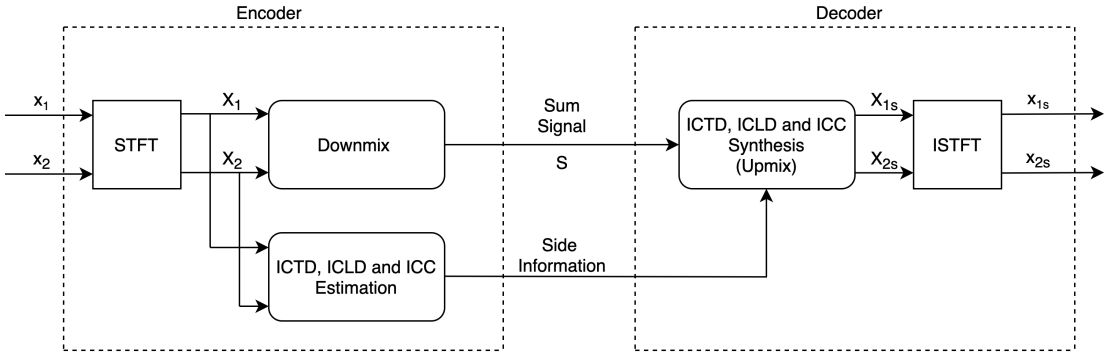
CHALMERS
UNIVERSITY OF TECHNOLOGY



Parametric Coding of Binaural Audio

An implementation and brief study of the low-complexity binaural cue coding scheme and its application on binaurally rendered and recorded signals

Master's thesis in Sound & Vibration

Viktor Bjärnklint

Department of Architecture and Civil Engineering
Division of Applied Acoustics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

Master's thesis 2020:ACEX30

Parametric Coding of Binaural Audio

An implementation and brief study of the low-complexity binaural cue coding scheme and its application on binaurally rendered and recorded signals

Viktor Bjärnklint

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2020

Parametric Coding of Binaural Audio
An implementation and brief study of the low-complexity binaural cue coding scheme and its application on binaurally rendered and recorded signals
Viktor Bjärnklint

Supervisor: Viktor Gunnarsson, Dirac
Examiner: Jens Ahrens, Division of Applied Acoustics

Master's Thesis 2020:ACEX30
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Block diagram of the low-complexity binaural cue coding scheme

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2020

Parametric Coding of Binaural Audio
An implementation and brief study of the low-complexity binaural cue coding scheme and its application on binaurally rendered and recorded signals.

Viktor Bjärnklint
Division of Applied Acoustics
Department of Architecture Civil Engineering
Chalmers University of Technology

Abstract

Binaural cue coding, BCC, presumes that the signals in the two channels differ primarily with respect to time, level, and coherence. The method then assumes that both channels can be synthesised by imposing the according time, level, and coherence differences onto a single-channel signal.

BCC can generally be split into two major parts: encoding and decoding. Encoding represents the analysis of the signal and investigates the signal's inter-channel time and level differences as well as the inter-channel coherence. The signal's channels are then summed, the signal is downmixed. The decoding process upmixes the single-channel signal to its original number of channels and restores the inter-channel time and level differences and uses the inter-channel coherence to mimic the inter-channel coherence of the original audio signal.

Binaural signals, signals consisting of two channels each filtered with head related transfer functions, HRTFs, are sensitive to compression algorithms since they depend on a highly delicate time and level differences between the two channels. Thus, the thesis is focused on the low-complexity implementation of the BCC scheme presented by Faller in his Phd thesis *Parametric Coding of Spatial Audio* and how well this scheme can synthesise the spatial cues present in binaural signals. The result and analysis is divided into two parts: a simple analysis used to investigate the general performance of the implemented algorithm and an informal listening test evaluating how well the program handles the synthesis of the binaural signals.

The synthesised audio exhibits a varied intensity of artefacts if the ICTD synthesis is incorporated in the BCC process. If the ICTD synthesis is left out of the BCC process, almost no artefacts can be perceived. However, this reduces the lateralization of the sounds/sound sources present in the audio. The conclusion that follow is thus that the implementation illustrates the potential the method has for transferring high complexity audio, such as binaural signals, at low bitrates.

Keywords: BCC, binaural, ICC, ICTD, ICLD, audio, sound.

Acknowledgements

I would like to thank my examiner Jens Ahrens and supervisor Viktor Gunnarsson for all the valuable feedback they have provided during the course of this project. I also extend my thanks to all my classmates and friends who have given me feedback, helped me with measurements and helped me reflect upon my work.

Viktor Bjärnklint, Gothenburg, June 2020

Contents

List of Figures

List of Tables

1

Introduction

This chapter presents the background, aim and limitations of the master's thesis.

1.1 Background

Recent research aims to use a microphone array to get the same binaural signal as would be recorded at the eardrums of a human head. Such a microphone array could be used for capturing 3D sound for virtual reality playback, e.g. when recording a concert or when holding a video conference in virtual reality (VR). Binaural Cue Coding, BCC, is a subject that describes how a multi-channel audio signal can be downmixed to a single channel and then upmixed again to the original number of channels, restoring the spatial impression of the original recording. This is done by analysing the original signal in terms of a set of psychoacoustically relevant cues like inter-channel level difference and inter-channel time difference and storing this as low bitrate side-information. The theory is that the side-information contains all perceptually relevant information to reproduce the spatial properties of the original sound stage.

1.2 Aim

The idea of this thesis is to apply BCC analysis to binaural audio. The binaural recording can come either from a microphone array, from a real (or dummy model) head or an audio track convolved with a head related impulse response. As such, the aim of the thesis is to implement a BCC scheme presented by Christof Faller in his Phd thesis *Parametric Coding of Spatial Audio* and evaluate how binaural signals are affected by the BCC processing.

1.3 Limitations

Christof Faller [1] presents a variety of processing methods for different audio frameworks within the limits of binuaral cue coding. This includes subjects as: hybrid BCC (which mixes BCC with more traditional audio coders to achieve a higher audio quality at the cost of higher bitrates), C-to-E BCC (which presents measures for mixing between a different numbers of channels), low-complexity implementation of

BCC (which presents a straight forward implementation of the coding scheme) and BCC for flexible rendering (which makes it possible for the decoder to render the desired auditory image).

The master's thesis has been focused on the low-complexity implementation of BCC since this method is well suited for the available programming environment and aligns with the aim of the project. The number of channels analysed with BCC has been restricted to two channels, a decision which is supported by type of signal that is to be analysed, binuaral signals, and simultaneously reduces the amount of code that needs to be created. As the focus of the thesis is centred around the input and output of the BCC, the bitrate will not be considered as defining property of the constructed code.

2

Theory

This section covers the concepts of human hearing, spatial hearing, introduces some of the key audio processing components and the concept of binaural cue coding.

2.1 Hearing

The human hearing can be split into two major concepts. How the sound is changed in its path to the ears and how the sound is processed in the inner ear.

2.1.1 The Human Ear

The human ear can be separated into three parts: the outer, the middle and the inner ear. The outer ear is composed of the pinna, the visible part of the ear, the ear canal and the eardrum. The pinna, in combination with the torso, contribute to the directional properties relevant for our spatial perception [2]. They influence mainly higher frequencies above 1 kHz. The middle ear transmits the vibrations of the eardrum to the inner ear. The inner ear analyses sound with the help of the cochlea and the basilar membrane.

2.1.2 Spatial Hearing

The ability to determine the origin of sound is a property of the human hearing and subsequently, the two ears. The term binaural hearing refers to that two ears are used to evaluate the sound [3]. The ability to localise sound is generally called localisation and is mainly based on three concepts:

- The inter-aural time difference, ITD
- The inter-aural level difference, ILD
- The spectral cues related to shape of the pinna and torso

[4]

When talking about spatial hearing the horizontal and median plane are of major concern. The planes are used to describe the position of a sound source. The horizontal plane refers to sound coming from the left, right, in front or from behind of the listener, and the median plane refers to sound coming from above, below,

behind or in front of the listener [3]. The position of a sound source is typically defined using angles. A sound source located in front of the listener is defined at 0° azimuth and as such, ±90° azimuth on the horizontal plane would mean a position to the far left/right directly perpendicular to the left/right ear.

The inter-aural level difference is a consequence of shadowing. Shadowing is caused by that the head itself acts as a barrier for one of the ears and as such the level is higher at the ear that is not shadowed [4]. The effect is more pronounced the further to the left/right the sound source is positioned and its maximum effect can be found at $\pm\,90°$ azimuth. However, as the size of the wavelength varies with frequency this effect is not as apparent for lower frequencies as these waves diffract around the head. In general the ILDs are negligible below 500 Hz but can be as big as 20 dB at higher frequencies.

The inter-aural time differences is a function of the distance between the ear's individual position and the position of the sound source. The maximum distance is achieved from a sound source positioned at $\pm\,90°$ azimuth on the horizontal plane, and is proportional to the distance between the ears. The distance correspond to an approximate travelling time of 690 μs. If a sinusoidal tone is studied the time difference can be related to the phase difference between the two ears. However, as the frequency increases it gets increasingly difficult for our ears to identify whether the phase difference is due to the distance in space or simply a consequence of the sinusoidal cycle. These difficulties start to appear when half the wavelength of interest is the same length as the distance between the ears, this is usually around 750 Hz (assuming an the average distance of 23 cm between the ears). Assuming a source position to the left, it gets difficult to determine if the sinusoidal tone at the right ear is half a wavelength ahead or behind the left ear. As the frequency increases, and the wavelength decreases, this process becomes even more difficult. Around 1500 Hz the ability to analyse the phase difference of the two signals is completely lost [4]. Instead, above 1500 Hz, the phase difference between the envelope at the right and left ear is used for determining the ITD [5].

The ILD, ITD and the spectral colouration (from the shape of the torso and pinnae) all contribute to our ability to localise sound. They can also be said to act as a filter that changes with the position of the sound source [6]. This is more generally termed head-related transfer functions, HRTF, in the frequency domain and corresponds to the head-related impulse response, HRIR, in the time domain. Since the HRTF is dependent on the anatomical configuration of each individual, these functions are individual and a HRTF defined for one person might not result in the same spatial image for another person. Through the use of HRTFs spatial localisation can be achieved for signals that is played through headphones. However, it has been found that this is highly dependent on accurate representation of inter-aural phase relations below 1000 Hz and inter-aural level differences over the whole frequency range [7].

2.2 Audio Processing Tools

The section describes the ideas and concepts of the tools used within BCC to process the audio signals.

2.2.1 The Short-Time Fourier Transform

The short-time Fourier transform, STFT, is commonly used to analyse a signal in time and frequency domain, giving the user information about how the signal's spectral properties changes over time [8]. The method is commonly defined according to Equation 2.1

$$X_m(\omega) = \sum_{n=0}^{N-1} x(n)w(n - mR)e^{-j\omega n} \tag{2.1}$$

where $x(n)$ is the input signal at time n, N is the length of the input signal, $w(n)$ is a window function of length M, R is the hop-size (the step-size between each block) between each successive Fourier transform and $X_m(\omega)$ is the Fourier transform of each window around time mR, where m indicates the current window. The sum of successive Fourier transforms equals the Fourier transform of the whole signal if the windows are designed such that a constant overlap-add equals one at hop-size R, in other terms, if Equation 2.2

$$\sum_{m=0}^{N-1} w(n - mR) = 1 \tag{2.2}$$

is fulfilled. The idea of the overlap-add criteria is visualised in Figure 2.1.

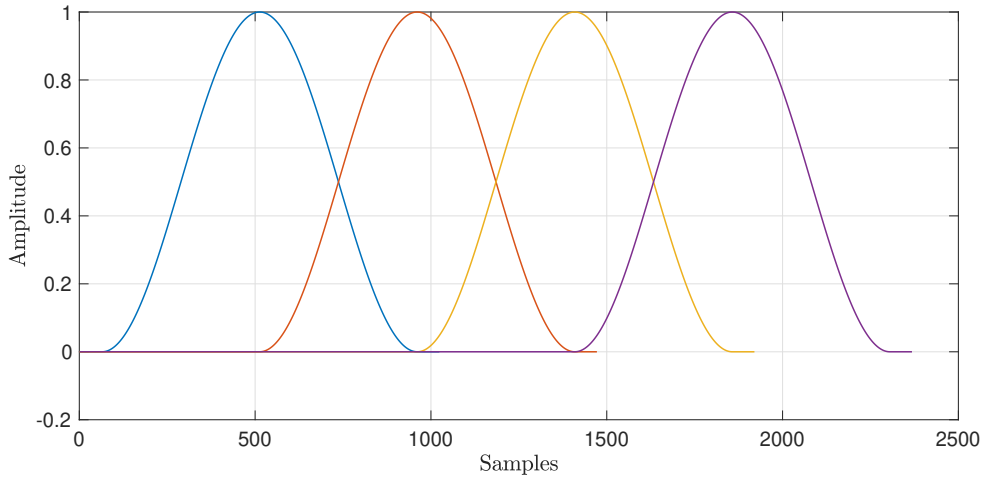


Figure 2.1: The idea of constant overlap-add visualised, each window can be seen to overlap in such a way that summation according to the hop-size would add up to an amplitude of 1.

2.2.2 Critical Bands and Equivalent Rectangular Bandwidth

Critical bands refers to a model of the human hearing and the human perception of sound. The model states that within a bandwidth, corresponding to a frequency of a tone, only noise within the bandwidth contributes to the masking of the tone [9]. This concept led to the creation of the Critical Band-Rate Scale, more commonly named the Bark scale. The Bark scale was determined using probe tones and narrow band maskers [8]. The scale was created with the intention of mimicking how humans process sounds, i.e, it was made such that a broad spectrum is analysed in several smaller parts, where these smaller part corresponds to the critical bands.

The Bark scale was later improved upon with the equivalent rectangular bandwidth, ERB, scale in an attempt improve the model of human hearing analysis [8]. The scale is similar to the Bark scale but is based on a different measurement method, notched noise method, which resulted in a slightly different appearance assumed to be a better approximate to the human analysis of sound.

2.2.3 Signal Comparison and Manipulation

A comparison between two signal levels is commonly done by calculating the level difference between the two signals. The inter-aural level difference is calculated by the same method, also shown in Equation 2.3 [3],

$$\Delta L(f) = 10 \log \left| \frac{X_1(f)}{X_2(f)} \right|^2 \tag{2.3}$$

where $X_1(f)$ and $X_2(f)$ represents the signals in frequency domain.

The inter-aural time difference between two signals can generally be determined by the inter-aural cross correlation [3]. The normalised inter-aural cross correlation is defined according to Equation 2.4:

$$\Psi_{x_{l,r}}(\tau) = \frac{\int_{t=-\infty}^{\infty} x_l(t) x_r(t+\tau) dt}{\sqrt{\int_{t=-\infty}^{\infty} x_l^2(t) dt \int_{t=-\infty}^{\infty} x_r(t)^2 dt}} \tag{2.4}$$

where τ is the internal delay. The maximum value of the cross correlation corresponds to the inter-aural coherence and the corresponding x-coordinate gives the ITD. The inter-aural coherence, IC, is a measure of how similar the two signals are to each other. The value can vary between -1 and 1, where 1 corresponds to a completely coherent set of signals, a value of zero indicates that the signals are not coherent and -1 that the signals are coherent but of opposite phase.

The time signature of a signal can be manipulated in frequency domain by using Equation 2.5:

$$X_2 = X_1 e^{\frac{-j2\pi i\tau}{N}} \tag{2.5}$$

where X_2 is the delayed signal, X_1 is the base signal, N is the signal length, τ is the delay in samples and $i \in (0 \leq i \leq N-1)$ is the index vector. A sinusoid delayed by 50 samples with this method is displayed in Figure 2.2. If the signal is not zero

padded by at least the same amount of samples as the intended delay, the signal will wrap around.

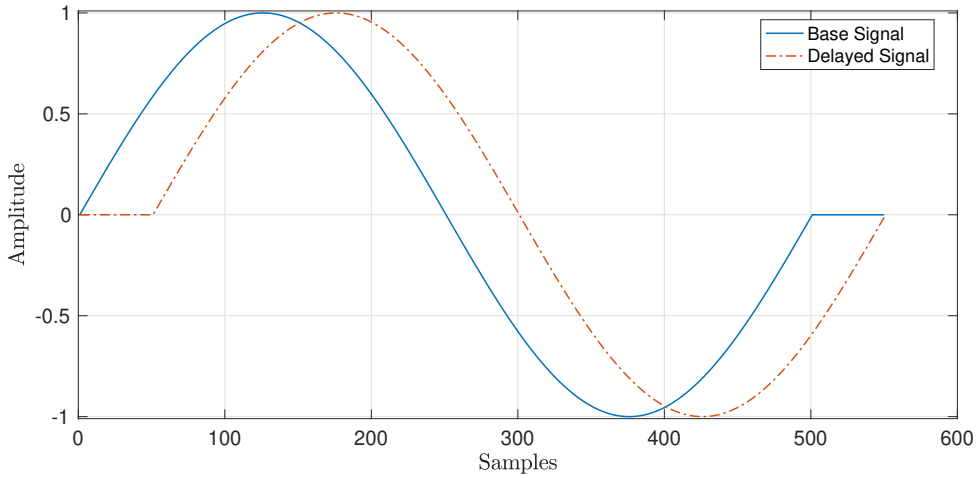


Figure 2.2: Two signals demonstrating the result of using Equation 2.5 to delay a signal.

2.3 Description of Key Components Within the Low-Complexity Implementation of BCC

Binaural cue coding makes use of the previously mentioned ILD, ITD and IC. The signal structure is analysed with these properties in mind, determining the difference in level, time and to which degree the channels within the signal are correlated. The properties are referred to as inter-channel level differences, ICLD, inter-channel time differences, ICTD, and inter-channel coherence, ICC. These properties are also referred to as side-information in the introduction. BCC is divided into two stages: encoding and decoding [1].

The encoding consists of three components:

1. Splitting the signal into blocks through the short-time Fourier transform, allowing the signal to be analysed in time and frequency domain and dividing each block into sub-bands using the ERB scale to mimic the human perception of sound.
2. Estimating the ICLD, ICTD and ICC cues.
3. Downmixing the channels to one channel.

The decoding consists of three components:

1. Generate spectral coefficients based on the analysed ICLD, ICTD and ICC cues.
2. Synthesise the output channels using the spectral coefficients and the sum signal.

3. Re-organising the signal into blocks, removing the sub-band structure, and preforming an inverse short-time Fourier transform to create a continuous signal.

A block diagram of the BCC scheme can be seen in Figure 2.3.

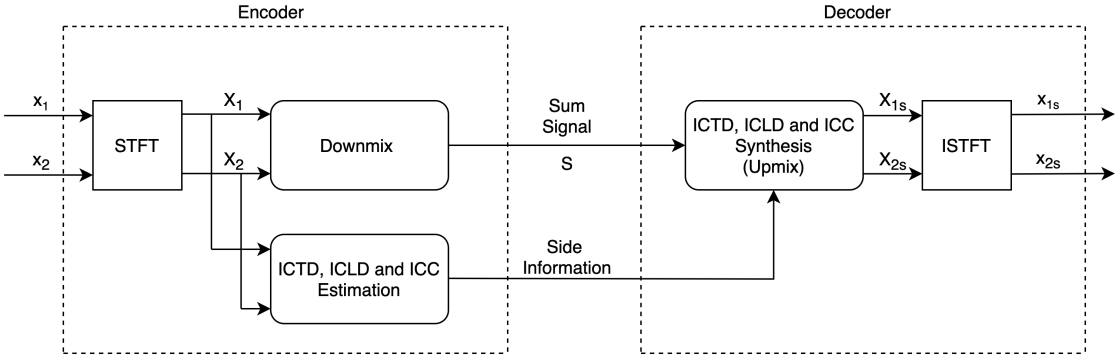


Figure 2.3: The low-complexity BCC scheme.

The low-complexity BCC algorithm estimates the ICLDs using the principles previously described in Equation 2.3. The ICTDs, however, is estimated in frequency domain and not in time domain according to the method described by Blauert in Equation 2.4. The low-complexity implementation proposes that the ICTDs should be estimated by evaluating the phase difference between the channels present in the signal and then converting it into phase delay, obtaining the delay in seconds. The process is done through two methods which is more thoroughly explained in method Section 3.1.4.

The process of estimating and synthesising the ICTD cues through phase delay has later been found to create artefacts in the audio. The reasons for this is not entirely clear but two main hypotheses has been made: 1) The processes requires accurate unwrapping of each bin present in the sub-band [10]. 2) It could be a consequence of blocks overlapping due to the variation of estimated time delays [11]. As such, various methods have been introduced to work around this problem. In some cases pure phase synthesis have been proposed (*Parametric Coding of Stereo Audio* [10]) and in other cases more complex algorithms has been developed which aims to correct the problems with the original implementation (*Improved Time Delay Analysis/Synthesis for Parametric Stereo Audio Coding* [11], *Parametric Binaural Audio Coding* [12]). However, since these methods are not included in the low-complexity implementation they are not explored further in this thesis.

3

Method

The section describes the implementation of the low-complexity BCC scheme presented by Faller and the method used to evaluate its performance.

3.1 Encoding

In order to evaluate the relation between the channels in the input signal and to save bitrate the signal is encoded [1]. Encoding implies that the signal is first evaluated in terms of inter-channel level difference, inter-channel time difference, inter-channel coherence and finally downmixed to a single channel. Each sub-band get one ICTD, ICLD and ICC coefficient to describe its properties.

3.1.1 Short-Time Fourier Transform

In order to divide the signal into blocks, and to enable reconstruction of the signal after it has been processed, a short-time Fourier Transform with constant overlap-add is used [8]. According to Faller [1] a Hann window with zero-padding on both sides, as can be seen in Equation 3.1,

$$w_a(n) = \begin{cases} 0, & \text{for } 0 \leq n < Z \text{ or} \\ & N - Z \leq n < N \\ \sin^2\left(\frac{(n-Z)\pi}{W}\right), & \text{for } Z \leq n < Z + W \end{cases} \tag{3.1}$$

is used. The zero padding enables non-circular time-shifts and linear filtering within the short-time Fourier transform domain [1]. The block length, N, is equal to $2Z + W$ where Z is the number of zeros present before and after the window, and W is the length of the window. After the signal has been processed through the short-time Fourier transform it is denoted as $X_c(k, i)$, where k indicates the current block and i indicates the current sample within the block. To enable the signal to be reconstructed without any losses a constant overlap-add is implemented with a hop-size of $W/2$.

3.1.2 Creating Sub-Bands

To save bitrate, each binaural cue is given to a set of frequencies instead of appending a binaural cue to each frequency bin [1]. Each set of frequencies is called a sub-band, the sub-band has a bandwidth equivalent to two times the ERB. The centre

frequencies of the bandwidth were chosen in such a way that no sub-band overlap the other. The bandwidths were calculated with equation 3.2,

$$ERB = 2(24.7(4.37f_c + 1)) \tag{3.2}$$

were ERB is the bandwidth in Hz and f_c is the centre frequency given in kHz [4].

The creation of the sub-bands creates a rather complex and un-intuitive signal structure. The signal structure is as follows: signal (x), block (k), sub-band (b) and bin ($i \in (I_0 \leq i \leq I)$, where I_0 is the first frequency bin of the sub-band and I is the last frequency bin). The signal structure is exemplified in Figure 3.1.

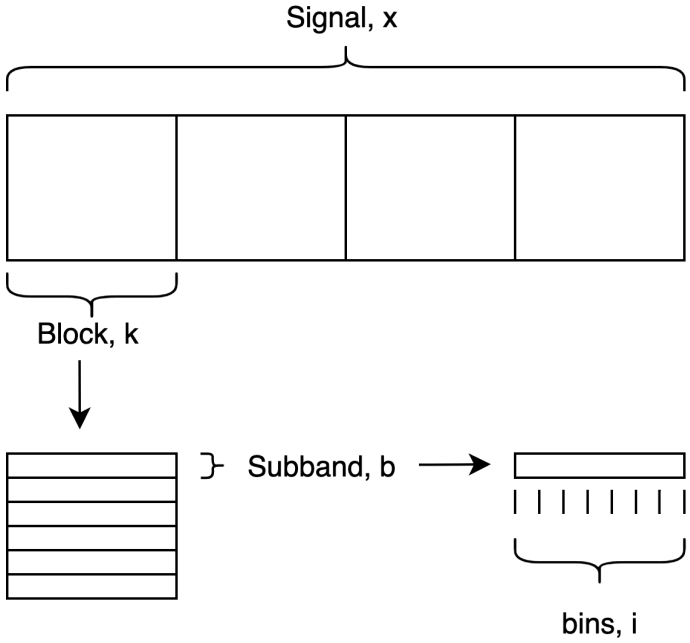


Figure 3.1: Structure of signal.

After the signal has been divided into sub-bands it is denoted as: $X_c(k, b, i)$.

3.1.3 Sub-Band Power Estimates

The sub-band power estimate is a summation of the power for all frequencies in one sub-band. The summation is done for both channels and for the total power of the two channels [1]. The sub-band power estimates are used as a tool to create the inter-channel level difference. The estimates are calculated according to Equation 3.3,

$$p_c(k, b) = \sum_{i=I_0}^{I} |X_c(k, b, i)|^2 \tag{3.3}$$

where $X_c(k, b, i)$ is the signal spectrum for each channel, block, sub-band, and $p_c(k, b)$ represents the sub-band power estimate for the current block and sub-band.

3.1.4 Estimation of ICTD, ICLD and ICC

The inter-channel level differences, $\Delta L_{12}(k,b)$, are estimated with the sub-band power estimates [1]. The sub-band power estimates are used to create a ratio expressed in dB values. The ratio is calculated according to Equation 3.4,

$$\Delta L_{12}(k,b) = 10\log_{10}\left(\frac{p_2(k,b)}{p_1(k,b)}\right). \tag{3.4}$$

The maximum and minimum level difference is set to $L_{12}(k,b) = \pm 18$ dB. The threshold roughly corresponds to the maximum level difference that occurs when one ear is completely shadowed by the head.

The inter-channel time differences, $\tau_{1c}(k,b)$, are typically estimated in two segments: one below approximately 1500 Hz and one above. Below 1500 Hz it is estimated using the mean of the phase delay between the two channels [1] and above using the slope of the linear regression of the phase difference, modelling the group delay of the signal. However, during the implementation of the BCC scheme it was found that lowering this limit to around 650 Hz generated more stable results and was thus used instead. The time difference is calculated according to Equation 3.5,

$$\tau_{12}(k,b) = \begin{cases} \text{mean}_{I_0}^{I}\left(\frac{-\frac{\angle_2(k,b,i)}{\angle_1(k,b,i)}}{2\pi f(i)}\right), & \text{for } f \leq 650\text{Hz} \\ \frac{1}{2\pi}\arg(\frac{\angle_2(k,b,i)}{\angle_1(k,b,i)}), & \text{for } f > 650\text{Hz} \end{cases} \tag{3.5}$$

where $\angle_c(k,i)$ is the angle of each frequency within a block and sub-band, and arg denotes the process of calculating the first order linear regression of the phase between the two channels and extracting the slope coefficient as an estimate of the time difference. The maximum and minimum time difference is set to $\tau_{12} = \pm 800$ ms, which is equivalent to roughly ± 35 samples for a sampling frequency of 44.1 kHz.

The inter-channel coherence, $c(k,b)$, is estimated according to Equation 3.6 through Equation 3.9 [1],

$$\Phi_{uv}(k,b,i) = \alpha X_u(k,b,i)X_v^*(k,b,i) + (1-\alpha)\Phi_{uv}(k-1,b,i) \tag{3.6}$$

$$\alpha = \frac{N}{T_{ICC} f_s} \tag{3.7}$$

$$c(k,b,i) = \sqrt{\frac{|\Phi_{uv}(k,b,i)\Phi_{uv}^*(k,b,i)|}{\Phi_{uu}(k,b,i)\Phi_{vv}(k,b,i)}} \tag{3.8}$$

$$c(k,b) = \frac{\sum_{I_0}^{I}[c(k,b,i)(\Phi_{uu}(k,b,i) + \Phi_{vv}(k,b,i))]}{\sum_{I_0}^{I}(\Phi_{uu}(k,b,i) + \Phi_{vv}(k,b,i)} \tag{3.9}$$

where u and v represent channel 1 and channel 2, α is a smoothing factor which is implemented with a time constant of $T_{ICC} = 50$ ms, f_s denotes the sampling frequency and * denotes the complex conjugate. The coherence values lie between 0 and 1.

3.1.5 Downmix

The downmix is preformed by summing the two channels and multiplying a gain factor to each bin in the current sub-band and block. The gain factor, $e_b(k, b, i)$, is calculated according to Equation 3.10,

$$e(k, b, i) = \sqrt{\frac{\sum_{c=1}^{C} |X_c(k, b, i)|^2}{|\sum_{c=1}^{C} X_c(k, b, i)|^2}} \tag{3.10}$$

and the sum signal, $S(k, b, i)$, according to 3.11,

$$S(k, b, i) = e(k, b, i) \sum_{c=1}^{C} X_c(k, b, i). \tag{3.11}$$

3.2 Decoding

Using the ICLD, ICTD, ICC cues and the generated sum signal a synthesis of the signal can be constructed, this is called upmixing. The upmixing attempts to create the exact same signal from these input parameters by creating time and level coefficients that are multiplied with the sum signal[1].

3.2.1 ICLD Synthesis

The ICLD synthesis uses the previously calculated ICLD cues, Equation 3.4, to create the level differences for all spectral components within each sub-band and block. The ICLD synthesis also applies a decorrelation effect to the spectral components if the value of the ICC cue for the current block and sub-band is less than 1 [1]. The reference channel, Channel 1, ICLD synthesis is preformed according to Equation 3.12,

$$F_1(k, b, i) = \frac{1}{\sqrt{1 + 10^{\frac{\Delta L_{12}(k,b) + r_{12}(b,i)}{10}}}}, \tag{3.12}$$

and for Channel 2 according to Equation 3.13,

$$F_2(k, b, i) = F_1(k, b, i) 10^{\frac{\Delta L_{12}(k,b) + r_{12}(k,b,i)}{20}} \tag{3.13}$$

where r_{12} is the decorrelation factor that is calculated in Equation 3.14,

$$r_{12}(k, b, i) = (1 - c(k, b))\hat{r}_{12}(b, i) \tag{3.14}$$

where $\hat{r}_{12}(b, i)$ is a uniformly distributed noise between -3 dB to 3 dB.

3.2.2 ICTD Synthesis

The ICTD synthesis uses the previously calculated ICTD cues, Equation 3.5, to create the time differences between the two channels for all spectral components within each sub-band and block. The ICTD cues are transformed from time values to

samples by multiplying the values with the sampling frequency. The time difference coefficients can be created according to Equation 3.15 and Equation 3.16,

$$G_1(k,b,i) = \exp\left(\frac{-j2\pi(i\tau(b) + g_1(k,b))}{N}\right) \tag{3.15}$$

$$G_2(k,b,i) = \exp\left(\frac{-j2\pi[i(\tau(b) + \tau_{12}(k,b)) + g_2(k,b,i)]}{N}\right) \tag{3.16}$$

where $\tau(b)$, $g_1(k,b,i)$ and $g_2(k,b,i)$ are defined according to Equation 3.17 and Equation 3.18,

$$\tau(b) = -\frac{1}{2}\tau_{12}(k,b) \tag{3.17}$$

$$g_c(k,b) = \frac{g_0}{2}(1 - c(k,b))\sin\left(\frac{2\pi(i - I_0 + 0.5)}{I - I_0} + \frac{2\pi(c-1)}{C}\right). \tag{3.18}$$

Equation 3.17 minimises the maximum phase modification in one channel, ensuring minimal signal distortion in Equation 3.15 and Equation 3.16 [1]. If the coherence, ICC, between the two channels is lower than 1, the decorrelation is handled by imposing a group delay according to the sinusoid described in Equation 3.18. The constant g_0 determines the range of the group delay and is set to 40 as this is recommended for stereo signals [1].

3.2.3 Signal Synthesis

The signal is finally synthesised by multiplying the level coefficients and the time coefficients to the sum signal according to Equation 3.19,

$$\hat{X}_c(k,b,i) = F_c(k,b,i)G_c(k,b,i)S(k,b,i). \tag{3.19}$$

The synthesised signal is then reorganised into blocks, k, and samples, n, and the inverse short time Fourier transform is applied to the partitioned signal in order to obtain the BCC processed signal.

3.3 Evaluating the Implemented Algorithm

The implemented algorithm was evaluated in two parts. Firstly the use of simple signals such as impulses, noise and modulated noise signals were used to determine whether or not the algorithm worked as intended. Secondly more advanced signals such as music, signals convolved with HRIRs and signals recorded with a KEMAR head and torso were used to investigate to what degree the output suffered from artefacts generated by the BCC processing.

To check if the algorithm could synthesise level differences between two channels within a signal, two stimuli were created. The first stimulus was created for the most fundamental problem, constant level difference. This was achieved by creating a white noise signal for channel 1, copying it and multiplying it with a scaling factor for channel 2. The second stimulus used the same approach except that the scaling factor was varied in time. The two generated stimuli can be seen in Figure 3.2.

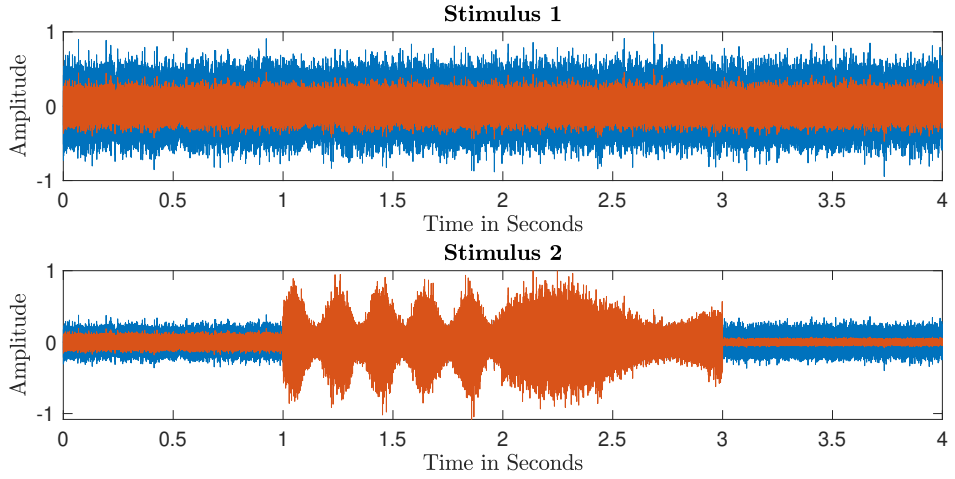


Figure 3.2: The two stimuli used to evaluate how well the program synthesised level differences. The blue signal corresponds to Channel 1 and the orange signal to Channel 2.

To investigate how well the synthesised amplitude difference corresponded to the original, the ratio between the two channels was studied. The ratio is, however, a very sensitive to error as soon as the value in the denominator comes close to zero. To mitigate this, the root mean square, RMS, ratio between the two channels was evaluated.

The stimuli used to evaluate the time differences were centred around delay and delaying different kinds of signals. These signals were: impulses, white noise and modulated white noise. The signals were all created in one channel and then delayed with a chosen delay in the second channel, usually between 1-40 samples. However, the summation utilised to generate the downmixed signal created some problems in the ICTD evaluation, since adding two identical signals, with the exception that one is delayed b samples generates a comb filter effect. To circumvent this, the sum signal was created from a signal containing two channels of the non-delayed signal.

In order to demonstrate what type of signal the BCC algorithm handles well, a signal was created with different signal structure in each channel. Channel 1 contained a sinusoid and Channel 2 contained white noise. The signal was processed with both ICTD and ICLD cues.

The signals, convolved with head related transfer functions, were designed in different ways in order to evaluate how effectively the BCC could synthesise their properties. Four types of audio were convolved: speech, a transient signal (drums), a fleeting audio (slowly playing guitar) and music (pop song).

The speech, drum and guitar signal were convolved such that the audio was perceived as if it was circling the head. The guitar and drum signals were also convolved such that they had fixed positions in space, the fixed positions were located around $\pm 80°$ azimuth. The signals were then added, creating a signal with two perceived sound sources, the guitar to the right and the drums to the left. The inverse was also

created, a signal with the guitar to the left and the drums to the right, to investigate if the eventual artefacts were a construct of the location or the sound. The dual sound source samples were also evaluated as single sound sources (without adding the two sources together) in order to assess how the downmix affects the audio.

Finally a conference meeting was simulated using recordings from a KEMAR head and torso (dummy head model). An artificial meeting was also created through the use of HRIRs and several speech samples. The purpose of simulating a conference meeting was to investigate how well several sources of sound, that might also interrupt and talk simultaneously, is preserved through the BCC processing. In the recorded meeting, people where sitting around a rectangular table with the dummy head on one short end, two people on the long end and one person on the other short end. People were talking on and off, sometimes interrupting each other. The artificial meeting was constructed spatially similar to the recorded meeting, but now the sources were positioned in a perfect semi-circle around the centre (listener). The meeting was constructed such that five different voices read segments of a sentence in chronological order one at a time until every voice had presented its location and characteristics to the listener. After this the voices began speaking on top of each other until everyone was speaking at the same time.

All samples evaluated in the informal listening test were processed in two ways:

1. With ICTD and ICLD synthesis
2. With only ICLD synthesis

This was done in order to investigate to what degree the ICTD analysis and synthesis affected the processed samples.

The samples created and evaluated in this report can be seen in Table 3.1. These samples were not evaluated in a user study due to the brief time that was left when the implemented algorithm reached an acceptable state.

The size of the blocks and the sampling frequency were chosen such that each window would be short enough to represent transient sounds but long enough provide a good spectral representation. This is usually achieved with a window duration between 15 ms and 35 ms for speech signals [13]. As such, the block length was chosen to 1024 samples and the sample frequency to 44.1 kHz, which results in an average window duration of 23 ms. The block length and sampling frequency was applied similarly to all processed signals.

Table 3.1: The samples used in the informal listening test.

Sample Name	Description of Sample Audio
Sample 1	Unprocessed Circling Speech
Sample 2	Processed Circling Speech
Sample 3	Unprocessed Circling Drums
Sample 4	Processed Circling Drums
Sample 5	Unprocessed Circling Guitar
Sample 6	Processed Circling Guitar
Sample 7	Unprocessed Pop Song, Short Segment
Sample 8	Processed Pop Song, Short Segment
Sample 9	Unprocessed Dual Source: Guitar Right, Drums Left
Sample 10	Processed Dual Source: Guitar Right, Drums Left
Sample 11	Unprocessed Dual Source: Guitar Left, Drums Right
Sample 12	Processed Dual Source: Guitar Left, Drums Right
Sample 13	Unprocessed Short Recorded Conference
Sample 14	Processed Short Recorded Conference
Sample 15	Unprocessed Long Recorded Conference
Sample 16	Processed Long Recorded Conference
Sample 17	Unprocessed Artificial Conference
Sample 18	Processed Artificial Conference
Sample 19	Unprocessed Long Pop Song
Sample 20	Processed Long Pop Song

4

Results and Analysis

The chapter presents the results and analyses the performance of the implemented BCC algorithm. The performance is analysed from three key perspectives: to what degree the program synthesises simple level differences, to what degree the program synthesises simple time differences and to what degree the program synthesis more advanced scenarios such as various signals manipulated with HRTFs.

4.1 Simple ICLD Synthesis

To analyse simple ICLD synthesis, signals such as the one showcased in Figure 3.2 were used. The signal length, number of blocks, number of sub-bands, block length and sampling frequency used can be seen in Table 4.1. The result from the synthesised constant level difference can be viewed in Figure 4.1 and Figure 4.2. The scalar used to manipulate the amplitude of the second channel was chosen to 0.5, the signal was then normalised to its maximum value.

Table 4.1: Properties of base signal and BCC processing for simple ICLD analysis.

Signal Length	176 400 Samples
Block Length	1024 Samples
Number of Blocks	393
Number of Sub-Bands	20
Sampling Frequency	44.1 kHz

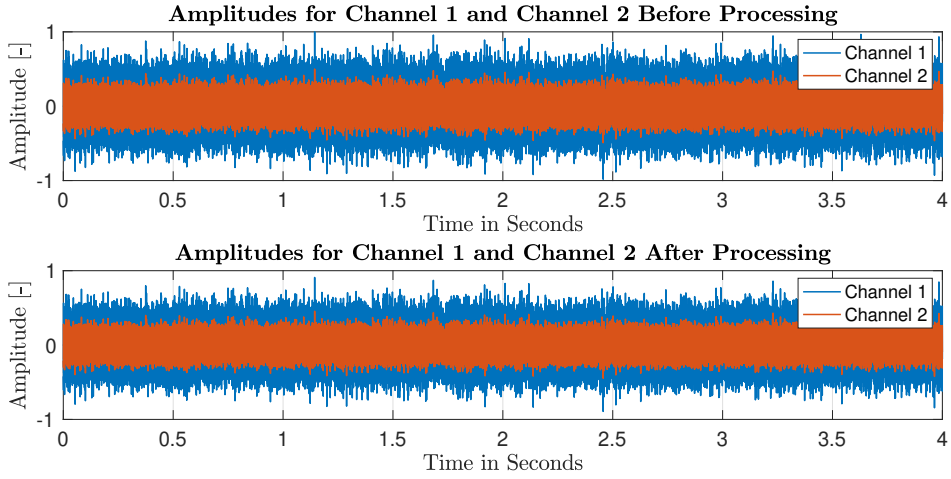


Figure 4.1: The amplitude of two channels where a constant amplitude difference of 0.5 is applied, before and after BCC processing.

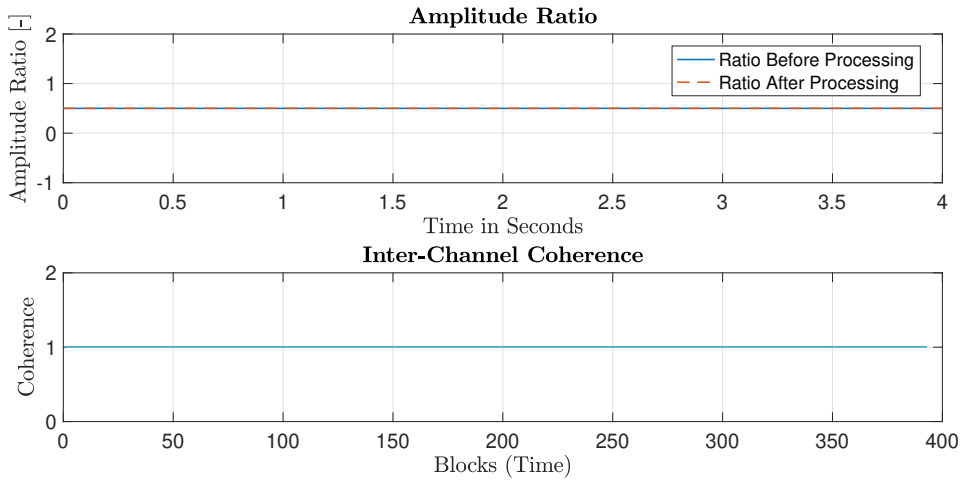


Figure 4.2: The amplitude ratio and the analysed ICC cues between two channels where a constant amplitude difference of 0.5 is applied, before and after BCC processing. The inter-channel coherence displays several frequency bands at once, however, at this specific instance it is not visible as all bands have taken the value 1.

In Figure 4.1 and Figure 4.2 it can be seen how the level difference between the channels have been completely synthesised. The inter-channel coherence is one for all sub-bands and blocks, and as such no decorrelation was used to achieve the result shown above.

The result from the synthesised varied level differences can be viewed in Figure 4.3 and Figure 4.4. The amplitude scalar was divided into three segments: the first segment was given a constant value of 0.5, the second segment was varied sinusoidally between one and three, and the third segment was given the constant

value 0.2. The signal was then normalised to its maximum value. The amplitude ratio was RMS-averaged over two samples.

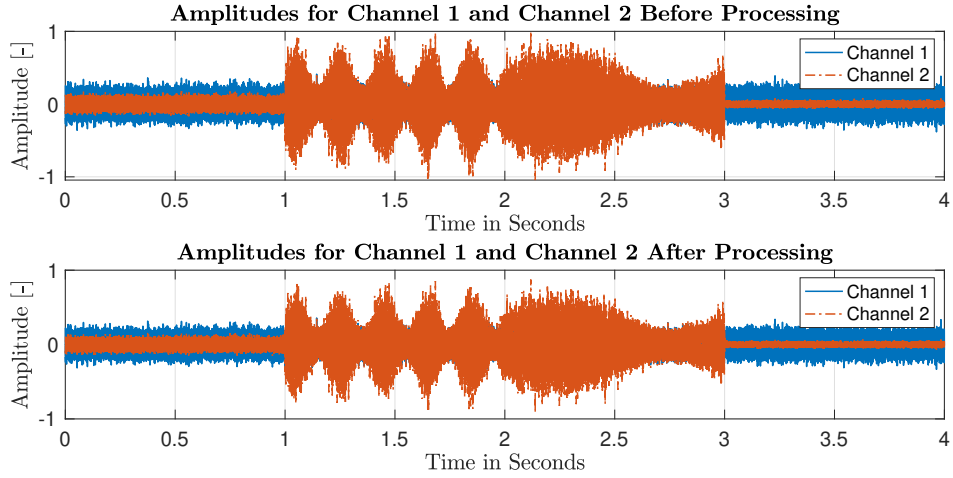


Figure 4.3: The amplitude of two channels with a varied amplitude difference, before and after BCC processing. The amplitude difference is varied in three segments: constant scalar of 0.5, modulated sinusoidally between 1 and 3, and a constant scalar of 0.2.

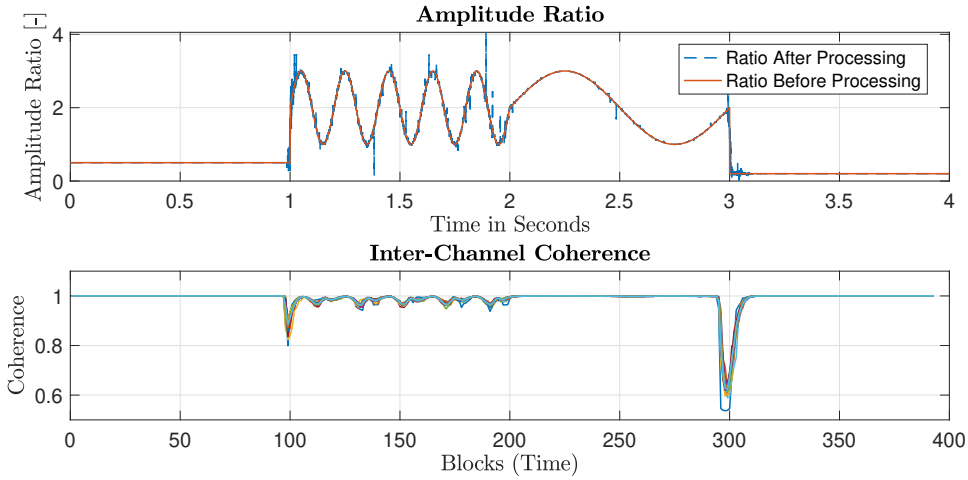


Figure 4.4: The amplitude ratio and analysed ICC cues between two channels with a varied amplitude difference divided into three segments: constant scalar of 0.5, modulated sinusoidally between 1 and 3, and a constant scalar of 0.2, before and after BCC processing. The inter-channel coherence displays several frequency bands at once.

Above, in Figure 4.3 and Figure 4.4, it can be seen how the signal with varied level differences is almost reconstructed perfectly. The discrepancies that exist is located at places where the coherence is lower than one. The behaviour is expected

as the method uses averaged values for synthesis of several spectral components. As such, it is overambitious to ask for perfect reconstruction of a signal that varies significantly in time.

In Figure 4.5 the level coefficients used to manipulate the magnitude spectrum of the sum signal can be seen.

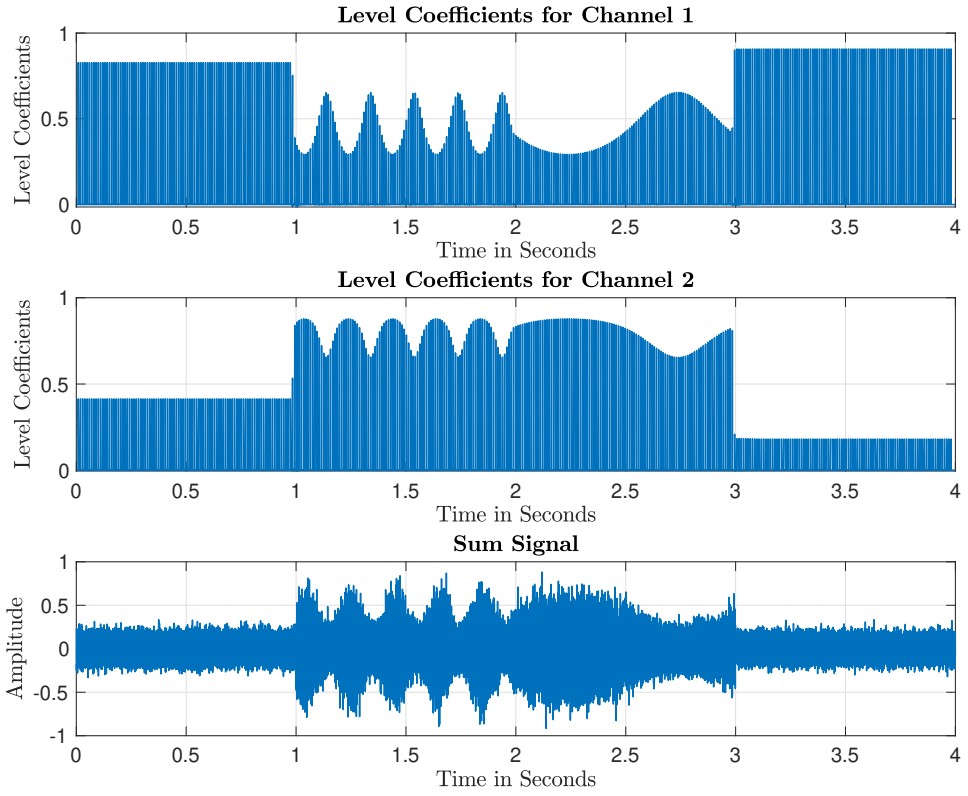


Figure 4.5: Level coefficients used to manipulate the magnitude spectrum of the sum signal in order to create the two channels.

The level coefficients, viewed in Figure 4.5 above, can be seen to behave as expected. Where the sum signal is supposed to be used to synthesise a constant lower/higher amplitude the coefficients have a lower/higher value.

4.2 Simple ICTD Synthesis

To analyse simple ICTD synthesis different signal types were used. These types were: impulses, noise signals and modulated noise signals. The signal length, number of blocks, number of sub-bands, block length and sampling frequency used can be seen in Table 4.2. During the analysis several different delays were used, however, only a delay of 10 samples is showcased here as no additional information is gained from showing any more.

Table 4.2: Properties of base signal and BCC processing for simple ICTD analysis.

Signal Length: Short Signal	1024 Samples
Signal Length: Long Signal	20 290 Samples
Block Length	1024 Samples
Number of Blocks: Short Signal	27
Number of Blocks: Long Signal	45
Number of Sub-Bands	20
Sampling Frequency	44.1 kHz

The result of a synthesised impulse delay can be seen in Figure 4.6 and in Figure 4.7 the analysed and synthesised delay is visualised.

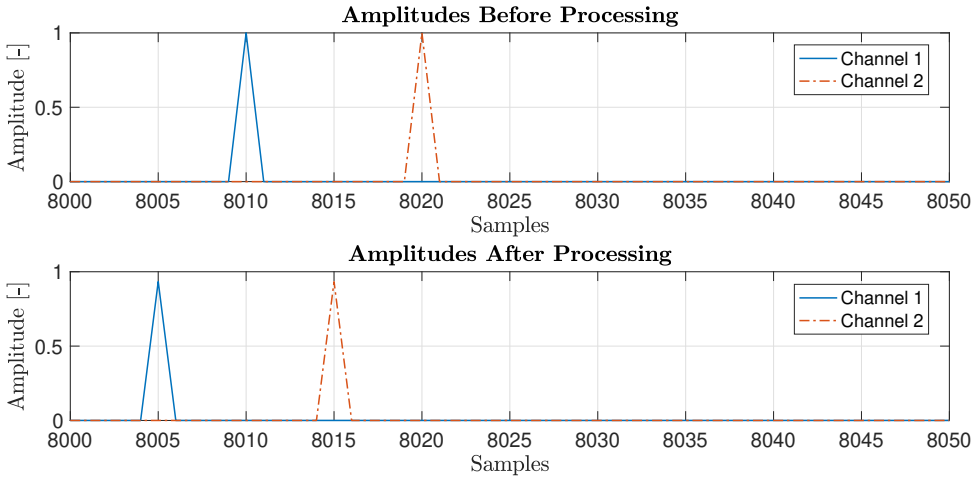


Figure 4.6: The unprocessed and processed impulse signal, with a delayed impulse in Channel 2. The delay between Channel 1 and Channel 2 in the original signal is 10 samples.

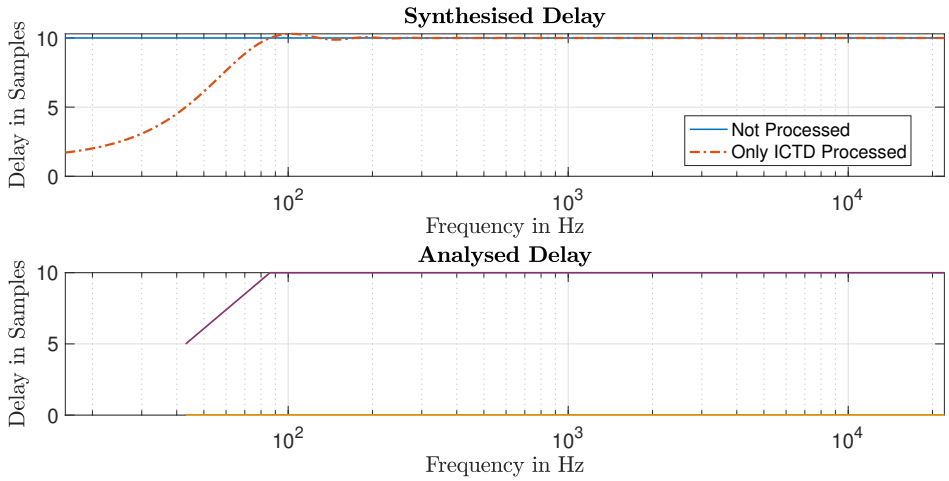


Figure 4.7: The analysed and synthesised delay for the impulse signal. The original delay between Channel 1 and Channel 2 is 10 samples. The different lines in the analysed delay represents different time instances (blocks).

As seen in the figures above, Figure 4.6 and Figure 4.7, the impulse delay is perfectly recreated with the exception of that it is shifted five samples earlier in time. The time shift is excepted as this is a measure taken in order to minimise the number of artefacts in the signal, in accordance to Equation 3.17. The analysed and synthesised delay can be seen to reach the expected delay of 10 samples for almost all frequencies. The lowest subband underestimates the delay during the ICTD analysis. This could be a consequence of the amount of samples in lowest sub-band being too few to successfully analyse the signal. However, this does not appear to have any major impact on the signal. The yellow line at the bottom, at zero delay, corresponds to all blocks where both signals contained zeros and as such no delay is present at these locations in time.

For the noise signal, the result of the synthesised signal can be seen in Figure 4.8, the synthesised delay in Figure 4.9.

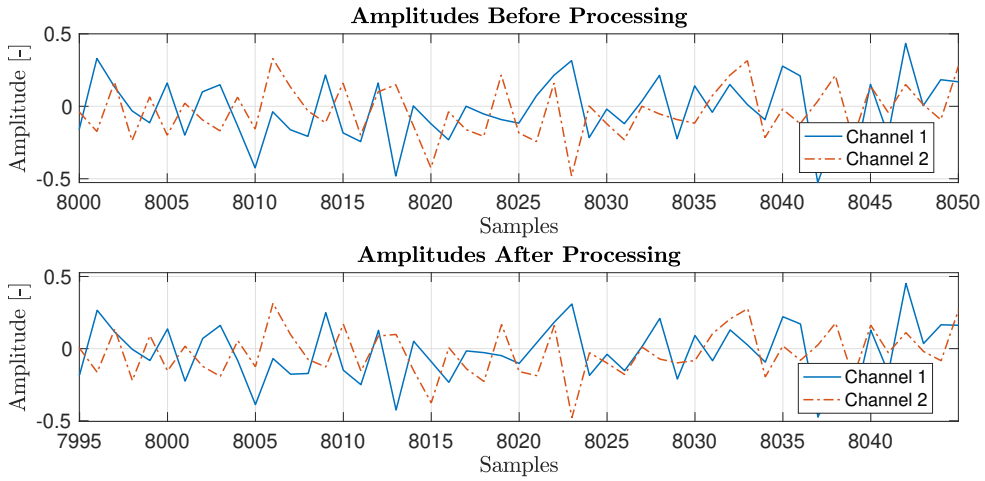


Figure 4.8: The unprocessed and processed noise signal, with a delayed noise in Channel 2. The delay between Channel 1 and Channel 2 in the original signal is 10 samples.

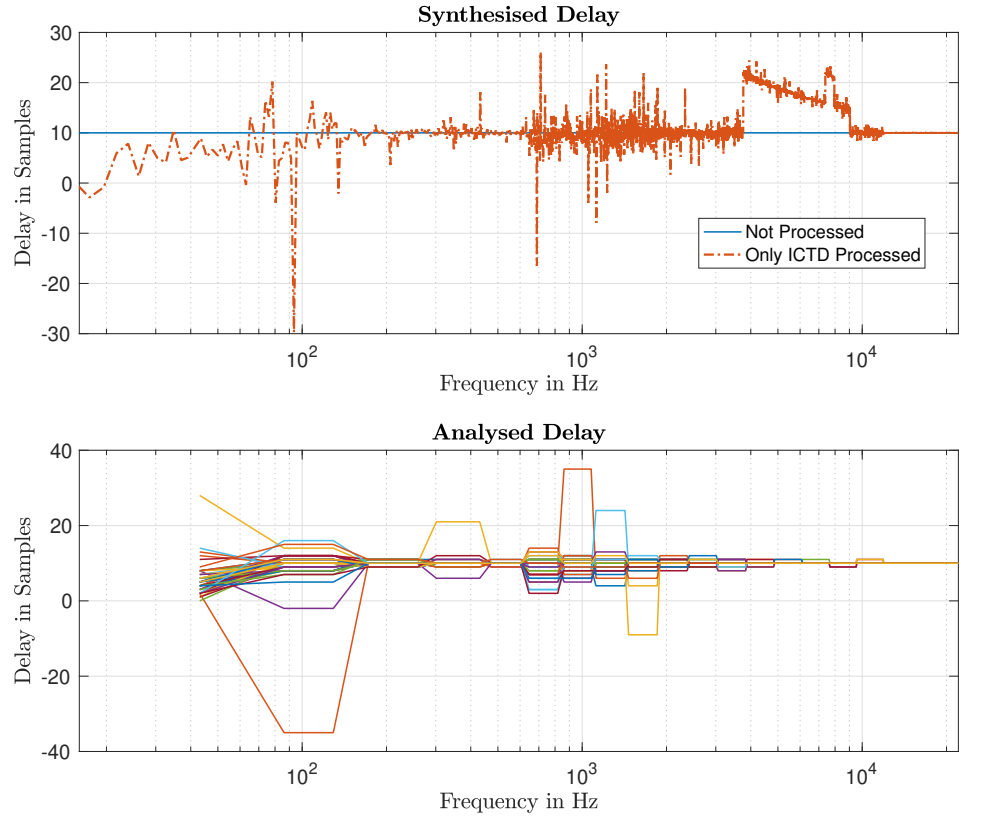


Figure 4.9: The analysed and synthesised delay for the noise signal. The original delay between Channel 1 and Channel 2 is 10 samples. The different lines in the analysed delay represents different time instances (blocks).

Above, in Figure 4.8, the original and synthesised signal appear to be very similar. However, according to the analysed delay in Figure 4.9 some frequencies at different time instances, each line represents one block in the STFT, is not analysed as a delay of 10 samples. The reason for this at lower frequencies (below 200 Hz) could be the low amount of values used to estimate the delay, at frequencies above this the windowing of the signal blocks might affect the phase and unstable unwrapping of the phase might affect what value the phase delay takes. Between the synthesised delay and the analysed delay some discrepancies can be seen, especially between 3000 Hz and 9000 Hz. The synthesised signal deviates a lot from the the intended 10 sample delay in this region, even though the analysed does not. A rigid explanation for this behaviour has not been found, it can simply be confirmed that the error varies with the size of the signal and what signal is used as input. This is showcased in Figure 4.10 where a modulated noise signal is used and in Figure 4.11 where a shorter noise signal is used.

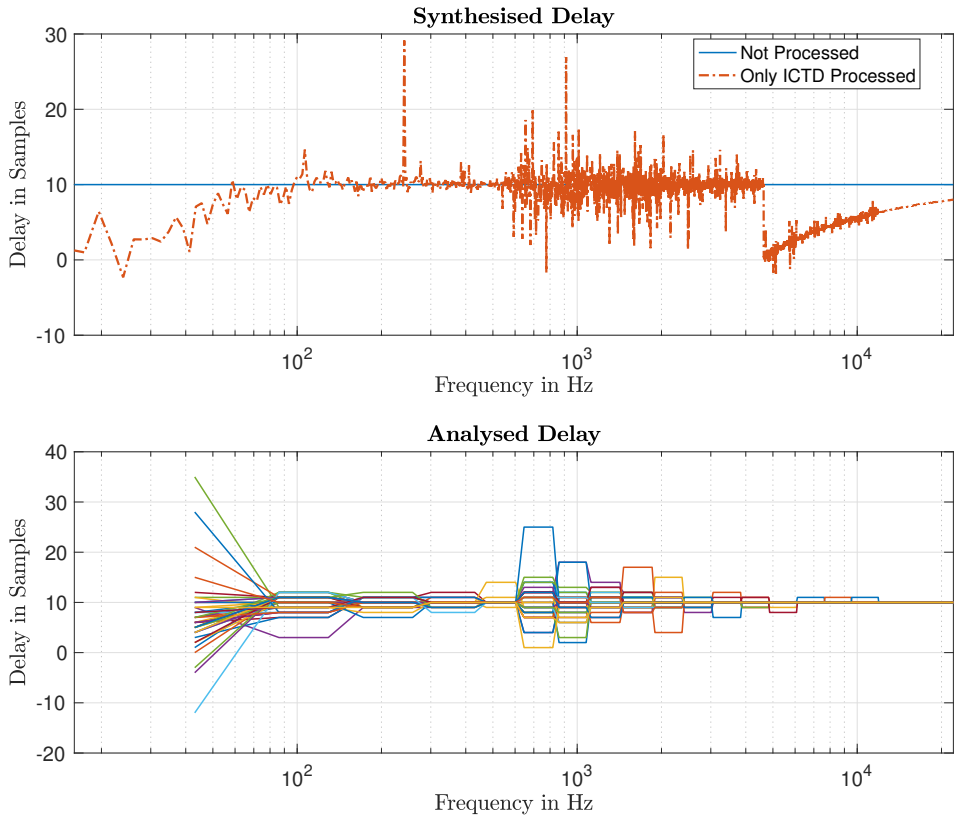


Figure 4.10: The analysed and synthesised delay for the modulated noise signal. The original delay between Channel 1 and Channel 2 is 10 samples. The different lines in the analysed delay represents different time instances (blocks).

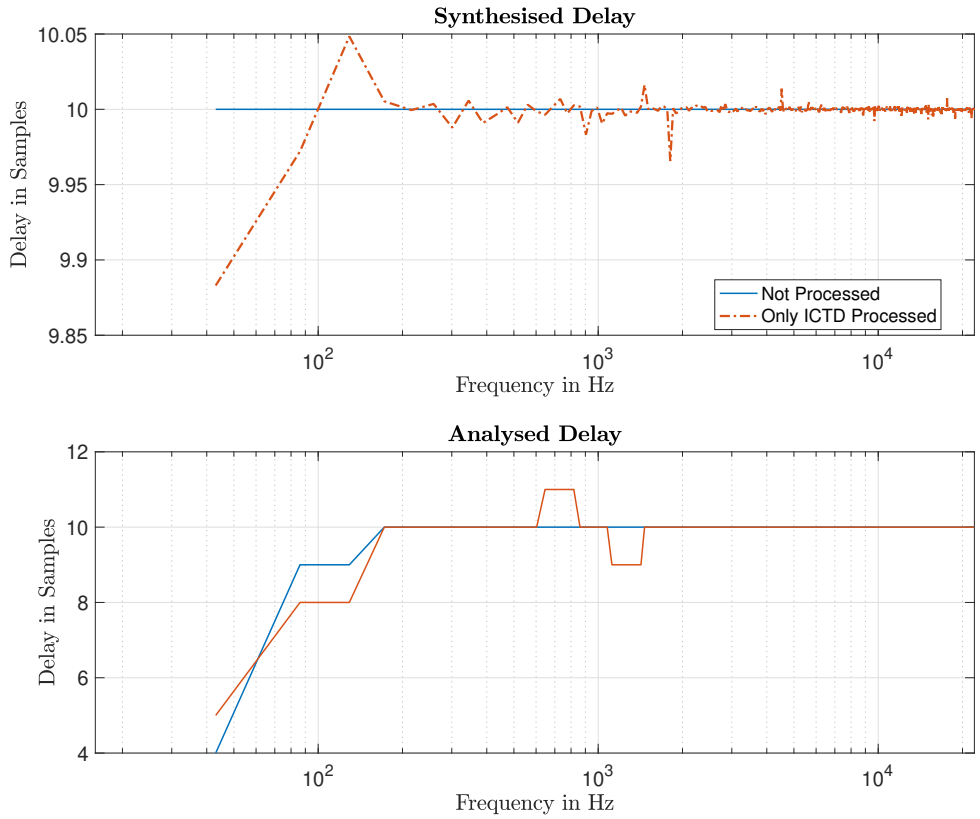


Figure 4.11: The analysed and synthesised delay for the shorter noise signal (compared to the one presented in Figure 4.8). The original delay between Channel 1 and Channel 2 is 10 samples. The different lines in the analysed delay represents different time instances (blocks).

During the analysis of the ICTD synthesis it was also found that the windowing can impact the calculation of a signals phase delay. This is shown in Figure 4.12.

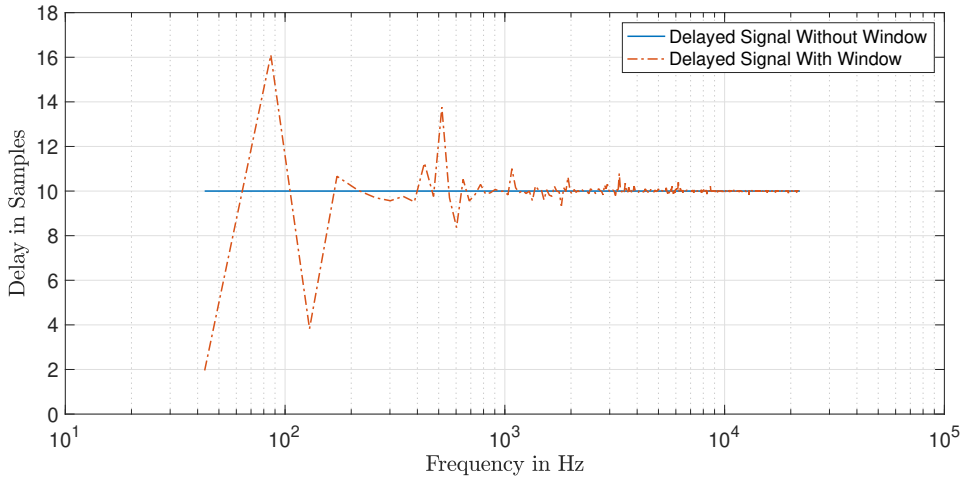


Figure 4.12: A comparison between calculated phase delay for a windowed signal and a signal who has not been windowed.

As seen above in Figure 4.12, the windowing can significantly impact the calculated phase delay within a signal.

4.3 Simple Synthesis: Fundamental Demonstration of BCC Properties

In order to highlight some of the fundamental features of the BCC algorithm, a signal with vastly different properties in Channel 1 and Channel 2 was created. Channel 1 contained a sinusoid and Channel 2 contained white noise. In Figure 4.13, Channel 1 and Channel 2 can be viewed before and after processing. In the figure it can be seen that the signal in Channel 1 has not been synthesised properly. The downmix procedure has added two fundamentally different signals together, creating the foundation for the synthesised signal. The result of this is that the single-channel is unable to represent the two original channels properties, which by extension means that each synthesised signal will be distorted by the other signal's properties. As such, Figure 4.13 can be said to demonstrate how BCC is dependent on that the signals in the two channels need to be similar. If the signal's channels deviate too much from each other, the synthesised content will be distorted due to the downmix procedure.

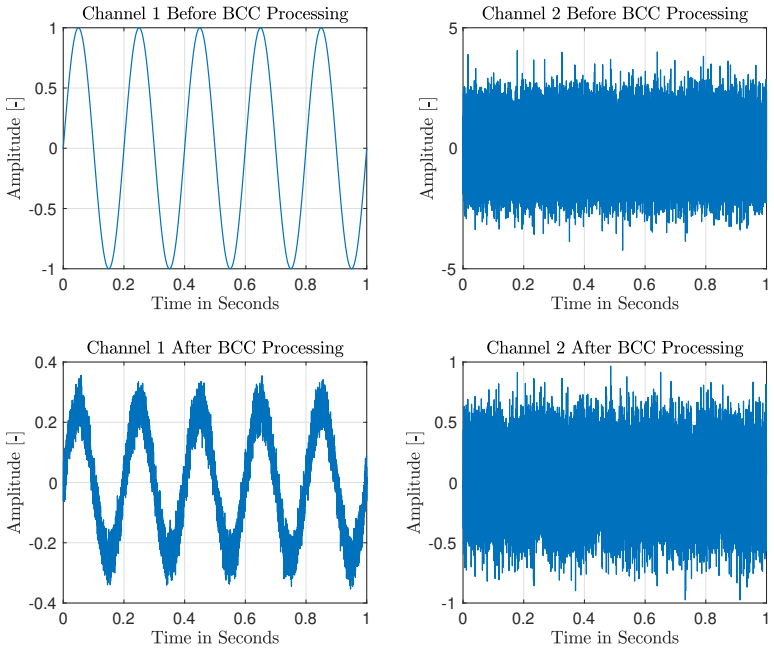


Figure 4.13: A comparison of Channel 1 and Channel 2, before and after processing, for a signal with vastly different properties in each channel.

4.4 Informal Listening Test

The advanced signals, signals convolved with different HRIRs and signals recorded with a KEMAR head and torso, were only evaluated through listening.

4.4.1 Circling Sound

The processed circling speech, drum and guitar signals all contained a low amount of artefacts. A noise, similar to what can be heard when an audio signal clips, could be heard most clearly in the guitar and speech signal. The noise was not heard in the drum signal, but it is possible that it was masked by the noisy characteristic of the signal. The noise is believed to be related to the ICTD processing. The basis for this belief stems from that signals processed without synthesising ICTD resulted in signals totally free from artefacts. However, removing the ICTD cues worsened the spatial reconstruction of the signals. The signals were no longer circling, but moved more like a pendulum between the ears.

4.4.2 Fixed Dual Sound Sources

The processed dual sourced signals exhibited more pronounced artefacts compared to the circling signals. The artefacts are most commonly found attached to the guitar source as a slight vibrato and distortion effect on longer notes and chords. The artefacts could be heard from the drum source when cymbals are struck, but

otherwise the source sounds clean. The artefacts might have been masked by the drums characteristics. When the samples were processed without ICTD analysis and synthesis, a vast amount of the artefacts were no longer present in the audio. However, some of the noise/distortion was still present on the guitar sound source. These artefacts may be inherent of the downmix procedure as two, very different, sound sources were combined.

To further expand upon the idea that the artefacts might originate from the downmixing procedure, the sound sources were processed individually (without adding them together). This resulted in signals with far less artefacts present. The most prominent difference, compared to the original dual sound source samples, could be found as the ICTD analysis and synthesis were excluded, meaning only ICLD was used to process the signal. In this case the processed audio exhibited almost a non-distinguishable processed audio signal (compared to the input signal), with the exception that the sound source was not spatially oriented in the same way as before (the sound was simply heard from the left/right, not the previously $\pm 80°$ azimuth). If ICTD was included in the processing, some artefacts could be heard.

The same artefacts were heard when the locations of the sources were switched, indicating that the ICTD process and/or the downmixing rather than the positioning of the sound sources is related to the amount of artefacts. The location of the guitar and the drums were spatially reconstructed to their original position.

4.4.3 Recorded Meeting

The processed recorded meeting suffered from wavering noise and distortion artefacts mainly in regions where the amplitude was larger (when people were speaking quietly less artefacts could be heard) and for persons located further to the right or left in the dummy head's perspective. The artefacts were not present in the audio processed only with ICLD analysis and synthesis. However, this format severely impacted the spatial reconstruction of the signal. The people speaking, in the ICLD processed audio, were positioned either to the left, right or in the middle of the head, compared to the ICLD and ICTD processed audio which placed each person on a semi-circle in front of the listener.

The audio processed only with ICLD did not exhibit any artefacts that could be related to the downmix procedure. This was probably a consequence of that the persons speaking in the sample rarely spoke at the same time, creating a signal with only "one" sound source, and that the character of each sound source is similar (in contrast to the character of a guitar and a drum set, which are very different from each other).

4.4.4 Artificial Meeting

The artificial meeting processed with ICLD and ICTD, mainly suffered from distortion artefacts on sources located further to the left or right. The artefacts were also enhanced as more people started speaking simultaneously. However, when only ICLD was used to process the audio these artefacts were no longer present in the

signal. Hence, the reasons for these artefacts may be (as speculated earlier) a result of the ICTD analysis and synthesis, and are probably not related to the downmix procedure. The spatial properties of the audio were not reconstructed to a satisfying degree without incorporating the ICTD analysis and synthesis.

4.4.5 Music

The processed pop song suffered from a lot of distortion, even though the spatial properties of the signal were well preserved. The increased amount of artefacts of the artefacts might be related to the difference between the two channels and/or the ICTD processing. The increased amplitude of artefacts are believed to be increased as the amount of sound present in a compressed pop song is quite much, and as such enables for a lot of sound to be distorted. The effect was the same for shorter and longer signals which indicates that the artefacts is not a consequence of something summing up throughout the audio, which Figure 4.9 and Figure 4.11 might indicate.

The same pop song processed without ICTD cues exhibited very slight, almost undetectable artefacts. As the spatial image of the signal is not as wide as the signals convolved with HRIRs or the ones recorded with the KEMAR head and torso, it does not suffer the same loss of its original spatial properties. Synthesising signals such as this, where spatial orientation is not as important, seems to work well with only ICLD cues.

4.5 Final Discussion of the BCC Implementation

In general a few observations of the BCC implementation should be given extra attention. The fact that artefacts seems to occur even though no ICTD processing was used highlights the importance of what types of audio that is processed. If an audio signal (with vast differences between the two channels) is processed, the downmixing procedure will inevitable be the source of some artefacts present in the synthesised audio signal.

Regarding the ICTD processing, it has been noted throughout the simple synthesis and the informal listening test, that this is a source for several of the artefacts seen/heard in the processed audio files. As mentioned earlier, in Section 2.3, this is a subject which has been the focus of several papers in an effort to improve the BCC method. It is possible that the artefacts found in this thesis may be caused by either faulty estimation of the phase delay (due to inaccurate unwrapping of the phase), as this has been noted as an unstable process in the simple synthesis, and/or from blocks overlapping unexpectedly due to the different time delay applied to each window. As there exist methods for solving problems such as this it is entirely possible that the experienced artefacts may avoided using any of the three methods mentioned earlier.

Another area which would interesting to investigate further, in relation to artefacts created in the ICTD process, is related to windowing and the analysed phase delay. In Figure 4.12 it can be seen that windowing, of a segment of a signal, can

significantly alter the analysed phase delay of that segment. This implies that the BCC method may be further improved upon by either investigating new ways of windowing or by applying the window at other stages in the processing to avoid corrupting the phase delay before it is analysed.

5

Conclusion

The BCC compression algorithm can be said to work very well for signals that do not heavily rely on ICTD cues, since these signals can be synthesised properly without the use of ICTD. For signals where the ICTD cues are of more importance, the synthesised signals lack depth and spatial width if ICTD is not included. However, even though including the ICTD restores the perception of the spatial soundscape it also increases the difficulty in synthesising audio without creating artefacts. The artefacts vary in intensity and is most notable in signals with a lot of sound. These artefacts are least noticeable for transient sound such as drums with a single sound source. Where the problems with the ICTD arises has been hard to localise. The hypothesis is that it may be related to the inaccuracies within the ICTD analysis, which is also supported by the different number of approaches that exist to model ICTDs, and the windowing of the blocks.

The method has potential for synthesising VR conference calls or similar high complexity audio where spatial width might help the listener differentiate between different sound sources. The listener(s) of the recorded (and later BCC processed audio) would easily hear and differentiate between who is talking, where this person is located and even follow a discussion between two other people accurately. If the artefacts can be removed, the algorithm would be a valuable tool for transmitting high complexity audio at low bitrates.

Bibliography

[1] C. Faller, "Parametric coding of Spatial Audio," Ph.D. dissertation, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2004.

[2] M. Kleiner, *ACOUSTICS and AUDIO TECHNOLOGY*, third edit ed. J. Ross Publishing, 2012.

[3] J. Blauert, *The technology of binaural listening*, 2013.

[4] B. C. Moore, *An Introduction to the Psychology of Hearing*, fourth edi ed. Academic Press, 1997.

[5] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, american ed. The Massachusetts Institute of Technology, 1983.

[6] X.-l. Zhong and B.-s. Xie, "Head-Related Transfer Functions and Virtual Auditory Display," 2014.

[7] R. M. Warren, *Auditory Perception: A New Analysis and Synthesis*. The Press Syndicate of The University of Cambridge, 1999.

[8] J. O. Smith, *Spectral Audio Signal Processing*, 2011. [Online]. Available: http://books.w3k.org/

[9] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*, second upd ed. Springer-Verlag, 1999.

[10] J. Breebaart, S. Van De Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *Eurasip Journal on Applied Signal Processing*, 2005.

[11] C. Tournery and C. Faller, "Improved time delay analysis/synthesis for parametric stereo audio coding," *Audio Engineering Society - 120th Convention Spring Preprints 2006*, vol. 2, pp. 990–998, 2006.

[12] P. Ojala, M. Tammi, and M. Vilermo, "Parametric binaural audio coding," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 393–396, 2010.

[13] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," in *4th International Conference on Signal Processing and Communication Systems, ICSPCS'2010 - Proceedings*, 2010.