Filtering with spatial parameters in B-format audio streams

Viktor Sannum
dat11vsa@student.lu.se

Department of Electrical and Information Technology
Lund University

Supervisor: Mikael Swartling

Examiner: Nedelko Grbic

June 19, 2018

Printed in Sweden
Tryckeriet i E-huset, Lund

Abstract

The B-format is an audio format capable of reproducing full spherical surround audio, meaning sounds can appear as if coming from any direction around the listener.

This thesis investigates an approach to quantify and manipulate the spatial information carried in B-format audio signals. It describes an analysis method and a corresponding model for quantifying the spatial data.

Two types of applications of the analysis method and model are then presented: a data mapping acting as a visualization of B-format audio signals and a set of filter types acting on B-format signals using the spatial properties from the analysis and model as input parameters.

The B-format has gotten a surge of interest recently, much due to the rise of virtual reality where the flexibility of the format has proven useful. Compact B-format compatibly microphone arrays allows for easy capturing real world scenes while the simplicity of the format makes for an an easily processable intermediate format.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background

In audio engineering and signal processing, one of the major tasks is to enhance or extract desired parts of a signal while removing or attenuate unwanted parts, or noise. There exist a multitude of techniques and tools for this depending on the types of signal, noise and desired quality of the result. Noise such as low frequency hum from the electricity grid or air condition systems can often be reduced with high-pass filters which attenuate the frequencies contaminated by the noise. By adjusting the equalizer in some music playing software, a listener can in a rough way adjust the balance of instruments in a tune to his or her liking. Or if a sound source in a signal is known to be periodic, sound from other sources can be suppressed by muting the signal while the source of interest is silent. Most of these techniques however, require some knowledge about the frequency content and timing of audio- and noise sources in the signal. This can be problematic, especially in real time applications where these parameters has to be known, either in advance, or by some analytic model capable of telling noise from signal in real time as well.

1.1.1 Spatial audio

A spatial audio signal contains not only the audio signal itself but also some spatial information about the sound scape as well. This may be things such as the position of sources which generated the audio signal and or information about the environment around them. The spatial information may be abstract, encoded as differences between multiple, concurrent audio channels or it may be transported out-of-band, describing the spatial information in a more direct manner. Mostly however, the term spatial audio refers to the former type. One basic example is the stereo hearing which is common among humans and animals with two ears. The two ears generate a signal each and the brain, having an internal model of the position and orientation of the ears, can process the audio signals to estimate directions of various sound sources in the vicinity. A dog bark which sounds louder in the left ear probably comes from a dog somewhere to the left of the subject.

Most standardized spatial audio formats are defined by how the signal is generated, how the channels of the signal are to be interpreted and possibly a set of

operations by which the signal can be modified or transformed into other formats. Two examples of such formats are two channel stereo and 5.1 surround sound, commonly used in music recordings and cinema. In the most basic sense these formats consists of two and six audio channels, respectively, intended to be played back on equally many speakers positioned in a specific configuration. The signals are usually produced by positioning one directional microphone per channel pointing in the same direction as the matching loudspeaker relative to the listener. This is usually enough to give the listener a satisfactory sense of audio source directionality.There are other spatial audio formats such as binaural stereo and DirAC, which are designed to reproduce the spatial sound information more accurately but at a higher cost of complexity and with specific requirements on the hardware configurations used in the signal capture and playback.

1.1.2 B-format audio

In the 1970's, a spatial audio technique was developed under the name of Ambisonics. The technique allowed for full spherical audio reproduction, which includes sound arriving from above and below. By individually decoding the speaker channels from an underlying transport format signal, an Ambisonics recording can be replayed using multiple different speaker configurations. Of particular interest is the Ambisonics B-format, one of the more commonly used transport formats, which encodes the whole sound field in as few as four channels. A large number of operations and transformation on the signals is also possible, making it possible to easily manipulate the soundfield to enhance or alter the listening experience. Ambisonics never quite caught traction however, and fell into obscurity in favor for the more common two channel stereo and 5.1 surround sound formats. Recently, much due to the flexibility in reproduction independent of loudspeaker configuration, there has been a renewed interest in the format, especially in areas related to virtual reality and 360 degree video.

1.2 Purpose and goals

The purpose of this thesis is to investigate the spatial properties of the B-format signal and try to develop an interpretation of the encoded spatial information, such as sound directionality, and how this relate to the sound sources captured in the signal.

- Can we quantify the spatial information in a useful way?
- As more and more, possibly infinite sound sources are added to the soundfield, how will the information lost affect the accuracy of the quantification?

We then wish to explore how we can use this interpretation of the spatial data in practical sense:

- Can we visualize the spatial data of the signal in a meaningful way?
- Can we use the spatial data for filtering purposes?

The end goal of this thesis should be to provide an analytical model of the spatial properties of a B-format signal and possibly some set of applications based on this analytical model.

1.3 Scope and limitations

This thesis will be limited to first-order B-format signals, partly for complexity reasons, but also due to additional costs involved in sourcing physical recording equipment and manpower for testing and verification purposes. For that reason, in the scope of this thesis, any references to the B-format in this thesis, unless explicitly specified, refers to the first-order B-format. Many other works and papers have been published on the subject of calibration and accuracy in B-format recording and reproduction. Thus, the amount of work focused on such activities should be minimized. The results should be interpreted in the light of such a context rather as a proof of concept and it is expected that work expanding on this thesis with such a focus should be able to achieve more accurate results.

Chapter 2
Theory

2.1 Basic audio theory and terminology

In the physical world, audio is the same as variation in pressure differences over time. In the common case of sounds reaching our ears, an air pressure difference exerts a pressure on tiny hairs in the ear canal. This pressure is transduced into electrical nerve signals which are then processed by the brain which interprets the change of pressure as sounds; fast changes as high pitched sounds and slow changes as low pitched sounds. Whenever an object produces a sound, we call this object a *sound source*. The pressure difference produced propagates outwards from the source in the shape of a *pressure wave*. As this wave reaches an entity capable of sensing the pressure wave we call this entity a *listener*. The pressure wave sensed by the listener may be described by a function over time $s(t)$ which we call the *signal* received by the listener from the source.

2.1.1 Spatial audio

As both sound sources and listeners exists in a space, they exhibit some set of spatial properties: such as a position and size. From the point of view of a listener, it means that sound sources exits in some directions at some distance from the listener. By measuring the change of pressure along the spatial directions, the source may estimate along what direction the pressure wave is propagating, and may therefore also estimate the direction of the audio source emitting the signal. We call this the *direction of arrival* or DOA of the signal. The idea of a sound source from the perspective of a listener is thus limited to the received signal and the directional *azimuth* and *elevation* angles relative to the listener. The abstract spherical space of signals from sound sources around a listener is often called the *soundfield* or, because of the meaninglessness of distances, the *far-field* around the listener.

2.1.2 Interference

As the pressure level at a point in space is affected by multiple pressure waves the resulting pressure level at that point becomes the sum of the effect of each individual waves. This is also called *interference*. A side effect of this is that a listener sensing the pressure level at one point can thus only sense one value for

all signals combined. The *mixing* of the signals causes a loss of information as less information about the individual pressure waves can be determined.

2.2 B-format audio

The B-format is a spatial audio format closely resembling the soundfield model described in 2.1 and describes the full spherical soundfield around one point in space. It is defined such that a signal s arriving from the DOA with azimuth θ and elevation ϕ is encoded as[1]:

$$\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ \cos\theta\cos\phi \\ \sin\theta\cos\phi \\ \sin\phi \end{bmatrix} s \tag{2.1}$$

The four channels w, x, y, z describes the complete spherical soundfield around the origin point at the listener. The channel w describes the absolute sound pressure at the origin point, whereas x, y and z describes the pressure gradient for the respective cardinal axes. The w channel is also said to be the *omnidirectional* channel as sounds arriving from all directions are attenuated equally. Similarly the, x, y and z channels are said to be the *directional* channels, favoring sources in their respective axis in a figure-of-eight pattern while attenuating all other directions. Also as sounds are positioned towards the negative side of the axis, the phases of the signals of said sources are encoded inverted.

2.2.1 Soundfield microphones

There are multiple ways of creating B-format signals. Monophonic signals may be encoded directly using (2.1) or a physical scene can be captured using special microphone arrays. A native B-format microphone array can be constructed by putting microphones with pickup patterns corresponding to the four channels close to the same point in space. That means one *omnidirectional* microphone and three *figure-of-eight* microphones at right angles. Another way is to construct a so called *Soundfield microphone* consisting of four *cardioid* microphone capsules facing in the directions perpendicular to the faces of an tetrahedron . Such an array outputs *A-format audio* which can readily be transformed to B-format using the following matrix:

$$\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} FLU \\ FRD \\ BLD \\ BRU \end{bmatrix} \tag{2.2}$$

where the A-format channels are $FrontLeftUp$, $FrontRightDown$, $BackLeftDown$ and $BackRightUp$ respectively. Additionally, in order to compensate for the fact that the microphone capsules cannot be positioned at the exact same position in space, the following equalizing filters should be applied to the converted complex

frequency space signal[2]:

$$F_w = \frac{1 + i\omega r/c - \frac{1}{3}(\omega r/c)^2}{1 + \frac{1}{3}i\omega r/c} \tag{2.3}$$

$$F_{xyz} = \sqrt{6}\frac{1 + \frac{1}{3}i\omega r/c - \frac{1}{3}(\omega r/c)^2}{1 + \frac{1}{3}i\omega r/c} \tag{2.4}$$

where i is the imaginary number, ω the angular frequency and r the distance of the capsules to the center of the microphone array. Since the capsules of the native B-format array and Soundfield microphones ideally should be positioned as close together as possible, these assemblies can be made very small. This can be advantageous in certain applications compared to spatial audio techniques where larger microphone arrays are necessary, such as Time-Difference arrival estimation.

2.2.2 Soundfield filtering and spatial transformations

One advantage of B-format audio is that there are many types operations which are fairly trivial to perform on the soundfield. Many common types of common wave shaping filters, such as low-pass, high-pass and notch filters can be applied given that all channels are processed equally. Another common operation, adding to the flexibility of the format, is the *virtual microphone*:

$$M(\theta, \phi, p) = p\sqrt{2}w + (1 - p)(\cos\theta\cos\phi x + \sin\theta\cos\phi y + \sin\phi z) \tag{2.5}$$

which projects the soundfield down to one channel with a signal signal matching that of a microphone pointing in the direction (θ, ϕ) with a polar pick-up pattern defined by $p$going from omnidirectional to cardioid to figure-of-eight. If we view the three directional channels x, y, z, as the basis vectors of a three dimensional space, we can also apply general affine transformations on the soundfield space. Three new channels $\hat{x}, \hat{y}, \hat{z}$ are derived by putting three virtual microphones along the base vectors of the transformation matrix. This effectively gives us a number of parameterized operations such as [4]:

- Reflection through the plane $ax + by + cz = 0$:

$$\begin{bmatrix} \hat{w} \\ \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 - 2a^2 & -2ab & -2ac \\ 0 & -2ab & 1 - 2b^2 & -2bc \\ 0 & -2ac & -2bc & 1 - 2c^2 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} \tag{2.6}$$

- Rotation with parameters $\phi, \theta, \psi$for rotation around the x-, y- and z-axis respectively:

$$\begin{bmatrix} \hat{w} \\ \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = \mathbf{R} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} \tag{2.7}$$

where the rotation matrix

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\phi & -\sin\phi \\ 0 & 0 & \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & 0 & \sin\theta \\ 0 & 0 & 1 & 0 \\ 0 & -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.8}$$

- Scaling along the x-, y- and z-axises with factors respectively:

$$\begin{bmatrix} \hat{w} \\ \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & s_x & 0 & 0 \\ 0 & 0 & s_y & 0 \\ 0 & 0 & 0 & s_z \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} \tag{2.9}$$

2.3 Soundfield analysis

One thing that we would like to do is to expand this this toolset of B-format operations. While some information is lost in the mixing process that happens when the soundfield captured, it is expected that some of the spatial information about the sound sources in the soundfield is contained within in the signal could be used for some other purposes.

2.3.1 The audio source model

In order to perform any meaningful spatial analysis of B-format soundfield we need a model of the audio sources of which the soundfield represents an image of. Given the definition of the B-format soundfield (2.1) we can conclude that sound sources have at least the properties of a position and sound signal. We also note that the encoding function is simply the mapping of the spherical coordinates θ, ϕ to the Cartesian coordinates of the orthonormal space spanned by the directional channels. A naïve estimate of the audio source properties case of a single audio source would then simply be the reverse mapping:

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = \begin{bmatrix} \arctan_2(y * w, x * w) \\ \arctan_2(z * w, \sqrt{x^2 + y^2}) \end{bmatrix} \tag{2.10}$$

Note the use of the $arctan_2$ function, which is the multi-valued inverse tangent function. It takes the sign of the operands in order to return the angular value in the correct quadrant. The sign in this case corresponds to whether the channel is in or out of phase with the signal received from the audio source. Since the phase of the omnidirectional channel is independent on DOA of the source we can compare the phase of the directional channels to that of w and get the sign for the channel, thus the multiplication with w.

2.3.2 Multiple audio sources and interference

The naïve case of the soundfield analysis (2.10) works well for soundfields of only one audio source. Since there is no information loss due to interference, the complete information for accurate estimation of the source's properties is contained

within the signal. However, as soon as the number of audio sources within the soundfield increases we start to run into issues due to interference and the analysis becomes inaccurate. We know from Fourier analysis that

$$S(t) = \sum_k (a_k \cos(f_k t + \omega_k)) \tag{2.11}$$

and the harmonic addition theorem states that

$$\sum_k (a_k \cos(ft + \omega_k)) = a \cos(ft + \omega) \tag{2.12}$$

or in other words: Any continuous signal can be represented as a sum of a set of sinusoids; The addition of sinusoids of one frequency results a sinusoid of the same frequency. Or: spectral content in one signal will only interfere with spectral content of the same frequency in other signals. Thus, we can safely reason about interference per frequency, or in the frequency domain which simplifies the matter somewhat. In practical terms, we do that by applying one of the versions of the Fourier transform to the signal:

$$s(t) = (w, x, y, z)(t) \overset{\mathfrak{F}}{\to} S(f) = (W, X, Y, Z)(f) \tag{2.13}$$

For the sake of convenience we define $\bar{W} = \sqrt{2}W$. Following this, it would be advantageous to state some assumptions about sound sources in a meaningful soundfield:

Assumption 1. *Signals from audio sources do not use the full bandwidth of the captured B-format signal.*

Assumption 2. *Signals from audio sources are spectrally unique and dominant.*

The first assumption is motivated by the fact that a full bandwidth signal is a continuous signal of white noise, analyzing the spatial properties of such a signal would be meaningless since the estimated values would be completely random. The second assumption is a bit trickier to motivate. Consider two sources emitting sound with the exactly same spectral content. Either they are part of the same entity (like singers in a choir) or through chance they happen to be the same. In any case, the limits of the first order B-format prevents any meaningful distinction between sub-entities. So by looking at a single frequency of the signal, we should now be able to determine the spatial properties of the audio source which radiates that frequency using a variation of (2.10):

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = \begin{bmatrix} \arctan_2(Y \cdot \bar{W}, X \cdot \bar{W}) \\ \arctan_2(Z \cdot \bar{W}, \sqrt{X^2 + Y^2}) \end{bmatrix} \tag{2.14}$$

In particular, note use of the dot product of the complex values, which means we scale the directional channels based on the phase alignment with $\bar{W}$, effectively giving the signed axis aligned amplitude. It is expected that in practice, most soundfields do not strictly follow the assumptions stated above. Noise is present in most audio recording systems and most audio source spectra overlaps within the

limited spectral resolution of any signal processing system performing the soundfield analysis. Thus, a best effort soundfield analysis is expected to be inaccurate in proportion to the presence of white noise and spectral overlap among audio sources. It would be practical to be able to calculate the proportion of energy at a given frequency is coming from the dominant audio source compared to other sources and noise. If we consider how the energy content at one frequency of the directional channels is affected as multiple audio sources are contributing:

$$\begin{aligned} D = \sqrt{X^2 + Y^2 + Z^2} = \\ \sqrt{(\sum_{i \in N} \cos\theta_i \cos\phi_i s_i)^2 + (\sum_{i \in N} \sin\theta_i \cos\phi_i s_i)^2 + (\sum_{i \in N} \sin\phi_i s_i)^2} = \\ \sqrt{\sum_{i,j \in N} s_i s_j (\cos\theta_i \cos\phi_i \cos\theta_j \cos\phi_j + \sin\theta_i \cos\phi_i \sin\theta_j \cos\phi_j + \sin\phi_i \sin\phi_j)} = \\ \sqrt{\sum_{i,j \in N} s_i s_j (\cos(\theta_i - \theta_j) \cos\phi_i \cos\phi_j + \sin\phi_i \sin\phi_j)} = \\ \sqrt{\sum_{i,j \in N} s_i s_j \cos\sigma_{ij}} \end{aligned} \tag{2.15}$$

as compared to how the total energy content at one frequency of the omnidirectional channel is affected:

$$\bar{W} = \sum_n s_n = \sqrt{(\sum_{i \in N} s_i)^2} = \sqrt{\sum_{i,j \in N} s_i s_j} \tag{2.16}$$

Take notice of how the signals $\bar{W}$ and D differ only by proportion of the angle $sigma_{ij}$, which is the central angle between the audio sources i and j as per the the spherical law of cosines.

Remark. *Since* $\cos(\sigma_{ij}) \leq 1$ *then* $W = D$ *if and only if* $\cos(\sigma_{ij}) = 1$ *for all* $i, j, s_i \neq 0$.

This means that as $\bar{W} = D$, all audible sound sources are positioned along the same DOA and as $\bar{W}$ and D become more distant, the more the sound sources are spreading out spatially. More specifically, $\bar{W}$ and D becomes more distant as sound energy is arriving from a larger spherical area. So, while we did not get a measurement on how much interference is affecting the signal per se, however, we have a measurement of how spread out the sound energy at a given frequency arriving at the listener is. We define the term *Relative spatial spread* (RSS) as:

Definition 1.

$$RSS = \left| \frac{\bar{W} - D}{\bar{W} + D} \right|$$

Since the value of the spatial spread is dependent on the magnitude of $\bar{W}$ and D we divide by said magnitudes to get a relative spatial spread value between 0, where there is no spread, all sound energy is arriving from the same direction, and 1, where the energy is arriving evenly from all directions. For the sake of convenience, we might want a general function mapping the the angular radius of a circle from which the equivalent amount of sound energy would have been evenly spread out to the RSS value. Such a function can be derived from (1) as:

$$RSS = \frac{\sqrt{2} - \sqrt{1 + \cos\frac{r}{2}}}{\sqrt{2} + \sqrt{1 + \cos\frac{r}{2}}} \tag{2.17}$$

Similarly, the inverted function mapping RSS to angular radius becomes:

$$r = 2\arccos(2 * \left(\frac{RSS - 1}{RSS + 1}\right)^2 - 1) \tag{2.18}$$

2.3.3 Audio source width

Now, so far we have assumed that sound sources are point like, that is, sounds from one sound source only arrives from one single exact DOA. In real life, this is rarely true. Things that makes sound have sizes: speaker cones have a diameter, guitar strings have a length and the resonance cavities that are human mouth openings tend to widen as they talk. One way to approximate sized audio sources is to simply use a large number of point like audio sources spread all over the sized audio source, each radiating a small fraction of the total audio energy emitted by the approximated sized source. Again, in assumption (2) we assumed that sound sources are dominant within their spectrum. Or at least, in a less strict sense, dominant enough in large enough portion of the spectrum that the soundfield analysis becomes meaningful. Since the RSS represents a size of the spatial distribution of many point like audio sources, it also means that we also can interpret the RSS as an indicator of the angular width of a sized audio source. So under assumption (2), the RSS should be proportional to the angular size, along equation (2.18), of the dominant sound source of that frequency.

2.4 Practical applications

Using the soundfield analysis theory presented in 2.3 it is then possible to transform a B-format signal into a domain of single frequency audio sources with a set of spatial properties.

$$s(t) = \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} \leftrightarrow S(f) = \begin{bmatrix} s \\ \theta \\ \phi \\ RSS \end{bmatrix} \tag{2.19}$$

Similarly it is possible to perform the reverse transformation using (2.1) and (2.17). The purpose of this section is to propose two types of practical applications based on the spatial analysis transformation into the *audio source domain*.

2.4.1 Soundfield visualization

One way we can use the results soundfield analysis is as parameters for an algorithm which draws a visualization of the B-format soundfield. Equation (2.18) lets us translate the RSS for a frequency into the radius of a circle on the spherical soundfield from where the sound energy at that frequency is arriving. As such, for each frequency we can draw a circle on a sphere with the center at the coordinates (θ, ϕ) and radius given by equation (2.18). We can fill the circle with a color mapped from the frequency to a suitable color gradient and map the amplitude of the signal at the frequency to the opacity or intensity of the circle. We can then project the sphere onto a rectangle using any common map projection and display the result on a screen, giving a sort of spheric map or camera image of the whole soundfield. If the B-format signal is split into windowed segments over time, we can use the short time Fourier transform to perform the soundfield analysis and visualization continuously over time, possibly several times per second in real-time, creating a sort of sonic camera.

2.4.2 Spatial filters

Another we can use the spatial domain analysis is to design filters which act in the audio source domain using the spatial properties of audio source as filtering parameters. One such filter type is a *direction-of-arrival filter* or DOA-filter:

$$\hat{s}(f) = s(f) * g(\theta(f), \phi(f)), g \to [0, 1] \tag{2.20}$$

which attenuates signals arriving from certain directions while preserving others. One use case for such a filter could be source separation, where the signals arriving from two spatially spread out audio sources are separated into the signals transmitted from each source respectively. If the direction to the sources of interest is known, their signals can be separated by attenuating audio arriving from all other directions. Another type of filter we can implement in the audio source domain is *width filters*:

$$\hat{s}(f) = s(f) * g(RSS), g : [0, 1] \to [0, 1] \tag{2.21}$$

which can be further classified into narrow-pass-filters and wide-pass-filters which attenuates signals arriving from wide and narrow audio sources respectively. One use case for such filters is background noise removal. Assuming that background noises, such as distant talking, air condition systems, electric grid hum et cetera, tend to appear wider as it propagates through walls and reverberates. Then such noises could be reduced by a narrow pass filter.

2.4.3 Practical considerations

Looking back at the assumptions made in 2.3.2, we reasoned that the strictness assumption could be loosened with the prediction of a lesser accuracy in the analysis. If we consider for instance equation (2.14), we can see how the influence of the interference of non-dominant signals from secondary audio sources may shift the result from the DOA-estimation towards the direction of a weighted mean of

direction of arrival of all influencing sources. Similarly, looking at equation (2.15) and the RSS definition (1), we see how the interference of non-dominant signals from secondary audio sources may influence the RSS to appear larger than that of the dominant audio source. In the case of complete non-dominance, that is, no source is particularly dominant on a frequency, the result of the analysis is expected to be useless or erroneous for that frequency. Again however, assuming that sources are mostly dominant in their spectra, a best effort analysis should give mostly accurate results. A filtered signal containing artifacts such as musical noise and leakage might still be a clear improvement over the unfiltered signal. Since random white noise from recording equipment and similar also exhibit random spatial properties, it is expected that such noise will degrade the accuracy of the analysis for all sources and frequencies in proportion to the signal to noise ratio.

Chapter 3
Method

We wish to investigate partly whether the theory and spatial analysis method laid out in chapter 2 is feasible even under real-world conditions. We also which to see how well the filter types presented in 2.4.2 perform in the use cases used as examples. In 2.4.3 we predicted that interference caused by the spectral overlapping in real-world sources and equipment noise would cause a degradation in analysis accuracy. As such we also wish to investigate to what degree these factors affect the performance in practical real-world applications of the soundfield analysis.

3.1 Setup

The theory laid out in chapter 2 was evaluated using a real world setup. To prevent the influence of outside factors all experiments took place within a well isolated anechoic chamber. For the sake of reproducibility all generated sounds were synthesized and sent to two speakers positioned inside the anechoic chamber. A soundfield microphone, as described in section 2.2.1 was built as per [3] and put in chamber at a point between the two speakers. The angle between the two speakers as seen from the microphone array was about 90 degrees horizontally. Additionally a 4 channel preamp was built to power the microphone array and amplify the low-level signal. A USB audio interface with 4 input and 4 output channels were used to connect the microphone array and loudspeakers to a computer positioned outside of the anechoic chamber.

3.1.1 Software

The main software used in the evaluation was written specifically for the purpose of this thesis. It is written in C and is designed to run on most Linux systems. GNU Octave was also used for offline processing of results for this thesis. However any signal processing as described in section 2 is carried out in real time in the main software. Primarily the features of the software is divided among two threads.

The main thread is responsible for the audio signal processing. It can be run in several modes, depending on desired input and type of processing to apply. The input audio signal may be captured from the microphone connected to the USB-interface using ALSA or it may be read from a B-format file stored on disk. It may also playback simulated audio sources using audio files stored on disk on

the connected loudspeakers. The input audio signal is continuously processed in windows of a configurable sample length (by default 1024 samples) according to the overlap-add-method. That is, first the last two windows are split into three overlapping windows. A hamming window function is applied to all three windows before applying the fast fourier transform. Processing in the frequency domain is then performed on all windows berfore being transformed back to the time domain. The three windows are then added together forming the resynthesized processed time-domain signal on the center window. The frequency domain processing of each window is configureable depending on the current mode. If the microphone is used as input, the filters described in 2.2.1 are applied first, white noise may also be applied to the microphone signal at this point, for the purpose of simulating equipment noise. After processing the microphone input the spatial analysis described in 2.3.2 is performed. The results of the analysis is then used when applying the filters as described in 2.4.2 if desired. The post-filter result may be written to a file on disk or sent to monitoring, depending on the selected mode. After the frequency domain processing the signal is yet again transformed into the frequency domain for the monitoring step, albeit this time only for one window. The spatial analysis is carried out again, followed by the a down mixing of the time-domain filtered signal for output to a pair of headphones by the computer.

The other thread is responsible for graphical monitoring of the processed audio signal, giving visual feedback to the operator. It implements the visualization algorithm described in 2.4.1. The graphics thread is fed the result most recently analyzed audio window from the monitoring step through a triple buffer. The triple buffer prevents the any data races from occurring due to both threads accessing the same buffers simultaneously. The graphics thread is designed in such a way that it is not a dependency of the audio thread. As such, if desired, it may be left out from being compiled into the resulting binary at all. This is desirable if the software is to run on an embedded system or small DSP-chip, considering that the graphics requires GPU processing.

3.1.2 Filter evaluation

The filter types described in 2.4.2 are evaluated in three different practical scenarios:

- Separation of background noises from foreground speaking sound source
- Separation of two simultaneously speaking sound sources
- Separation of two simultaneously speaking sound sources in noisy environment

The first scenario evaluates the effectiveness of using a width filter to separate the mixed audio of a narrow talking sound source from wide background noise. The second scenario the effectiveness of using two DOA-filters configured to separate the mixed audio of two narrow sound sources of human speech while the third scenario is a combination of the two previous scenarios, separating the three mixed audio streams using a combination of both types of filter.

For the purpose of reproducibility, three base audio recordings are used for all scenarios, the spectrograms in Figure 3.1 gives an idea of the spectral content of each base recording:

1. talk1.wav - a man reading a passage from a book in finnish
2. talk2.wav - another man reading a passage from a book in english
3. back.wav - featuring various background noises from a public space containing mumbles and some brass instruments tuning

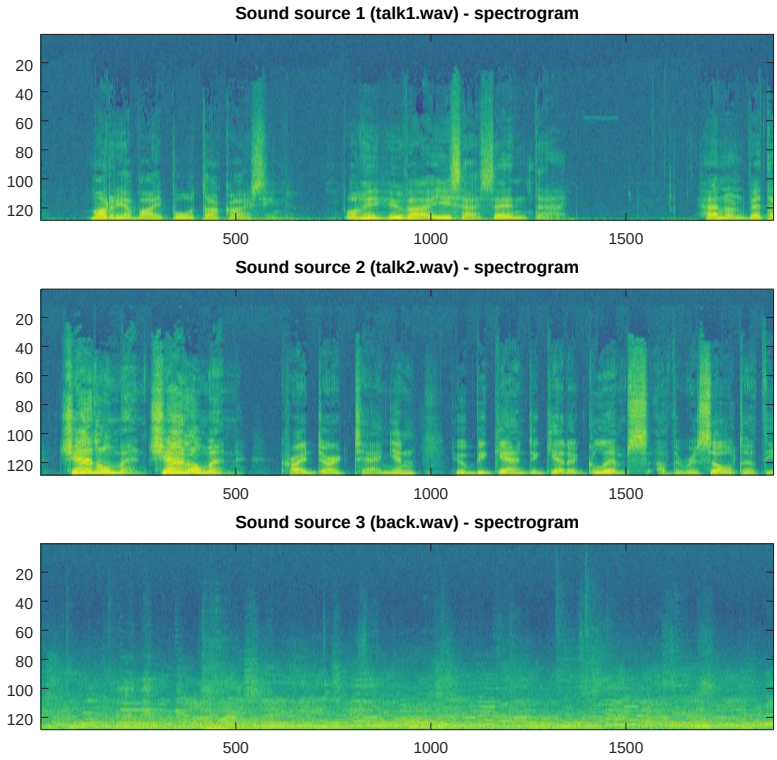

Figure 3.1: Spectrograms of the spectral content over time of the base audio recordings used in the evaluations

The base audio recordings are mono-channel wave files. Simulation of the different spatial properties of the sound sources are done by replaying the recordings at different volumes in the two speakers. By playing a recording in one speaker only, a narrow sound source at the position of the speaker is achieved while playing a recording in both speakers simultaneously achieves the effect of a wider audio source positioned between the speakers. Sound sources are simulated by replaying the base audio recording on the loudspeakers in a way that achieves the required spatial properties. Meanwhile, the soundfield microphone positioned between the loudspeakers captures the audio in the chamber. Any processing is applied to the

captured and the result is stored in a new 4-channel audio file for off-line processing of results.

Evaluation procedure

Each scenario is carried out in the following manner: First, each sound source is played back, one by one, in turn and the unprocessed captured signal is stored in a reference recording. This is done twice for each source, resulting in two reference recordings of each source. These recordings are called *ref_ i_ 1* and *ref_ i_ 2*, where i is the source index. Then, all sounds are played back at the same time, mixed. The software stores the captured unfiltered recording, called *mix*, before applying the corresponding filters implemented for each source. The result of the filtering is also stored in a corresponding recording for each audio source, called *res_ i*. The filter parameters are acquired by performing the spatial analysis on *ref_ i_ 1*, taking the weighted mean of the spatial parameters. See figure for an principal illustration of the filter curve derived from the spatial parameters. A recording of the chamber without any source being played back is also done for measuring noise levels, called *noise*.

GNU Octave is then used for comparing the recorded files using cross correlation ($\star$) [5]. Cross correlation takes two signals and return a similarity value between 0 and 1, where 0 means there is no similarity and 1 means the signals are identical. The following cross corellations are performed:

- $c_s = ref\_i\_1 \star ref\_i\_2$
- $c_m = ref\_i\_1 \star mix$
- $c_r = ref\_i\_1 \star res\_i$

c_s is the likeness of the two reference recordings. That is after any system noise have affected the signals. This is the best possible reconstruction of the reference signal we can expect. c_m is the likeness of the unfiltered mix recording and the reference recording. If the result of the filtering is lower than this, the filtering is actually making a worse job in reconstructing the reference signal than no filter at all. c_r is the likeness of the filtered source signal to the reference signal. It is a measurement of the filter efficiency. Ideally this should be somewhere in the range of $c_m < c_r < c_s$, with the closer c_r is to c_s, the better the filter is reconstructing the reference signal.

The scenarios are repeated several times with additional white noise added to the signal captured by the microphone. The filter parameters are always derived from a run of the scenario with the least possible noise level (no added white noise). The actual signal to noise ratio in the system at each iteration of the evaluation is calculated by comparing the maximum audio level of the *noise* and *res_ i_ 1* recordings. This allows us to see the performance of the filters as a function of the signal to noise ratio in the system.

3.1.3 Soundfield visualization

Designing formal experiments testing the effectiveness or usefulness of a data visualization technique is a difficult task and out of scope of this thesis. Also, the

audiovisual coherence of a real-time audio visualization cannot quite be conveyed through printed medium alone. However, some example screenshots of the visual output of the software described in 3.1.1 will be provided in chapter 4. The subjective experience of the audiovisual feedback provided by said software will also be reflected on and discussed in chapter 5

Chapter 4
Results

The results from the evaluation performed in section 3.1.2 are presented in this section. The results are presented as a graph of the three cross correlations c_s, c_m and c_r over the signal to noise ratio in the system. The plots can seen as the best possible reconstruction, least improving reconstruction and actual resulting reconstruction as functions of signal to noise ratio. Each graph is also displayed as a relative improvement of the signal by the filters given by the formula $y = \frac{c_r - c_m}{|c_s - c_m|}$. This transformation can be interpreted as if stretching the space between the top and bottom plot in non-relative graph to fill the relative graph space, showing the relative improvement of the signal reconstruction as a normalized value between 0 and 1.

4.1 Scenario 1

This scenario evaluated the performance of a pair of width-filters as described in 2.4.2 tasked with separating and reconstructing the signal arriving from one talking narrow audio source and one wide background noise audio source under varying signal to noise ratios in the system.

4.2 Scenario 2

This scenario evaluated the performance of a pair of DOA-filters as described in 2.4.2 tasked with separating and reconstructing the signals arriving from two simultaneous audio sources under varying signal to noise ratios in the system.

4.3 Scenario 3

This scenario evaluated the performance of a combination of DOA-filters and width-filters as described in 2.4.2. In this scenario, three sources off different widths and position were to be separated under varying signal to noise ratios in the system.

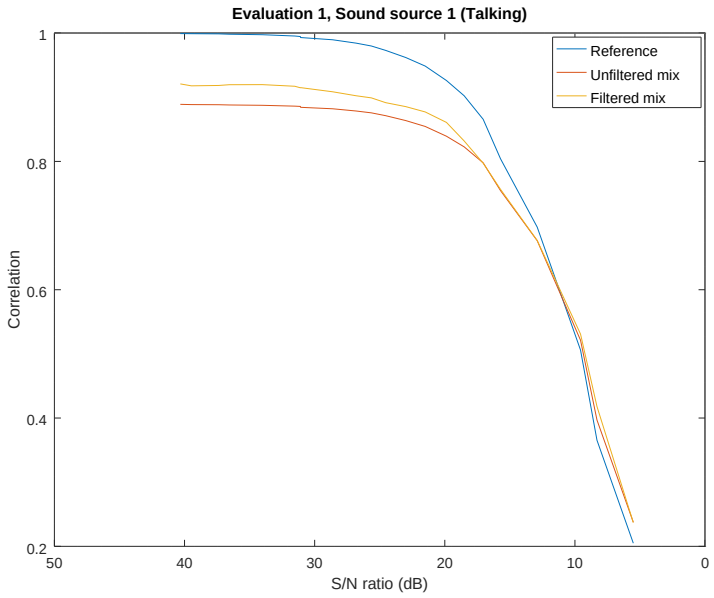

Figure 4.1: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

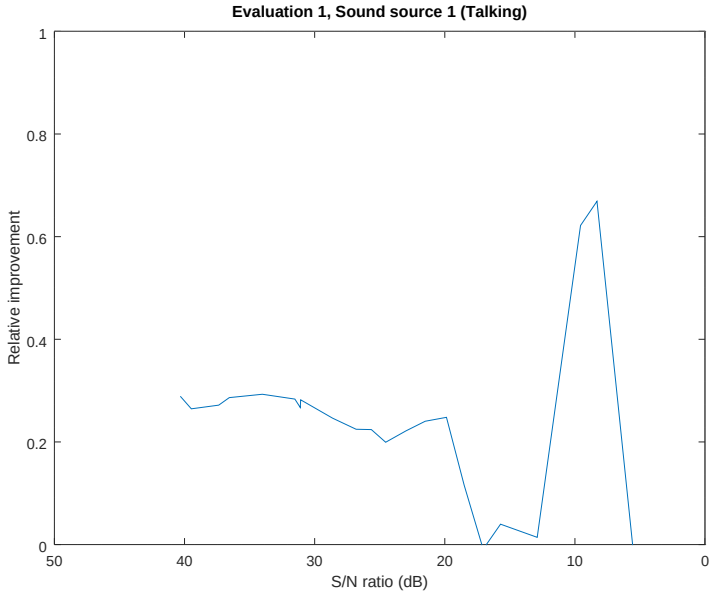

Figure 4.2: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

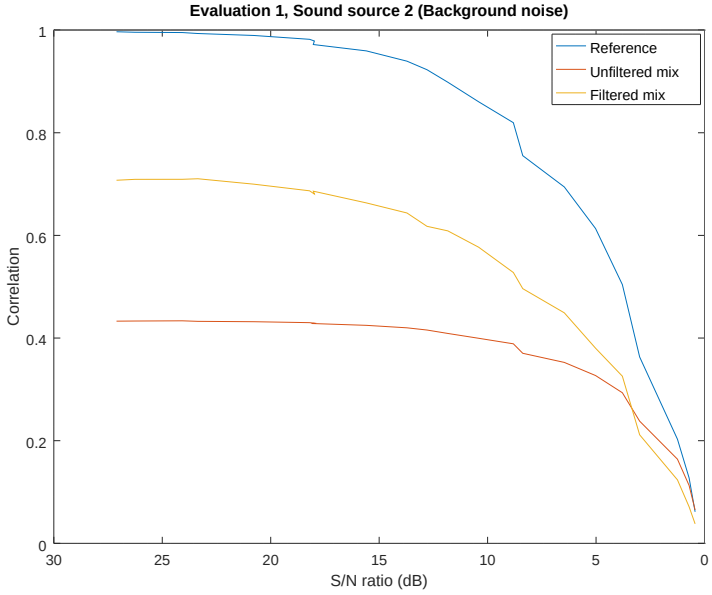

Figure 4.3: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

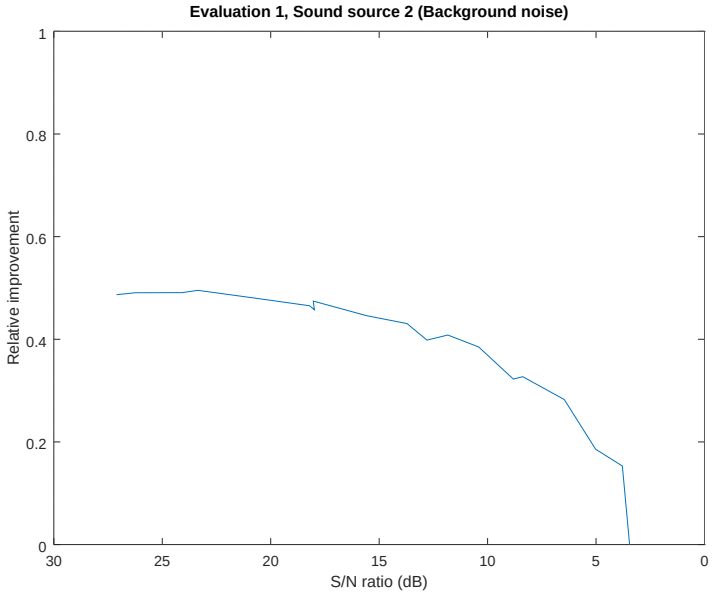

Figure 4.4: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

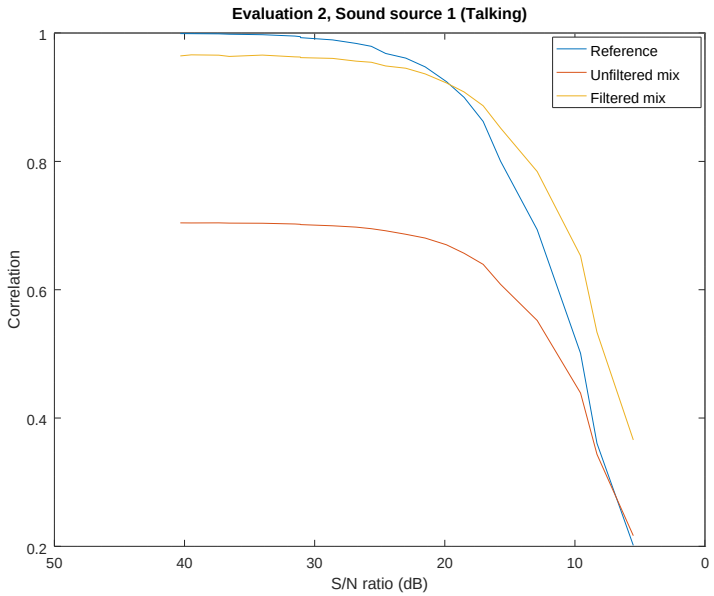

Figure 4.5: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

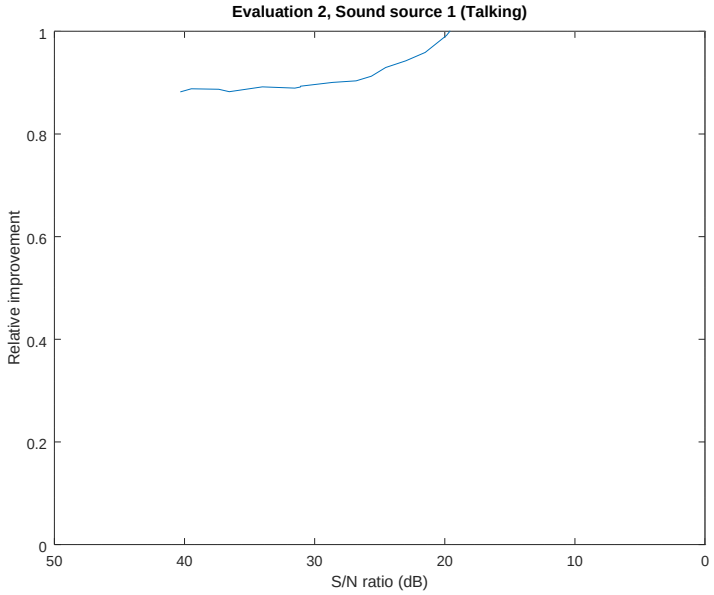

Figure 4.6: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

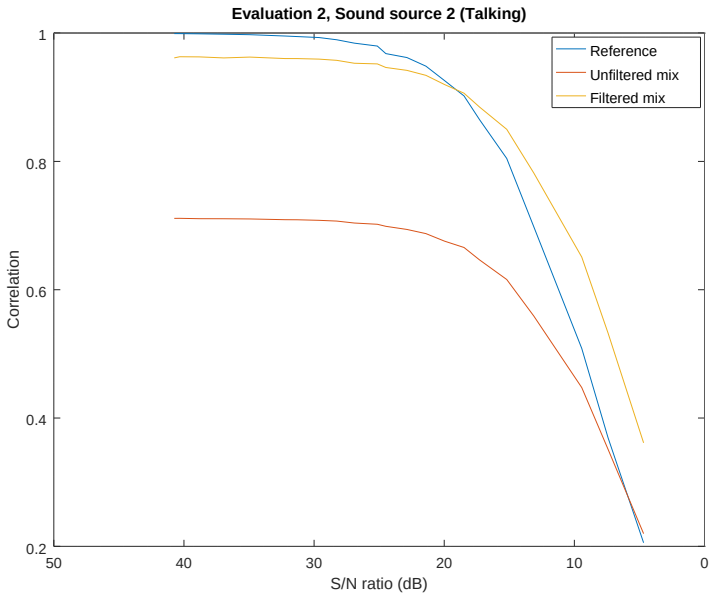

Figure 4.7: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

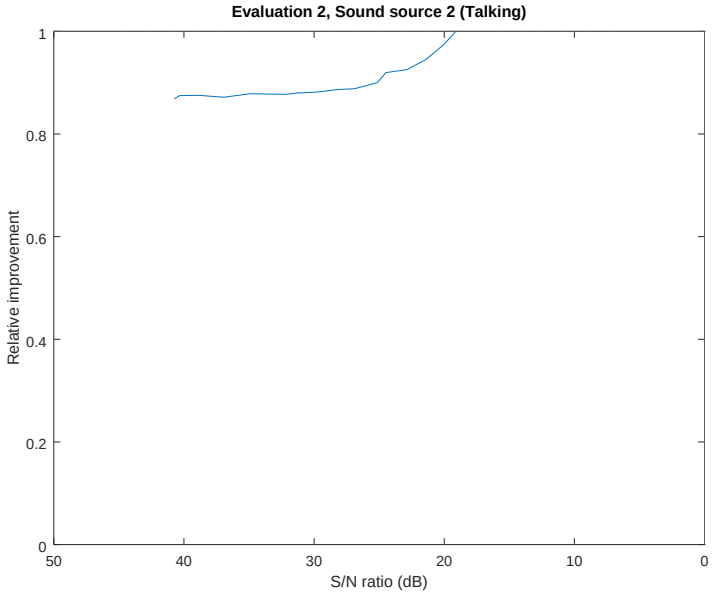

Figure 4.8: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

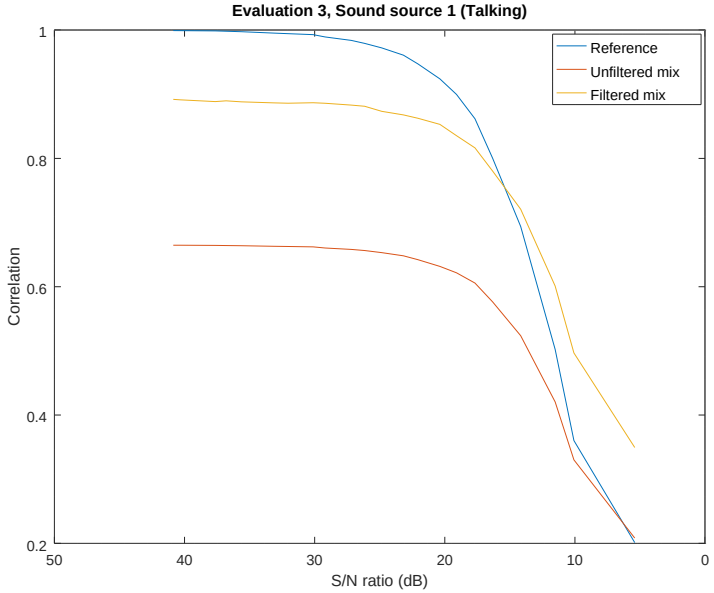

Figure 4.9: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

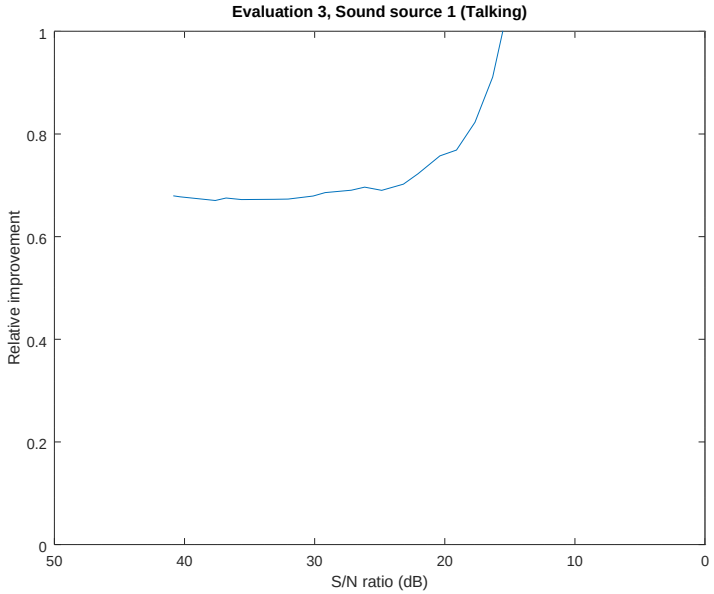

Figure 4.10: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

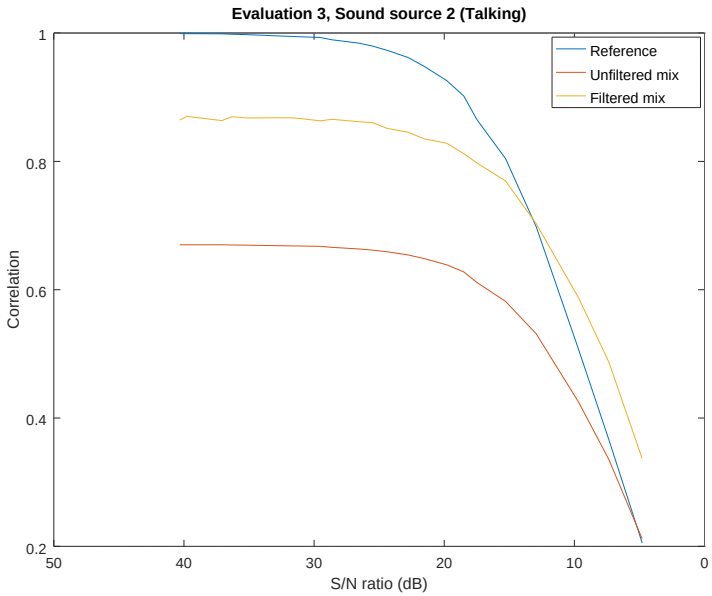

Figure 4.11: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

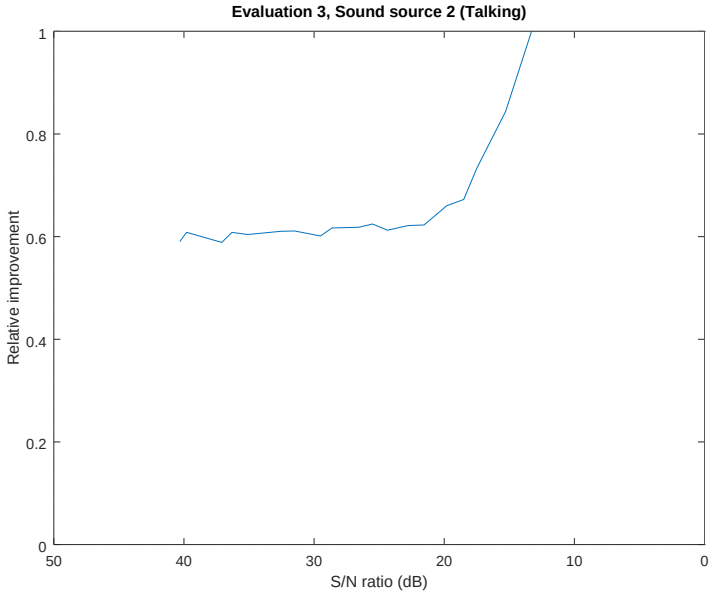

Figure 4.12: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

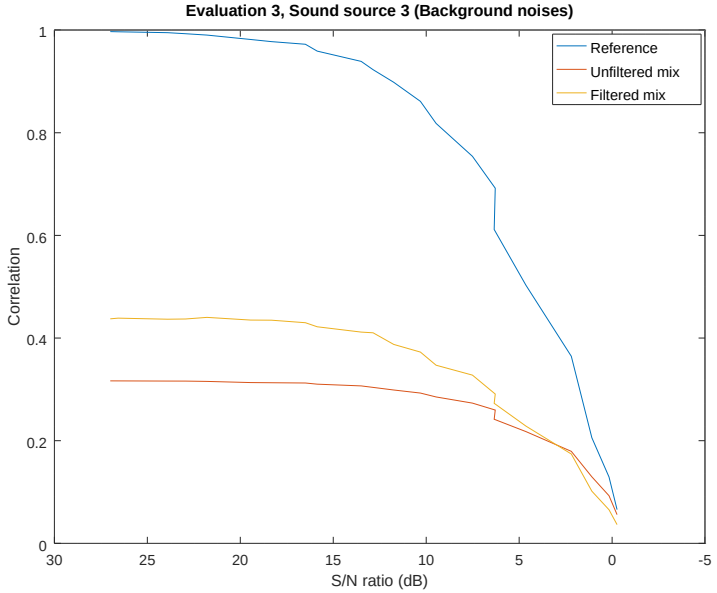

Figure 4.13: Correlation between signal reconstructed by filter and the reference. Reference correlation on the top, result correlation in the center and mix correlation at the bottom

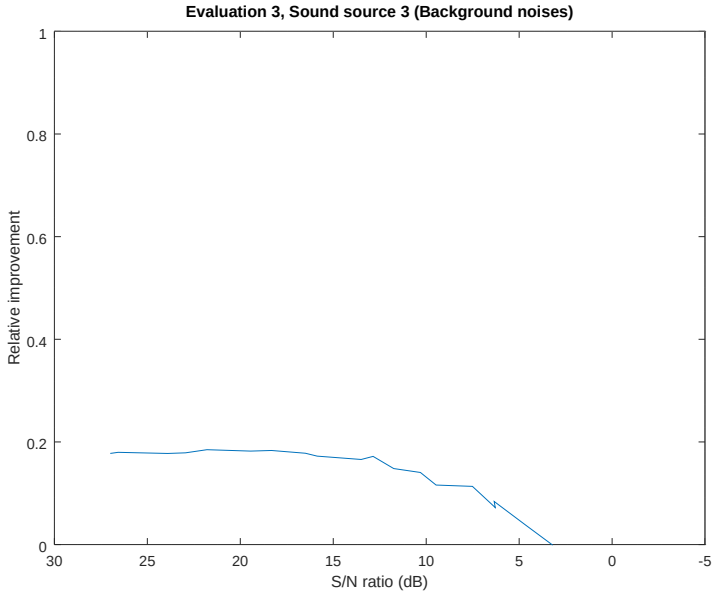

Figure 4.14: Relative improvement of resulting filtered reconstruction, as compared to unfiltered signal and best possible reconstruction

4.4 Visualization

In this section some screenshots of the visual output from the visualization algorithm is presented. The images depicts a visual representation of the soundfields of instants in various recordings. The captions of the images tries to describe the audible sounds heard in the recordings at the moments the images were taken. As described in section 2.4.1, the images show a projection of the whole spherical soundfield, similar to a cylindrical projection of a world map. The center of the image is the front direction, the thick vertical lines at a quarter from the edges are the directions 90 degrees to the left and right respectively and the top and bottom the directions of up and down. The direction towards the read wraps around on the left and right edges of the image. Circles of blue-ish color is sound energy in the low frequency spectrum whilst mid frequencies tend towards green and red at higher frequencies. The darker the image, the more energy is arriving from that direction.

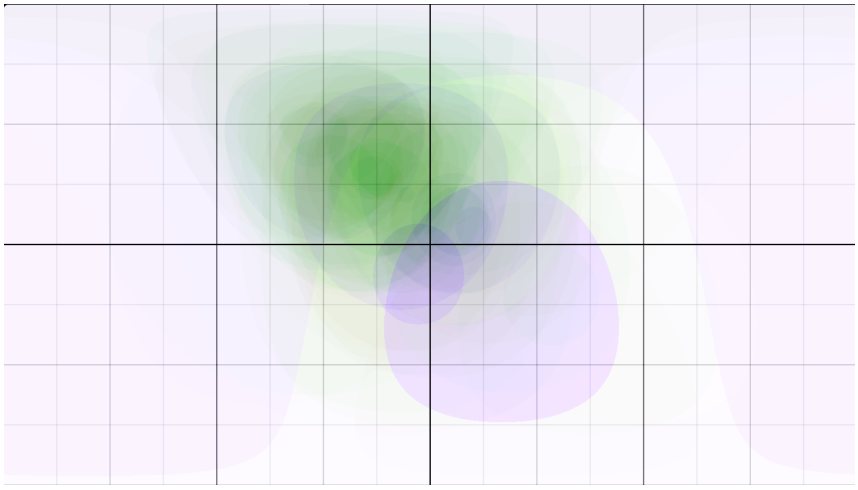

Figure 4.15: This screenshot was taken whilst replaying a B-format recording of a number of spitfire airplanes flying past the listener. In the instant depicted in the image, one of the airplanes is heard approaching the listener from the left towards the right.

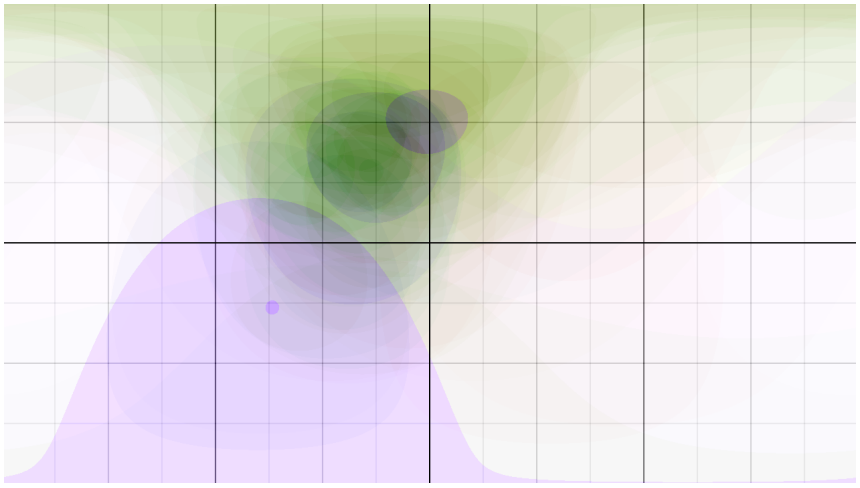

Figure 4.16: From the same recording as figure 4.15, here the airplanes has just passed the listener from the right towards the left.

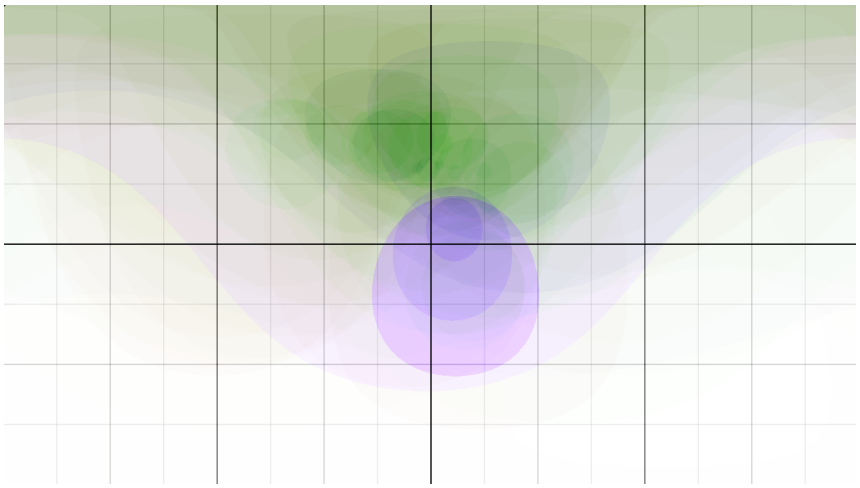

Figure 4.17: This screenshot is taken from a recording of a fireworks display, a short book from one of the firework pieces has just been heard from the front of the listener.

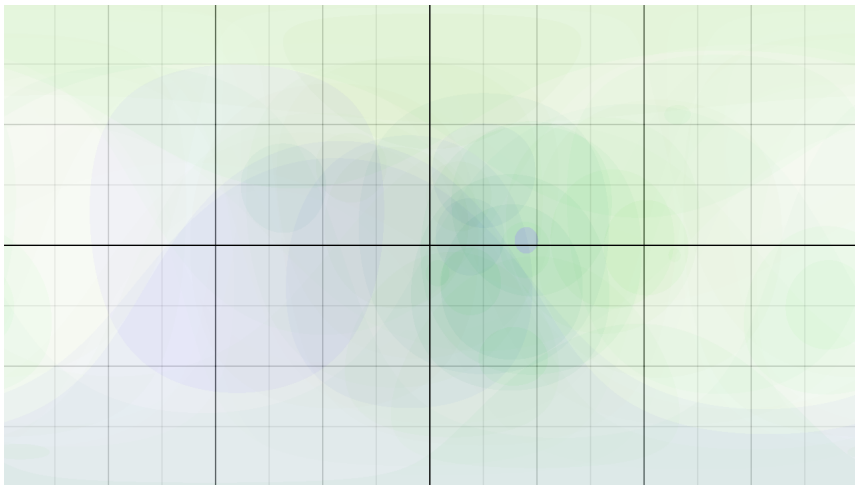

Figure 4.18: This screenshot is taken from a recording of a piano performance in a concert hall. The pianist is positioned to the front of the listener. A lot of reverberant sounds from the piano is heard all around.

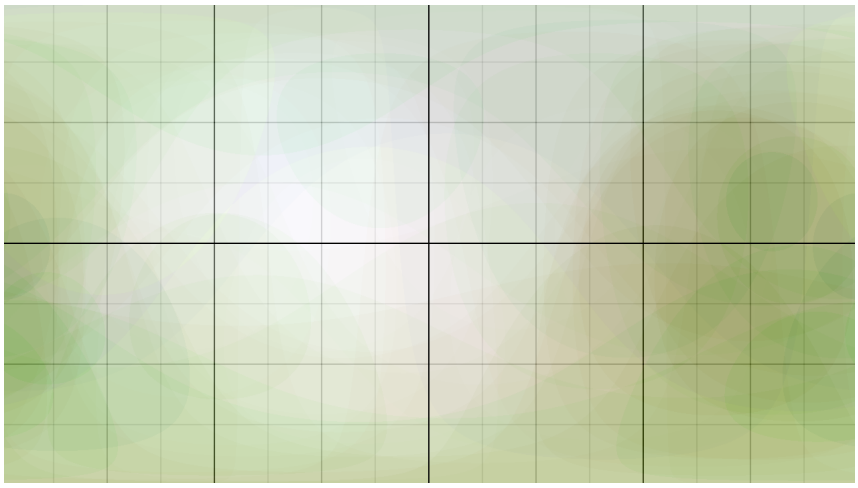

Figure 4.19: Later in the recording from figure 4.18, the piece is over and the audience is heard applauding loudly behind the listener.

Chapter 5
Discussion and conclusions

5.1 Spatial filter evaluation

One of the most clear conclusions that be made when looking at results from the filter evaluation in chapter 4 is that there is a clear improvement in the reconstruction of the reference signal by the filter for all sources in all three scenarios. Another overall conclusion that can be made is that all improvements are quite stable in regards to the signal to noise ratio. However, at a certain point around 15dB to 20dB S/N ratio, the noise in the system is to destructive to the signal, meaning that no real improvement can be made to the signal anymore and the soundfield analysis becomes pointless. However, note that a signal with 15dB S/N is a very noisy signal.

5.1.1 Scenario 1

One of the lowest improvements seen is that of the talking audio source from scenario 1 with a relative improvement of around 30% (figure 4.2). This can be explained with the fact that the unfiltered mix signal already is quite similar to the reference signal (figure 4.1), possibly because the sound file of the talking source is louder than that of the background noise that is filter out. As such it is hard to achieve a higher improvement than the achieved absolute correlation of around 95%. The wider source in this scenario, the background noise, on the other hand reach an improvement of around 50% (figure 4.4). The result is high but still rather limited, possibly due to the same fact that the source is quieter than the filtered source, meaning that the whenever the two sources interfere, the damage caused to the more quiet source becomes more significant.

5.1.2 Scenario 2

In this scenario we see some of the greatest improvement results of all scenarios, both sources could be reconstructed with about a 60% to 70% improvement (figure 4.6 and 4.8). One factor behind could be the fact that human voices have very wide but sparse spectra but also speech is very temporally sparse, meaning sound energy mainly comes in short burst, as can be seen in the spectrograms in figure3.1. This lowers the chance of interference and thus allows for better signal reconstruction.

5.1.3 Scenario 3

In the final scenario we see somewhat a mix of the results of the two other scenarios. Both the louder talking sources have improvements of about 60% to 70% (figure 4.10 and 4.12) while the more quiet background noise gets even more drowned out by the other sources (figure 4.14). Another explanation behind the lower improvement in the background noise may be the lack of DOA-filtering for this source. In the processing of the background noise source, only a width-filter was used to reconstruct the signal. A result of this is that any interference between the talking sources would cause those frequencies to appear wider, thus bleeding into the widepass-filter used for the background noise reconstruction.

5.2 Soundfield visualization

It is hard to objectively convey the audiovisual experience of using the soundfield visualization in conjunction to listening to the same soundfield in real time. As such, this section will contain some subjective observations and discussion on the subject on account of the author of this thesis and said visualization algorithm. The biggest impression of the visualization is that it responds fairly well to the auditory cues of the signal. Short burst of sound (as in the fireworks example of figure 4.17) pop by quickly, whilst longer sounds tend to focus energy in the region of the sound direction. In the examples with airplanes flying around the listener (figures 4.15 and 4.16), one could clearly see how the sounds of airplanes moved across the auditory field in tandem with the direction they were heard as moving in. Another interesting observation is that of how reverberations and echoes affect the resulting image. Short, immediate reverberations (close to the timeframe covered by a single analysis frame, 21ms) tended to diffuse the directionality of sounds, as can be seen in how 'big' the sounds in the piano example (figure 4.18 appear. Similarly, some reverberations could be observed affecting certain frequencies more than other, as an example, in the recording of airplanes, larger, blue circles were moving around below the main body mass of sound energy, suggesting that sounds of low frequencies both reached the listener directly as well as via reflections off the ground. Longer reverberations, as heard separately after short bursts of sound, as in the fireworks example (figure 4.17, could be clearly seen a few frames later arriving from other directions than the direct sound (not depicted in any figure). In some cases, someone listening to a B-format recording, may be limited to listening to a smaller portion of the full soundfield sphere, such as when using stereo headphones instead of a fully spherical loudspeaker setup or encoding the soundfield to a binaural signal. In these cases the visualization algorithm may provide information about audible cues and sounds in directions difficult to discern through sound alone.

5.3 Conclusions

The results of this thesis show clear improvements in correlation between the reconstructed signals from the filters and reference signals, especially considering

the low quality and lack of calibration of the equipment used in the evaluation process. Transforming the B-format signal into the audio source domain using the proposed analysis method provides an intuitive abstraction for manipulating the spatial properties of soundfields. The filter parameter abstraction of the audio source domain arguably provides an intuitive interface to work with the spatial properties of the audio sources and signal itself. The types of filters presented are also shown to be useful in practical application such as sound source separation and foreground/background separation.

5.4 Future work

It should be noted that the practical evaluations of the filters proposed in this thesis were performed with less than ideal equipment with very limited calibration applied. It is expected that performing the evaluations using professional and tuned equipment could generate even better results. Oppositely, the same evaluations were carried out in an anechoic environment and few sound sources. Investigating the effects of reverberation and more sound sources on the analysis could be the subject of further research on the subject. The visualization technique proposed in the thesis provides an intuitive way to visualize the soundfield at a point in time similar to how spectrograms may visualize the spectral content of recordings over time. Some audio editing tools, such as the discontinued *Adobe Soundbooth*, allows the user to draw areas in the spectrogram of a sound file and apply gain or attenuation in the limited time-frequency scope of said areas, as if drawing in the spectrogram. It would be interesting to see a similar tool for visually manipulating B-format soundfields using the soundfield visualization as a user interface for width- and DOA-filters.

References

[1] M.A. Gerzon, *Periphony: With-height Sound Reproduction*, Journal of the Audio Engineering Society, Vol. 21 No. 1 Jan/Feb 1973 pp.2-10

[2] Gerzon, M.A. *The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound*, Presented at the 50th AES Convention, London, March 1975.

[3] D.T. Hemingson *Construction of an Experimental Tetrahedral Ambisonic Microphone*, http://danh.coffeecup.com/pdf/Exp2%20tetrahedral%20frame%20wiring.pdf

[4] M, Kronlachner *Spatial Transformations for the Alteration of Ambisonic Recordings* https://iem.kug.ac.at/fileadmin/media/iem/projects/2013/kronlachner.pdf

[5] H. L. Kennedy, *A New Statistical Measure of Signal Similarity*, Conference: Information, Decision and Control, 2007. IDC '07 https://www.researchgate.net/publication/4256278_A_New_Statistical_Measure_of_Signal_Similarity