The initial attack and decay can clearly be seen. The point at which the resonance begins can be extracted by finding the point where the first derivative of the spectral flux crosses a threshold after the minimum. From visual inspection of ζ' for a variety of tom-tom recordings and informal listening tests of the results a threshold of $\zeta' > -10$ was chosen. The position for this tom-tom is indicated by a dashed vertical line. The audio data after this point in time is used as the sympathetic tom-tom resonance.

7.3.3 Snare drum

This section outlines processing performed specific to when simulating snare drum bleed into the tom-tom microphone.

Resonance filter

For an object to sympathetically resonate, the resonant frequencies have to be excited. In relation to this research, this means that for a tom-tom to sympathetically resonate, the resonant frequencies must be produced by the snare drum [Rossing, 1992]. After listening to and analysing real tom-tom bleed recordings it became apparent that for low tom-toms, the fundamental frequencies are not excited by the snare drum hit but are excited when the tom-tom is hit directly. Therefore using the resonance of the tom-tom from a direct hit, as described in the previous section, will not be accurate for the simulation since it will contain frequencies which ordinarily would not be excited.

To mitigate this the extracted resonance is processed with a high pass filter with a cut off point taken from the peak frequency of the direct recording of a snare hit. It is assumed the snare drum will not produce significant amplitude frequencies below the peak frequency. In this implementation a 4th order Butterworth filter was used. The result of this is a more convincing low frequency tom-tom simulation where the fundamental frequencies are attenuated but the higher modes and any rattle of the tom-tom is retained.

Gain

Analysis of the real data shows that the peak amplitude of the direct snare hit has a linear relationship to the peak amplitude of the tom-tom bleed resonance. As mentioned previously, the position of the drums is unknown and therefore the gain cannot be directly estimated.

Through trial and error it was found that scaling the extracted resonance by a factor that is proportional to the difference in peak frequency of the snare drum and peak frequency of the extracted resonance produced audibly satisfactory results. This means that a large difference in peak frequency will result in a large gain factor and more attenuation as less modes are being excited, also reducing the low frequency mode level.

The steps of the method are outlined in Figure 7.4.

7.3.4 Kick drum

The kick drum produces much lower frequencies than the snare drum and will resonate lower frequencies of the tom-tom. Therefore filtering of the extracted resonance is not required. The extracted resonance is scaled by a single value for all tom-toms in comparison to the peak amplitude of the direct kick drum.

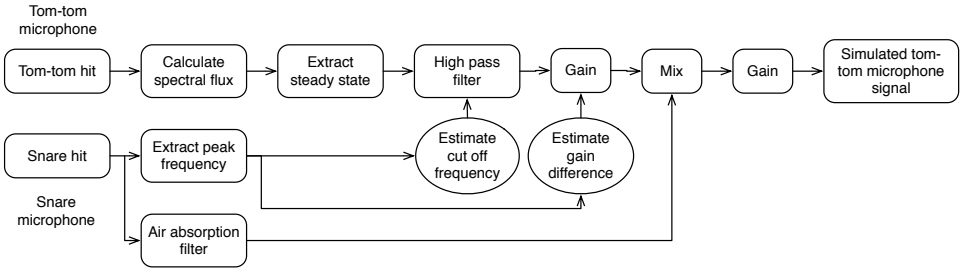


Figure 7.4: Block diagram of the method to simulate snare drum bleed in a tom-tom microphone.

7.4 Evaluation

The effectiveness of the simulations was established through a subjective listening test. We had available full recordings including bleed of four drum kits. The bleed was also simulated for these kits, using only the direct recordings of the snare, kick and tom-toms. The simulations were then compared to the real recordings. Both kick and snare bleed was simulated for every tom-tom in each kit. The kits each had six, three, three and four toms respectively. For this test a single hit velocity of the kick and snare drums was used, resulting in 32 audio samples available to analyse and simulate. The velocity of the hit used was in the mid range of the available velocities to test the algorithm on an average sample.

7.4.1 Subjective analysis

Description

A listening test was designed to ascertain whether a participant was able to distinguish the real recording from the simulation. The null hypothesis was that participants are unable to discern between real and simulated recordings.

A pairwise comparison listening test [Bech and Zacharov, 2006] was designed and implemented online. The test was conducted online to reach a wider audience and to attract more participants. The url was only distributed to those considered experts in the field of audio who had experience of critical listening, which resulted in 35 participants. The users were asked to indicate their experience in audio (audio engineer, software developer, student etc) and to rate their specific experience at listening to drum recordings on a scale of 1 to 10.

As a control test, the participant was firstly presented with two sounds; one direct snare signal and a snare signal mixed with the real tom-tom microphone with snare bleed and were asked to indicate which sound contained bleed. If the participant was unable to hear the bleed they were not included in the analysis.

The majority of participants were able to detect the bleed. The participant was then presented with a training page where they could listen to all the sounds which would be used in the listening test to familiarise themselves with the sounds.

The participants were presented with an interface with two buttons labelled 'Sound A' and 'Sound B' which when clicked would play the corresponding sound. In the majority of trials the real recording and simulation of the same recording would be randomly assigned as either A or B. 10 additional pairs were included where A and B were the same sound files, randomly chosen from the dataset, as a control to ensure the participant could establish when the sounds were the same or different. The order of pairs was randomised and therefore the test was double-blind.

After listening to both sounds, the user was given four options to choose from:

1. Sound A is a real recording.
2. Sound B is a real recording.
3. The sounds are different but either sound could be the real recording.
4. The sounds are the same.

Option 3 was included after pilot tests suggested it was common for a participant to identify the sounds were different but that both sounded like a real recording. Option 4 was included to establish if any simulations were good enough to be considered the same sound. The user was also given the opportunity to add any other comments about each pair.

Results

The results were analysed assuming a Binomial distribution as an adaptation of traditional ABX listening tests [Boley, 2009]. 25 of the participants correctly identified 7 out of the 10 identical pairs and were used for the following analysis.

Processing of the responses resulted in four possible outcomes for each pair trial:

- Correct identification of the real recording.
- Incorrect identification of the simulation as the real recording.
- Incorrect identification that the sounds are the same.
- Identifying the sounds are different but no preference which is the real recording.

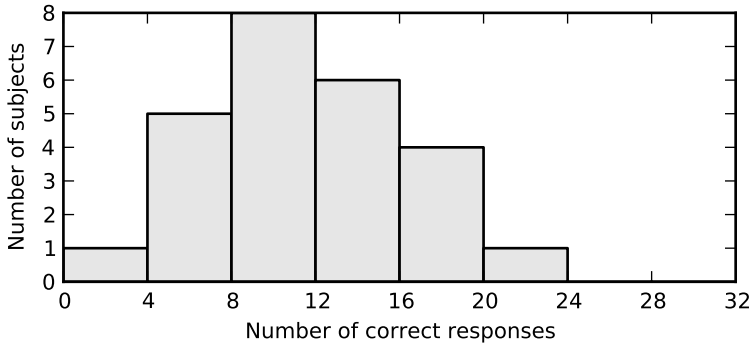


Figure 7.5: Histogram of the number of correct responses per subject.

To reject the hypothesis that participants are unable to distinguish between real and simulation recordings the users would have to correctly identify the real recordings with a high confidence.

For the 32 different pairs, the number of correct responses for each user is shown as a histogram in Figure 7.5. The mean number of correct response was 11.1, a probability of 0.35 of total responses, with a sample standard deviation 4.7.

Taking the probability of correctly identifying the real recording as 0.25 by chance, 9 subjects, or 37.5%, correctly identified the real recording with a confidence interval of $p <= 0.05$. As the users have been filtered by those that could identify the equal pairs, it can be assumed that the participant was highly unlikely to incorrectly identify the sounds are the same. If the probability of a user selecting the correct answer is now 0.33, 5 subjects, or 21%, correctly identified the real recording with a confidence interval of $p <= 0.05$.

The results therefore fail to reject the hypothesis that users are unable to identify the real recording from the simulation as only 5 participants out of 32 are able to correctly identify the real recordings with a statistical significance higher than 95%. This leads to the conclusion that the simulation is convincing in the majority of cases.

Figure 7.6 shows the number of correct responses against the signal-to-distortion (SDR) ratio between the real and simulated signal. The SDR was calculated using a modified version of performance measurements used in blind source separation [Vincent et al., 2006] and gives an indication of the perceptual difference between two signals. Table 7.1 shows the Pearson's Correlation Coefficient (PCC) and p-value for each pair. This shows there was a negative correlation between SDR and the number of correct responses and a positive correlation between the number of responses that the sounds are the same and

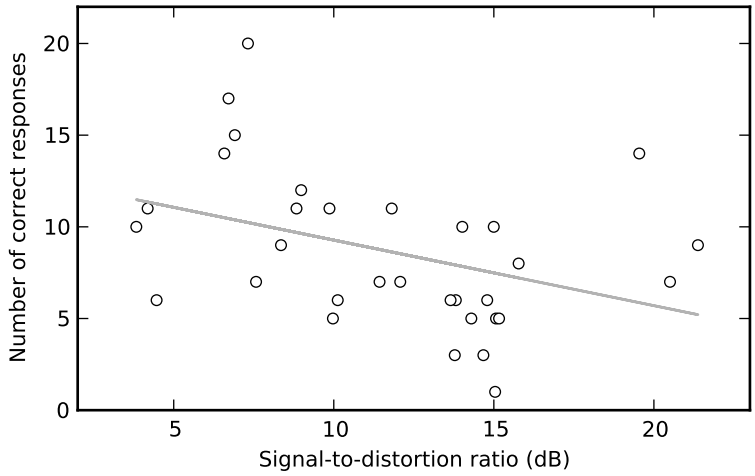


Figure 7.6: SDR plotted against the number of times that the real recording in each pair was correctly identified.

Response	*PCC*	*p-value*
Correct	-0.387	0.029
Incorrect	-0.046	0.804
Same	0.380	0.032
No preference	0.054	0.771

Table 7.1: Pearson's Correlation Coefficient of each response against SDR for each pair of sounds.

SDR. This was as expected, since it suggests that with pairs that are very different i.e. the simulation sounds different to the real recording, the real recording was more likely to be correctly identified. Equally, if the SDR is high and the pair sounds similar, they were likely to incorrectly respond that the sounds were the same. There is little correlation to the other responses. Although this suggests the participants were able to hear the difference, it is a fairly weak negative or positive correlation at around $\pm$ 0.4.

The results were also analysed using only participants that rated their experience as 6 out of 10 or higher. There was no significant difference between the results, which suggested the results were representative of audio experts with experience in drums and audio experts without.

7.5 Discussion and conclusions

In this chapter we have presented a method for simulating snare and kick drum bleed into tom-tom microphones from existing data. The bleed instrument part of the bleed signal is simulated by attenuating and filtering the direct bleed

instrument recording to simulate air absorption. The sympathetic resonance of the tom-tom by the bleed instrument is simulated by extracting the resonance from the direct tom-tom recording and applying a filter dependent on the peak frequency of the bleeding drum.

The simulation was subjectively tested using a pairwise comparison listening test and analysed using variations on analysis for ABX listening tests. Subjects were presented with pairs of sound, one of which was the real recording and one which was the simulation. The subjects were asked to indicate which sound was real or if the sounds were the same. The results were not statistically significant to reject the hypothesis that subjects were unable to distinguish the difference between the real and simulation. This suggests listeners were unable to identify the real recording in the majority of cases.

The simulation can be extended by simulating some of the finer details, such as rattle between tom-toms and the effects of groups of instruments on the resonance. A machine learning approach could be taken by processing recorded data to extract features that may be different between the direct recorded data and the bleed data.

The listening test can be extended by presenting subjects with the real and simulated recordings in a drum loop instead of single hits and simulating many different velocity layers.

This chapter has shown that it is possible to simulate microphone bleed in MDW drum loops purely from analysis of the audio and using audio sample which would be available to use. Although it is shown to be possible, this chapter does not investigate whether bleed in these cases is actually required, although it is assumed it will be optional as to whether the bleed is included and to what extent.

Chapter 8

Conclusions and future perspectives

In this chapter we summarise the outcomes of the thesis and suggest possible future directions for the research.

In this thesis we have set out to answer the question of whether microphone artefacts in live sound can be reduced by using digital signal processing techniques with no prior knowledge of microphone placement. This was achieved by either automatically emulating processes a human sound engineer would go through or by applying novel methods that could not be achieved by a human. This has been realised for the proximity effect, comb filtering and microphone bleed.

8.1 Proximity effect

In Chapter 3 we presented a novel method for detecting and correcting the proximity effect with no prior knowledge of the signal or source to microphone distance. This was achieved through analysis of the microphone signal for audio features that indicate the proximity effect and through dynamic filtering to reduce the effect.

Techniques to reduce the proximity effect rely on the skills and knowledge of the sound engineer using static equalisation or on specialist microphone construction. Section 3.1 outlined the literature on automatically reducing the proximity effect, which relies on knowledge of the source to microphone distance. The method we have shown in this thesis assumes the source to microphone distance is unknown and will change over time.

We have shown that the algorithm we researched was able to detect the proximity effect in test material recorded with a directional microphone with both a white noise and male vocal source. We were then able to correct the proximity effect using the same sources and a variety of types of movement.

8.1.1 Future perspectives

In this research we assumed that the proximity effect affected all frequencies below 500Hz equally. A new direction of research would be to investigate how the proximity effect changes with source to microphone distance and how to adapt the method we have proposed to reduce the proximity effect with adaptable filters.

The main assumption we make is that the sound engineer has already applied corrective equalisation to the source at a static distance, which is assumed to be the mean source to microphone distance when the source is moving. This is a possible future area of research to investigate other assumptions that can be made about how a source moves in front of a microphone and to find new ways of deciding on a baseline to aim correction towards.

We also assumed the proximity effect was only occurring due to the source to microphone distance decreasing. Another potential area of research is to investigate how the proposed method can be applied to the proximity effect due to changing angle of incidence.

The proposed method also assumes that the signal content does not change by a large amount in the low frequencies. The method could be extended with more research to take this into account.

8.2 Comb filter reduction

In Chapter 4 we have discussed using the GCC-PHAT method of delay estimation to inform compensating delays to reduce the effect of comb filtering in single source, multiple microphone configurations.

Using the GCC-PHAT on musical input signals had not been fully investigated in the prior literature. A survey of the literature in Section 4.1 also suggests there was little justification for the window shape used in the calculation.

We have provided an analysis of how the accuracy of the GCC-PHAT is correlated to the bandwidth of the incoming signal. We have shown that using a Blackman window increases the mean accuracy of the GCC-PHAT over a sample set of 20 different musical instrument recordings by 50% compared to the rectangular window. We have concluded that windows that taper to zero at the extremities, for example the Hann or Blackman window, are most appropriate for arbitrary musical sources.

8.2.1 Future perspectives

There are a number of areas concerned with the GCC-PHAT and comb filtering which can be further researched.

In the simulations used in Chapter 4 the sources are assumed to be point sources. In the real recordings analysed, loudspeakers were used to output different musical instrument sources as the goal was to investigate the effect of different signal bandwidths. Therefore this did not investigate the effect of the different sound radiation patterns of different instruments. For example, in close proximity to a large instrument such as a piano, the source radiates from across the full width of the instrument where the hammers hit the strings. Therefore there is no specific area of sound transmission.

In some instruments the area of sound transmission can also change depending on how it is played. The result of this is that there may in fact be different delays for different parts of the instrument played at different times. Early research by the author has tested the GCC-PHAT algorithm on a recording of a clarinet with successful results. This research can be extended to other instruments.

We also assume in this research that all of the microphones reproducing the same source are the same type and model with the same directivity pattern. As we described in Chapter 2, different microphones can have different characteristics and this is a future area of research. A microphone behaves as a filter on the source signal which will exhibit group delay which will cause different times of arrival for different frequencies. Being able to counteract this and still estimate the delay is a potential area of future research.

We also assume only linear changes to the source between the two microphones. Further research is required into the effect of non linear filtering on one of the microphone signals, for example through the use of audio effects. For example a use of delay estimation was suggested as being between a guitar recording directly through a DI box and through a microphone reproducing a guitar amplifier. Amplified guitars can have an effect applied, or even just the effect of the amplifier itself on the signal. The effect this has on the GCC-PHAT is an area of further research. Distortion in this case can be a particular problem, along with any effects which may change the phase of the signal.

We also assume that the delays we are concerned with estimating are of an integer number of samples. We have not discussed the use of sub sample delays, which is another future research topic.

8.3 Microphone bleed reduction

In Chapters 5 and 6 we presented research into reducing microphone bleed in multiple source, multiple microphone configurations.

We presented an extension of CTRANC, a technique for noise cancellation with crosstalk from telecommunications that had not previously been applied in a live sound context. We proposed combining CTRANC with delay estimation to improve the accuracy of the method. In anechoic conditions the inclusion of the centred adaptive filters proved to improve the Signal-to-interference ratio by as much as 18.2dB whilst also adding less artefacts than the original CTRANC method. In reverberant conditions the centred adaptive filters improved the Signal-to-artefact ratio by a maximum of 8.1dB but at the detriment of interference reduction. The centred CTRANC proved to be computationally complex and to only improve interference reduction in low reverberation configurations.

In Chapter 6 we implemented CTRANC in the frequency domain to become FDCTRANC. From this we found that there were issues with comb filtering in the method which had not been discussed in the literature. We proposed iterating over the method to reduce the comb filtering effects. Analysis of test audio samples in simulated reverberant conditions showed that the proposed method produced a maximum Signal-to-interference ratio of 40.6dB compared CTRANC at 31.9dB. We have also shown that FDCTRANC is significantly faster than CTRANC, taking less than 1 second to process 10 seconds of audio compared to a mean time of 13.6 seconds for CTRANC, while still producing similar perceptual results, shown through a listening test.

We then expanded FDCTRANC to the overdetermined case by introducing a selection stage to determine whether multiple microphones were reproducing the same source. The selection process was shown to improve the results of the FDCTRANC, resulting in as much as 32dB Signal-to-interference ratio improvement over the FDCTRANC with selection stage in simulated overdetermined configurations and outperforming the FDCTRANC with selection in all overdetermined configurations tested.

8.3.1 Future perspectives

There is potential for in depth future research into the FDCTRANC method. This research was concerned with applying the method to the live sound configuration, to which it had not previously been applied. An interesting future research area is to investigate the frequency dependent step size in more depth to discern how it affects the accuracy and convergence of the method and how it can be exploited for a variety of input signals.

In the selective FDCTRANC method, the selection is achieved through fre-

quency domain correlation. It is shown that this is suitable for the simulated configurations tested in this research. It would be interesting in the future to test this on real recordings with higher levels of noise and reverberation, as the selection may not perform as well as expected and other methods of selection could be employed, such as using self similarity matrices from MIR.

Like many audio processing methods, the methods we have presented for bleed reduction become less effective when more reverberation is introduced. The adaptive filter based methods we have presented are able to perform some reduction in reverberant conditions and on informal listening tests are able to reduce the level of the direct sound and some early reflections but often leave the late reverberation. One of the reasons we want to remove the bleed is that it can cause comb filtering. This then leads to a potential research project to investigate the effect reverberation has on the perception of the target signal and whether complete removal of all of the reverberation by other methods has detrimental effects. Although difficult to answer, another question posed is that of preference, that is whether some late reverberation left in the signal is adequate or complete removal with artefacts on the target source is preferred.

8.4 Microphone bleed simulation

In Chapter 7 we presented research into simulating microphone bleed in multi-sampled drum workstations. This research was conducted by the author while based at an industry partner, therefore the outcome is specific to their product.

Despite this the algorithm developed for this research holds and listening tests show that expert listeners were not able to discern the simulated bleed from the real recording with statistical significance. This has also not been achieved in other products.

8.4.1 Future perspectives

This research opens up more questions regarding the perception of microphone bleed. We included the microphone bleed to enhance the realism of a simulated drum kit recording. It might be the case that this can also be applied to other simulations and synthesised sounds. It also asks whether microphone bleed in a real recording is desired or not. There is no definitive answer to this. If the bleed is causing a problem such as comb filtering or problems with mixing, it would be desirable to have it removed. But it is also possible that the "problems" are in some cases what makes a recording realistic.

The bleed simulation work could also be extended by including more acoustic theory into the method, or using the simulated positions of the virtual drums to

make some inferences about the audio processing that needs to be done, rather than from analysis of the audio with no other information.

8.5 Overall future perspectives

The research in this thesis is concerned with reducing the proximity effect, comb filtering and microphone bleed in live sound. They are related in terms of being caused by microphone position errors and some of the approaches to reduce them share common ground, such as using delay estimation to reduce comb filtering and also to centre the adaptive filters in Chapter 5.

With regards to the research as a whole the first area to pursue is looking at other artefacts that were described in Chapter 2. Dereverberation is a current area of research with some interesting results. Like the bleed reduction/source separation field, there is a compromise between accurate dereverberation and retaining the target source. An interesting area may be to look at trying to reduce the level of distinct, high amplitude echoes that can cause comb filtering of the target signal.

Another area of research is to investigate how the proposed methods work together when applied to a complex configuration of microphones and sources. For example investigating how delay estimation between microphones is affected by microphone bleed from other sources and how this is improved by delay estimation. It is possible that if there are any changes to the phase of each signal through the bleed reduction, the delay estimation may not be as accurate.

With regards to extending the research presented in the thesis, the overall future direction is to include more testing of each method in more reverberant and noisy environments. A factor of live sound performance is that there will inevitably be an audience in the same space. For the purposes of this thesis we assumed the only sources were those expected in a musical performance and research is required to thoroughly test each method.

As mentioned in Chapter 2, microphone technology appears to be going in the digital direction and manufacturers are increasingly able to include digital signal processing (DSP) within the microphone itself which is tuned to that specific microphone. It is the author's opinion that as DSP becomes more efficient and chips become smaller and more affordable, the manufacturers of said equipment will exploit the capabilities more. This leaves an area open for research possibilities.

From this thesis we have learned that audio signal processing can be used for reducing microphone artefacts in live sound with no prior knowledge of the sources or microphones. The reduction in artefacts makes a considerable difference to the microphone signals and has implications for future audio signal processing in the live sound domain.

Appendix A

Analysis of vocal recording in proximity effect correction

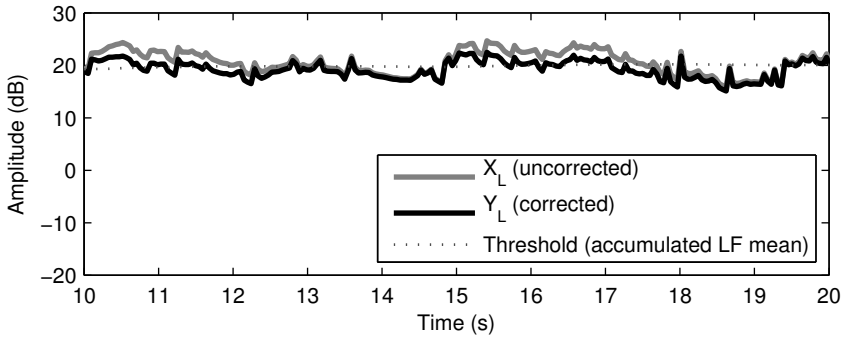


Figure A.1: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(1) with male vocal input.

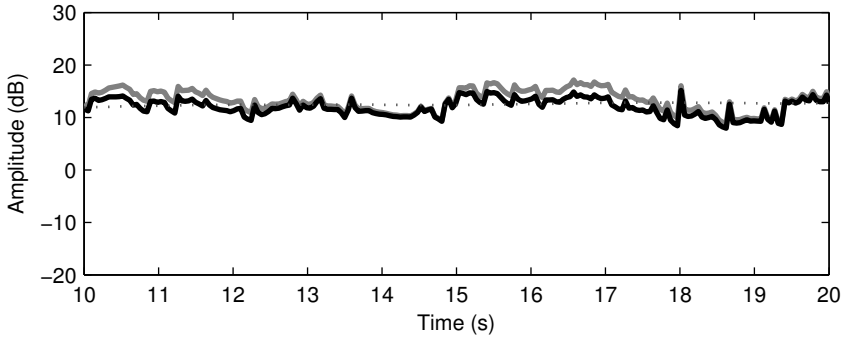


Figure A.2: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(2) with male vocal input.

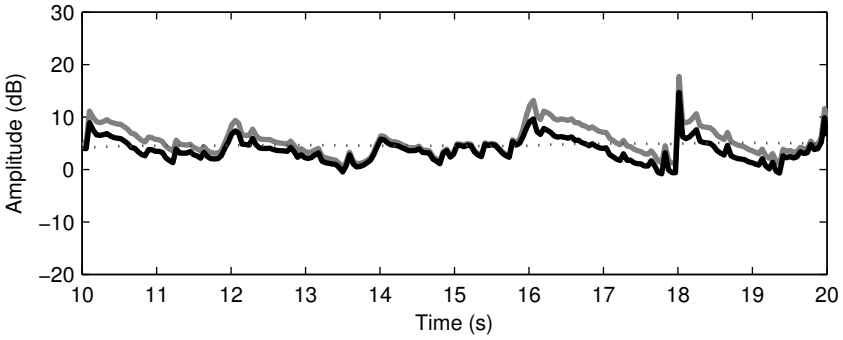


Figure A.3: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(4) with male vocal input.

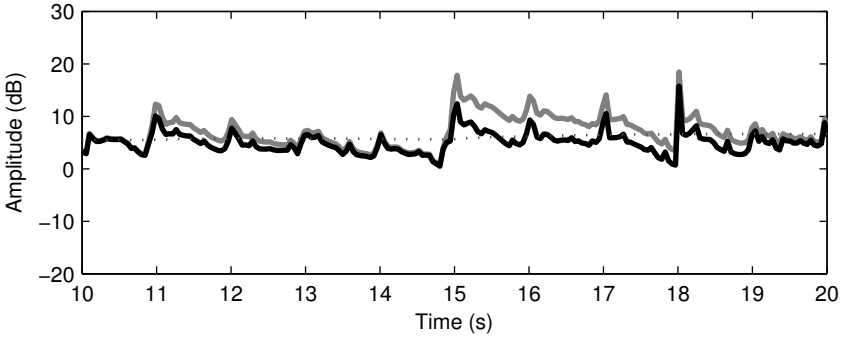


Figure A.4: Low frequency amplitude before and after proximity effect correction for the movement described in Figure 3.8(5) with male vocal input.

Appendix B

Comparing the GCC-PHAT to the Impulse Response with Phase Transform method

A common method of delay estimation between microphones reproducing the same sources is the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [Knapp and Carter, 1976]. Perez Gonzalez and Reiss [2008c] also suggests delays between microphones can be estimated using a method to estimate the impulse response by Meyer [1992] and applying the Phase Transform to that, referred to here as the IR-PHAT.

Here we show that the GCC-PHAT is equivalent to the IR-PHAT. In Section 4.2 we showed that the GCC is calculated by

$$\Psi_G[k] = X_1^*[k] \cdot X_2[k] \tag{B.1}$$

where X_1 and X_2 are the microphone signals x_1 and x_2 in the frequency domain and k is the frequency bin where $k = 0, \ldots, N-1$ where N is the length of the signal.

The Phase Transform is achieved by making $|\Psi_G[k]| = 1$ for all k.

From Meyer [1992] the impulse response is calculated by

$$\Psi_I[k] = \frac{X_2[k]}{X_1[k]} \tag{B.2}$$

and the same Phase Transform can be applied. As we have normalised the magnitude, we can show that $Arg(\Psi_I[k]) = Arg(\Psi_G[k])$.

From (B.1), the complex conjugate multiply means that

$$Arg(\Psi_G[k]) = Arg(X_2[k]) - Arg(X_1[k]). \tag{B.3}$$

From (B.2), through complex division

$$Arg(\Psi_I[k]) = Arg(X_2[k]) - Arg(X_1[k]) \tag{B.4}$$

therefore

$$Arg(\Psi_I[k]) = Arg(\Psi_G[k]). \tag{B.5}$$

Bibliography

Aichner, R., Buchner, H., Yan, F., and Kellermann, W. A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Processing*, 86(6):1260–1277, 2006. Applied Speech and Audio Processing.

Anazawa, T., Takahashi, Y., and Clegg, A. H. Digital time-coherent recording technique. In *Proceedings of the 83rd Audio Engineering Society Convention*, 1987.

Araki, S., Mukai, R., Makino, S., Nishikawa, T., and Saruwatari, H. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, 11 (2):109–116, Mar 2003.

Assous, S. and Linnett, L. High resolution time delay estimation using sliding discrete fourier transform. *Digital Signal Processing*, 22(5):820–827, 2012.

Assous, S., Hopper, C., Lovell, M., Gunn, D., Jackson, P., and Rees, J. Short pulse multi-frequency phase-based time delay estimation. *Journal of the Acoustical Society of America*, 127(1):309–315, 2009.

Azaria, M. and Hertz, D. Time delay estimation by generalized cross correlation methods. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32 (2):280–285, Apr 1984.

Baeck, M. and Zölzer, U. Real-time implementation of a source separation algorithm. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, 2003.

Balan, R., Rosca, J., Rickard, S., and O'Ruanaidh, J. The influence of windowing on time delay estimates. In *Proceedings of the International Conference on Information Sciences and Systems*, 2000.

Barry, D., Coyle, E., and Lawlor, B. Real-time sound source separation: Azimuth discrimination and resynthesis. In *Proceedings of the 117th Audio Engineering Society Convention*, Oct 2004.

Bech, S. and Zacharov, N. *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley, 2006. ISBN 0470869232.

Bechler, D. and Kroschel, K. Considering the second peak in the gcc function for multi-source tdoa estimation with a microphone array. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, 2003.

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, Sep 2005.

Benesty, J., Sondhi, M., and Huang, Y. *Springer Handbook of Speech Processing*. Springer, 2008a.

Benesty, J. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America*, 107(1): 384–391, 2000.

Benesty, J., Chen, J., Huang, Y., and Dmochowski, J. On microphone-array beamforming from a mimo acoustic signal processing perspective. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1053–1065, 2007.

Benesty, J., Chen, J., and Huang, Y. *Microphone Array Signal Processing*. Springer, Germany, 2008b.

Bilbao, S. Time domain simulation and sound synthesis for the snare drum. *The Journal of the Acoustical Society of America*, 131(1):914–925, 2012.

Björklund, S. and Ljung, L. An improved phase method for time-delay estimation. *Automatica*, 45(10):2467–2470, 2009.

Boley, Jon; Lester, M. Statistical analysis of abx results using signal detection theory. In *Proceedings of the 127th Audio Engineering Society Convention*, Oct 2009.

Brandstein, M. S. Time-delay estimation of reverberated speech exploiting harmonic structure. *The Journal of the Acoustical Society of America*, 105(5): 2914–2919, 1999.

Brandstein, M. S. and Silverman, H. F. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, 1997a.

Brandstein, M. S. and Silverman, H. F. A practical methodology for speech source localization with microphone arrays. *Computer, Speech and Language*, 11(2):91–126, 1997b.

Brunner, S., Maempel, H.-J., and Weinzierl, S. On the audibility of comb-filter distortions. In *Proceedings of the 122nd Audio Engineering Society Convention*, 2007.

Brutti, A., Omologo, M., and Svaizer, P. Comparison between different sound source localization techniques based on a real data collection. In *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2008.

Cannam, C., Landone, C., and Sandler, M. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference* 1467–1468, Oct 2010.

Cardoso, J.-F. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

Carey, M., Parris, E., and Lloyd-Thomas, H. A comparison of features for speech, music discrimination. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 149–152, Mar 1999.

Champagne, B., Bédard, S., and Stéphenne, A. Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, 4(2):148–152, 1996.

Chen, J., Benesty, J., and Huang, Y. Performance of gcc and amdf based time-delay estimation in practical reverberant environments. *EURASIP Journal on Applied Signal Processing*, 1:25–36, 2005.

Chen, J., Benesty, J., and Huang, Y. A. Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, 2006:1–19, 2006.

Chen, L., Liu, Y., Kong, F., and He, N. Acoustic source localization based on generalized cross-correlation time-delay estimation. *Procedia Engineering*, 15: 4912–4919, 2011.

Cho, N. and Kuo, C.-. J. Underdetermined audio source separation from anechoic mixtures with long time delay. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, 2009.

Choi, S. and Eom, D. Minimizing false peak errors in generalized cross-correlation time delay estimation using subsample time delay estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 96(1):304–311, 2013.

Comon, P. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. Robust localization in reverberant rooms. In Brandstein, M. and Ward, D., editors, *Microphone Arrays: Techniques and Applications*. Springer-Verlag, 2001.

Donohue, K. D., Hannemann, J., and Dietz, H. G. Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments. *Signal Processing*, 87(7):1677–1691, 2007.

Dooley, W. and Streicher, R. The bidirectional microphone: A forgotten patriarch. *Journal of the Audio Engineering Society*, 51(4):211–225, 2003.

Duxbury, C. Separation of transient information in musical audio using multiresolution analysis techniques. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-07)*, 2001.

Eargle, J. *The Microphone Book*. Focal Press, Oxford, UK, 2004.

Elko, G. W., Meyer, J., Backer, S., and Peissig, J. Electronic pop protection for microphones. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '07)* 46–49, 2007.

Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, Sep 2011.

Etter, W. Distance-based automatic gain control with continuous proximity-effect compensation. In *Proceedings of the 133rd Audio Engineering Society Convention*, Oct 2012.

Faller, C. and Erne, M. Modifying stereo recordings using acoustic information obtained with spot recordings. In *Proceedings of the 118th Audio Engineering Society Convention*, 2005.

Févotte, C., Gribonval, R., and Vincent, E. Bss eval toolbox user guide. Technical report, IRISA Technical Report 1706, Rennes, France, Apr 2005.

Fletcher, N. H. and Rossing, T. D. *The Physics of Musical Instruments*. Springer, 1998. ISBN 0387983740.

FXpansion,. BFD2 product page, 2013. Last visited on 15/01/2013.

Georganti, E., May, T., Van der Par, S., Harma, A., and Mourjopoulos, J. Speaker distance detection using a single microphone. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1949–1961, 2011.

Geravanchizadeh, M. and Rezaii, T. Y. Transform domain based multi-channel noise cancellation based on adaptive decorrelation and least mean mixed-norm algorithm. *Journal of Applied Siences*, 9(4):651–661, 2009.

Giannoulis, D., Massberg, M., and Reiss, J. D. Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 2013.

Gnann, V. and Spiertz, M. Comb-filter free audio mixing using stft magnitude spectra and phase estimation. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, 2008.

Gottlieb, D. and Shu, C.-W. On the gibbs phenomenon and its resolution. *SIAM Review*, 39(4):644–668, 1997.

Gross, J., Etter, D., Margo, V., and Carlson, N. A block selection adaptive delay filter algorithm for echo cancellation. In *Proceedings of the 35th Midwest Symposium on Circuits and Systems*, volume 2 895–898, Aug 1992.

Habets, E. and Benesty, J. A two-stage beamforming approach for noise reduction and dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):945–958, 2013.

Hadei, S. A. and Iotfizad, M. A family of adaptive filter algorithms in noise cancellation for speech enhancement. *International Journal of Comptuer and Electrical Engineering*, 2(2):1793–8163, 2010.

Harris, F. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

Hassab, J. and Boucher, R. Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3):549–555, Jun 1981.

Haykin, S. *Adaptive Filter Theory*. Prentice Hall, 4th edition, 2001. ISBN 0130901261.

Hedayioglu, F., Jafari, M., Mattos, S., Plumbley, M., and Coimbra, M. Separating sources from sequentially acquired mixtures of heart signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '11)* 653–656, 2011.

Hetherington, P., Paranjpe, S., and Pennock, S. Stereo acoustic echo cancellation for telepresence systems. In *Proceedings of the 129th Audio Engineering Society Convention*, Nov 2010.

Howard, D. and Angus, J. *Acoustics and Psychoacoustics*. Focal Press, Oxford, UK, 2000.

Huang, Y., Benesty, J., and Chen, J. Identification of acoustic mimo systems: Challenges and opportunities. *Signal Processing*, 86(6):1278–1295, 2006. Applied Speech and Audio Processing.

Huber, D. M. and Runstein, R. E. *Modern Recording Techniques*. Focal Press, 2005. ISBN 0240806255.

Hyvärinen, A., Hoyer, P. O., and Ink, M. *Independent Component Analysis*. Wiley, 2001.

International Telecommunication Union,. Method for the subjective assessment of intermediate quality level of coding systems. Technical Report ITU-R BS.1534-1, International Telecommunication Union, 2003.

Izhaki, R. *Mixing Audio: Concepts, Practices and Tools*. Focal Press, 2007. ISBN 0240520688.

Jillings, N., Clifford, A., and Reiss, J. D. Performance optimization of gcc-phat for delay and polarity correction under real world conditions. In *Proceedings of the 134th Audio Engineering Society Convention*, 2013.

Josephson, D. A brief tutorial on proximity effect. In *Proceedings of the 107th Audio Engineering Society Convention*, 1999.

Jourjine, A., Rickard, S., and Yilmaz, Ö. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP '00.*, volume 5, 2000.

Katz, B. *Mastering Audio, Second Edition: The Art and the Science*. Focal Press, 2007. ISBN 0240808371.

Kendall, G. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4):71–87, 1995.

Knapp, C. H. and Carter, G. C. Generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

Kokkinis, E., Reiss, J. D., and Mourjopoulos, J. A wiener filter approach to microphone leakage reduction in close-microphone applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):767–779, 2011.

Kokkinis, E. K. and Mourjopoulos, J. Unmixing acoustic sources in real reverberant environments for close-microphone applications. *Journal of the Audio Engineering Society*, 58(11):907–922, 2010.

Kuttruff, H. *Acoustics: An Introduction*. Spon Press, 2006. ISBN 0415386802.

Kwon, B., Park, Y., and sik Park, Y. Analysis of the gcc-phat technique for multiple sources. In *Proceedings of the International Conference on Control Automation and Systems (ICCAS '10)* 2070–2073, Oct 2010.

Laird, J. *The Physical Modelling of Drums Using Digital Waveguides*. PhD thesis, University of Bristol, 2001.

Lartillot, O. and Toiviainen, P. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, 2007.

Lehmann, E. A. and Johansson, A. M. Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America*, 124(1):269–277, 2008.

Leonard, T. Time delay compensation of distributed multiple microphones in recording: An experimental evaluation. In *Proceedings of the 95th Audio Engineering Society Convention*, 1993.

Lepauloux, L., Scalart, P., and Marro, C. Computationally efficient and robust frequency-domain GSC. In *12th IEEE International Workshop on Acoustic Echo and Noise Control*, Aug 2010.

Lepaulox, L., Scarlart, P., and Marro, C. An efficient low-complexitiy algorithm for crosstalk-resistant adaptive noise canceller. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009)*, 2009.

Lim, J. and Oppenheim, A. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

Lu, Y., Fowler, R., Tian, W., and Thompson, L. Enhancing echo cancellation via estimation of delay. *IEEE Transactions on Signal Processing*, 53(11): 4159–4168, 2005.

Maddams, J., Finn, S., and Reiss, J. D. An autonomous method for multitrack dynamic range compression. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-11)*, 2012.

Madhavan, G. and de Bruin, H. Crosstalk resistant adaptive noise cancellation. *Annals of Biomedical Engineering*, 18:57–67, 1990.

Makino, S., Lee, T., and Sawada, H., editors. *Blind Speech Separation*. Springer, 2007. ISBN 1402064780.

Mansbridge, S., Finn, S., and Reiss, J. D. An autonomous system for multitrack stereo pan positioning. In *Proceedings of the 133rd Audio Engineering Society Convention*, Oct 2012a.

Mansbridge, S., Finn, S., and Reiss, J. D. Implementation and evaluation of autonomous multi-track fader control. In *Proceedings of the 132nd Audio Engineering Society Convention*, Apr 2012b.

Margo, V., Etter, D., Carlson, N., and Gross, J. Multiple short-length adaptive filters for time-varying echo cancellations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '93)*, volume 1 161–164, Apr 1993.

McCarthy, B. *Sound Systems: Design and Optimization: Modern Techniques and Tools for Sound System Design and Alignment*. Focal Press, 2006. ISBN 0240520203.

Meyer, J. Precision transfer function measurements using program material as the excitation signal. In *Proceedings of the 11th International Conference of the Audio Engineering Society: Test and Measurement*, 1992.

Meyer Sound,. *SIM System II V.2.0 Operation Manual*. Meyer Sound, 1993.

Millot, L., Elliq, M., Lopes, M., Pelé, G., and Lambert, D. Revisiting proximity effect using broadband signals. In *Proceedings of the 122nd Audio Engineering Society Convention*, 2007.

Mirchandani, G., Zinser, R.L., J., and Evans, J. A new adaptive noise cancellation scheme in the presence of crosstalk [speech signals]. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 39(10): 681–694, 1992.

Mitianoudis, N. and Davis, M. E. Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 11(5):489–497, 2003.

Moir, T. and Harris, J. Decorrelation of multiple non-stationary sources using a multivariable crosstalk-resistant adaptive noise canceller. *International Journal of Adaptive Control and Signal Processing*, 27(5):349–367, 2012.

Moorer, J. A. About this reverberation business. *Computer Music Journal*, 3 (2):13–28, 1979.

Nagel, F., Sporer, T., and Sedlmeier, P. Toward a statistically well-grounded evaluation of listening tests—avoiding pitfalls, misuse, and misconceptions. In *Proceedings of the 128th Audio Engineering Society Convention*, May 2010.

Nesbit, A., Plumbley, M. D., and Davies, M. E. Audio source separation with a signal-adaptive local cosine transform. *Signal Processing*, 87(8):1848–1858, 2007. Independent Component Analysis and Blind Source Separation.

Nikolov, E. and Milanova, E. B. Proximity effect frequency characteristics of directional microphones. In *Proceedings of the 108th Audio Engineering Society Convention*, 2000.

Nikolov, E. and Milanova, E. B. Proximity effect of microphones. In *Proceedings of the 110th Audio Engineering Society Convention*, 2001.

Novotny, M. and Sedlacek, M. The influence of window sidelobes on dft-based multifrequency signal measurement. *Computer Standards and Interfaces*, 32 (3):110–118, 2010.

Nuttall, A. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):84–91, 1981.

Olson, H. *Acoustical Engineering*. Professional Audio Journals, 1991.

Paliwal, K. K. and Alsteris, L. D. On the usefulness of stft phase spectrum in human listening tests. *Speech Communication*, 45(2):153–170, 2005.

Pan, C. Gibbs phenomenon removal and digital filtering directly through the fast fourier transform. *IEEE Transactions on Signal Processing*, 49(2):444–448, Feb 2001.

Parra, L. and Spence, C. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, May 2000.

Parsa, V., Parker, P., and Scott, R. Performance analysis of a crosstalk resistant adaptive noise canceller. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 43(7):473–482, 1996.

Pedersen, M. S., Larsen, J., Kjems, U., and Parra, L. C. A survey of convolutive blind source separation methods. In *Springer Handbook on Speech Processing and Speech Communication*. Wiley, 2007.

Perez Gonzalez, E. and Reiss, J. D. Automatic mixing: Live downmixing stereo panner. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, 2007.

Perez Gonzalez, E. and Reiss, J. D. Improved control for selective minimization of masking using inter channel dependancy effects. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, 2008a.

Perez Gonzalez, E. and Reiss, J. D. An automatic maximum gain normalization technique with applications to audio mixing. In *Proceedings of the 124th Audio Engineering Society Convention*, 2008b.

Perez Gonzalez, E. and Reiss, J. D. Determination and correction of individual channel time offsets for signals involved in an audio mixture. In *Proceedings of the 125th Audio Engineering Society Convention*, 2008c.

Perez Gonzalez, E. and Reiss, J. D. Automatic gain and fader control for live mixing. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)* 1–4, 2009.

Perez Gonzalez, E. and Reiss, J. D. A real-time semiautonomous audio panning system for music mixing. *EURASIP Journal on Advances in Signal Processing*, 2010(5), 2010.

Perez-Lorenzo, M., Viciana-Abad, R., Reche-Lopez, P., Rivas, F., and Escolano, J. Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. *Applied Acoustics*, 73(8):698–712, 2012.

Peters, R., Smith, B., and Hollins, M. *Acoustics and Noise Control (3rd Edition)*. Pearson Education Canada, 2011. ISBN 0273724681.

Ramadan, Z. M. A three-microphone adaptive noise canceller for minimizing reverberation and signal distortion. *American Journal of Applied Science*, 5 (4):320–327, 2008.

Reed, F., Feintuch, P., and Bershard, N. Time delay estimation using the lms adaptive filter - static behaviour. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3):561, 1981.

Rickard, S., Balan, R., and Rosca, J. Real-time time-frequency based blind source separation. In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA '01)* 651–656, 2001.

Rossing, T. D. Acoustics of drums. *Physics Today*, 45(3):40–47, 1992.

Rubo, Z., Guanqun, L., and Xueyao, L. A time-delay estimation method against correlated noise. *Procedia Engineering*, 23:445–450, 2011.

Rui, Y. and Florenico, D. Time delay estimation in the presence of correlated noise and reverberation (2004). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

Rumsey, F. and Mccormick, T. *Sound and Recording: An Introduction (Music Technology)*. Focal Press, 2005. ISBN 0240519965.

Salvati, D. and Canazza, S. Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments. *IEEE Signal Processing Letters*, 20(5):507–510, 2013.

Salvati, D., Canazza, S., and Roda, A. A sound localization based interface for real-time control of audio processing. In *Proceedings of the 14 th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

Savage, S. *The Art of Digital Audio Recording: A Practical Guide for Home and Studio*. Oxford University Press, USA, 2011. ISBN 0195394100.

Shapton, D. Digital microphones: A new approach? *Sound on Sound*, Mar 2004.

Shure,. *KSM42 User Guide*. Shure Incorporated, 2010.

Shure,. Shure SM58 dynamic microphone specification, 2013.

Shynk, J. Frequency-domain and multirate adaptive filtering. *IEEE Signal Processing Magazine*, 9(1):14–37, Jan 1992.

Soo, J.-S. and Pang, K. Multidelay block frequency domain adaptive filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(2):373–376, Feb 1990.

Sporer, T., Liebetrau, J., and Schneider, S. Statistics of mushra revisited. In *Proceedings of the 127th Audio Engineering Society Convention*, Oct 2009.

Tamin, N. S. M. and Ghani, F. Techniques for optimization in time delay estimation from cross correlation function. *International Journal of Engineering and Technology*, 10(2):69–75, 2003.

Torio, G. Understanding the transfer functions of directional condenser microphones in response to different sound sources. In *Proceedings of the 13th Audi Engineering Society Conference on Microphones and Loudspeakers*, 1998.

Torio, G. and Segota, J. Unique directional properties of dual-diaphragm microphones. In *Proceedings of the 109th Audio Engineering Society Convention*, Sep 2000.

Tourney, C. and Faller, C. Improved time delay analysis/synthesis for parametric stereo audio coding. In *Proceedings of the 120th Audio Engineering Society Convention*, 2006.

Uhle, C. and Reiss, J. D. Determined source separation for microphone recordings using iir filters. In *Proceedings of the 129th Audio Engineering Society Convention*, Nov 2010.

Van Gerven, S. and Van Compernolle, D. Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness. *IEEE Transactions on Signal Processing*, 43(7):1602–1612, Jul 1995.

Vincent, E. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):91–98, Jan 2006.

Vincent, E., Gribonval, R., and Fevotte, C. Performance measurement in blind audio source separation. *IEEE transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, Jul 2006.

Vincent, E., Gribonval, R., and Plumbley, M. D. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8):1933–1950, 2007. Independent Component Analysis and Blind Source Separation.

Vincent, E., Jafari, Maria, G., Abdallah, Samer, A., Plumbley, Mark, D., and Davies, Mike, E. Probabilistic modeling paradigms for audio source separation. In Wang, W., editor, *Machine Audition: Principles, Algorithms and Systems* 162–185. IGI Global, 2010.

Wan, X. and Wu, Z. Sound source localization based on discrimination of cross-correlation functions. *Applied Acoustics*, 74(1):28–37, 2013.

Ward, D., Reiss, J. D., and Athwal, C. Multitrack mixing using a model of loudness and partial loudness. In *Proceedings of the 133rd Audio Engineering Society Convention*, Oct 2012.

Westner, Alex; Bove Jr., V. M. Applying blind source separation and deconvolution to real-world acoustic environments. In *Proceedings of the 106th Audio Engineering Society Convention*, May 1999.

Widrow, B., Glover, J. J., McCool, J., Kaunitz, J., Williams, C., Hearn, R., Zeidler, J., Eugene Dong, J., and Goodlin, R. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716, 1975.

Yilmaz, Ö. and Rickard, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, Jul 2004.

Zeng, Q. and Abdulla, W. H. Speech enhancement by multichannel crosstalk resistant anc and improved spectrum subtraction. *EURASIP Journal on Applied Signal Processing* 1–10, 2006.

Zinser, R., J., Mirchandani, G., and Evans, J. Some experimental and theoretical results using a new adaptive filter structure for noise cancellation in the presence of crosstalk. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '85)*, volume 10 1253–1256, Apr 1985.

Zölzer, U., editor. *DAFX - Digital Audio Effects*. Wiley, UK, 2002.