HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation
Acoustics and Audio Signal Processing

Oskari Porkka

Modification of multichannel audio for non-standard loudspeaker configurations

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, Apr 6, 2009

Supervisor:	Prof. Matti Karjalainen
Instructor:	D.Sc. (Tech) Aki Härmä

HELSINKI UNIVERSITY OF TECHNOLOGY			ABSTRACT OF THE MASTER'S THESIS
Author:	Oskari Porkka		
Name of the thesis:	Modification of multichannel audio for non-standard loudspeaker configurations		
Date:	Apr 6, 2009	**Number of pages:**	71
Faculty:	Faculty of Electronics, Communications and Automation		
Professorship:	S-89		
Supervisor:	Prof. Matti Karjalainen		
Instructor:	D.Sc. (Tech) Aki Härmä		

In this thesis, analysis and decomposition methods for multichannel audio are studied. The objective of the work is to transform multichannel recordings to new reproduction systems so that the spatial properties of the sound are preserved. Spatial hearing of the human auditory system, signal-based similarity and localization measures, and information-technological source separation methods are described as background theory. Then, different multichannel audio transform methods are reviewed. The experimental part of the work starts with an analysis of DVD recordings to gain helpful information about the production methods of such recordings for further development of audio transform methods. The test reveals that the three frontal channels do not usually share common sound sources with the two rear channels. The properties of compact loudspeaker systems are investigated in two listening tests. The first test studies the differences between three-channel loudspeaker layouts, which exploit the reflections of sound waves from room boundaries. The latter one of the tests applies three transform methods known from the literature to widen the spatial dimensions of a three-channel compact loudspeaker system in comparison to a reference stereo system. These methods are a stereo signal transform method based on signal powers and interchannel cross-correlations, a primary-ambient signal decomposition based on principal component analysis (PCA), and directional audio coding (DirAC). The methods were ranked in this descending order of preference by the test subjects.

Keywords: Multichannel audio, stereo, spatial sound, spatial hearing, localization, audio coding, correlation, blind source separation, principal component analysis, format conversion, compact loudspeaker systems

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Oskari Porkka		
Työn nimi:	Monikanavaäänen muokkaaminen epästandardeja kaiutinkokoonpanoja varten		
Päivämäärä:	6.4.2009	**Sivuja:**	71
Tiedekunta:	Elektroniikan, tietoliikenteen ja automaation tiedekunta		
Professuuri:	S-89		
Työn valvoja:	Prof. Matti Karjalainen		
Työn ohjaajat:	TkT Aki Härmä		

Tämä diplomityö käsittelee monikanavaäänen analyysi- ja hajotelmamenetelmiä. Työn tavoitteena on pystyä muokkaamaan monikanavaäänityksiä uusille kaiutinkokoonpanoille siten, että äänen tilaominaisuudet säilyvät. Teoriataustana työssä ovat ihmiskuulon tilahavainnointiominaisuudet, äänisignaaleihin perustuvat samankaltaisuusmitat sekä suunta-arviot ja informaatioteknologian lähde-erottelumenetelmät. Työ käy läpi kirjallisuudesta löytyviä monikanavaäänen muokkausmenetelmiä. Diplomityön kokeellisen osuuden aloittaa DVD-levyjen analyysi, jolla pyrittiin saamaan tietoa levyjen äänituotannossa käytettävistä menetelmistä myöhempää äänimuunnostekniikoiden kehittämistä varten. Koe osoitti, että kolmen etukanavasignaalin ja kahden takakanavasignaalin välillä on vain harvoin yhteisiä äänikomponentteja. Kompaktien kaiutinkokoonpanojen ominaisuuksia tutkittiin kahdessa kuuntelukokeessa. Ensimmäinen koe tarkasteli eroja eri kolmikanavaisten kaiutinasettelujen välillä. Tavoitteena näissä toistosysteemeissä oli hyödyntää ääniaaltojen heijastuksia huoneen seinistä. Jälkimmäinen kuuntelukoe sovelsi kolmea tunnettua äänimuunnosmenetelmää kolmikanavaiseen kompaktiin kaiutinkokoonpanoon, jonka toistosta saatavaa tilahavaintoa pyrittiin laajentamaan. Kahden metodeista havaittiin parantavan tutkittuja tilaominaisuuksia.

Avainsanat: Monikanavaääni, stereo, tilaääni, tilakuulo, suuntakuulo, äänikoodaus, korrelaatio, lähde-erottelu, pääkomponenttianalyysi, äänen toistojärjestelmän muunnos, kompaktit kaiutintoistojärjestelmät

Acknowledgements

This master's thesis has been done for Philips Research. The research work was conducted in Eindhoven, the Netherlands during the year 2008.

I want to thank Aki Härmä, my instructor at Philips Research, for all the guidance and support for my work. His professional expertise has helped me in solving numerous theoretical and practical problems. I am grateful for his comments on my thesis text. My gratitude also goes to Werner de Bruijn, Ronald M. Aarts and John Lamb for their aid. Thanks to Bert den Brinker for lending his spatial audio processing book for my whole stay.

I wish to thank professor Matti Karjalainen from Helsinki University of Technology for his feedback and help in the writing process.

I would also like to thank the interns who shared the office with me. They achieved to maintain a stimulating working environment. All my numerous flatmates from Eindhoven deserve compliments for friendship.

Finally, I would like to thank my family for their support, especially my beloved Sonja, who patiently waited for me at home while I was away in my distant location.

Espoo, April 6, 2009

Oskari Porkka

Contents

Abbreviations

ASW	Auditory source width
BCC	Binaural cue coding
BSS	Blind source separation
DFT	Discrete Fourier transform
DirAC	Directional audio coding
ERB	Equivalent rectangular bandwidth
FFT	Fast Fourier transform
HRTF	Head-related transfer function
IACC	Interaural cross-correlation
IC	Interchannel coherence
ICA	Independent component analysis
ICC	Interchannel cross-correlation
ICLD	Interchannel level difference
ICTD	Interchannel time difference
IFFT	Inverse fast Fourier transform
IID	Interaural intensity difference
ILD	Interaural level difference
IPD	Interaural phase difference
ITD	Interaural time difference
LEV	Listener envelopment
LFE	Low frequency effect
PC	Principal component
PCA	Principal component analysis
SCA	Sparse component analysis
SIRR	Spatial impulse response rendering
SPCA	Sparse principal component analysis
STFT	Short-time Fourier transform
VBAP	Vector base amplitude panning

Chapter 1

Introduction

Nowadays, vivid sensations of spatial sound can be generated by loudspeaker reproduction systems. The first audio recordings were one-channel mono signals, and a big step boosting the spatiality of the sound recordings was when the second loudspeaker was introduced to the playback system. The stereo sound system introduced by Blumlein in 1931 [6] could produce phantom sound images, virtual sound sources that are localized between the two loudspeakers. Recently, the number of loudspeakers in reproduction systems has grown, which has led to increasing surround sensation in the sound playback. The surround sound systems are especially used to give realistic video and movie watching experiences. A popular surround audio reproduction system is the 5.1-surround [28] but even a 22.2-surround sound system has been proposed for the future high definition audio-visual content [47].

The mixing of commercially available audio material is usually optimized for playback from standardized loudspeaker layouts. The recordings are therefore format-dependent and they should be played back using these specified loudspeaker layouts. The standards specify the correct number and placement of loudspeakers, but the audio reproduction systems of consumers do not always meet these standards. All the loudspeakers may not be in their standardized positions or the number of loudspeakers might be even smaller or larger than what is required for faithful reproduction of the audio material. With the current audio and movie players, the signal is fed to the loudspeakers without much of configuration-specific treatment, and therefore the use of an incompatible playback system modifies the intended spatial image of the audio material.

Practical reasons can make the proper placement of loudspeakers complicated. Room dimensions or furnishing aspects may lead to compromises over the playback system. For example, the conventional 5.1 system requires five speakers placed in a circle around the listener. The demanded loudspeaker layout is difficult to realize in small living rooms, and the consumers place the loudspeakers often in non-standard configurations, or even

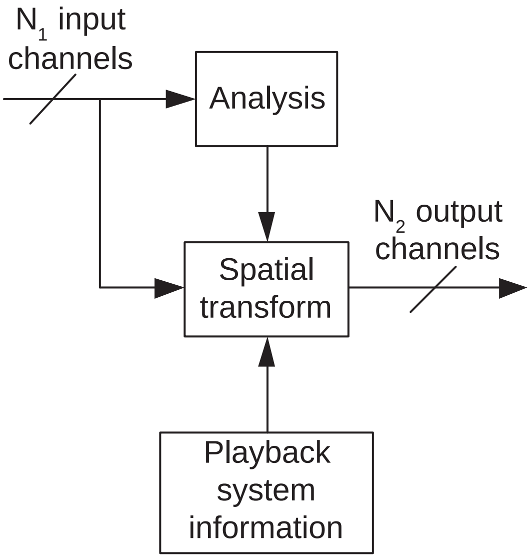


Figure 1.1: Overall block diagram of the spatial transform system.

leave some loudspeakers uninstalled. Many manufacturers have developed compact audio systems for the surround sound reproduction. These systems have smaller number of loudspeakers and they have less strict placement requirements. In the compact loudspeaker systems, the surround sound sensation is achieved by utilizing room reflections together with a combination of stereo dipole processing and acoustic properties of the loudspeaker system. This may lead to an enveloping listening experience but also to an inaccurate stereo image, when a sound source that is clearly positioned in the stereo image can become a part of the diffuse surround sound field.

The mismatch between the standards and the actual loudspeaker configurations motivates to develop new methods that transform audio content to a more suitable form considering the available reproduction system. This can be considered as a spatial remixing of the audio recording by first reverse mixing the audio content and then constructing a new mixture from the unmixed sources for the target system. A block diagram of such a transform system is illustrated in Figure 1.1. In the block diagram there are N_1 input channels which are analyzed in the analysis block. The analysis results and information about the current playback system are fed as parameters to the transform block, which finally produces N_2 appropriate output channels. In the general system any number of output channels can be chosen.

This report studies signal processing techniques that can be used for the modication of multichannel audio for non-standard loudspeaker configurations. Throughout the work, only audio formats and reproduction systems having the loudspeakers on a single horizontal plane are considered. This restriction comes from the fact that the mostly used consumer loudspeaker systems, namely the two-channel stereo and five-channel surround, have this property. The low frequency effect (LFE) channels of the multichannel audio formats are not concerned, because of the poor localization ability of the human auditory system at low frequencies. Special interest is given to compact loudspeaker systems as a non-standard system that aims at producing a surround sound field.

The following chapter describes analysis methods that are used to measure spatial properties from the audio signals. The spatial hearing of the human auditory system is first discussed, and after that emphasis is put on more signal-oriented analysis methods. Chapter 3 covers current spatial transform techniques used for stereo and multichannel audio. In Chapter 4, a statistical analysis of commercially available DVD audio recordings is presented. The analysis was conducted to gain better understanding of the mixing techniques used for the surround sound production of the present day multichannel audio. Chapter 5 covers the listening experiments that are part of this work. First an initial listening test studying the properties of compact loudspeaker systems is presented. The later experiments study the properties of the multichannel audio transform techniques presented in Chapter 3. Finally, conclusions are made in Chapter 6.

Chapter 2

Spatial audio analysis

This work is a study on the modifications of multichannel signals from a format to another. The spatial properties of the multichannel audio signals are wanted to be preserved as much as possible in the transform. These properties need to be analyzed before the required spatial decomposition can be done. The objective of this chapter is to introduce the fundamental topics of the spatial audio analysis to the reader. We may define the sound image as a spatial representation of sound sources and acoustics perceived by a listener in a listening point. Therefore, it is natural to start this chapter by an overview of spatial hearing, which is the spatial analyzer of the human auditory system. The spatial hearing is discussed in Section 2.1, and the primary focus of the section is on giving an overall description of the mechanisms and especially the cues that the human auditory system uses to localize sound events and to sense ambience from the sound environment. More extensive reviews on these topics can be read from the books by Blauert [4] and Moore [36].

Another question is how the spatial properties can be analyzed from audio signals that have two or more signal channels. The recording engineer intends to produce a certain type of spatial sound image by choosing the appropriate recording and mixing techniques. Before the audio mixtures can be analyzed for the modification purposes, it is good to have some knowledge about the production of the recordings. Discrete sound sources are normally desired to be perceivable as coming from various directions in the production of multichannel audio signals. The loudspeaker layout is often sparse, however, and the sound sources need to be placed somewhere between the loudspeaker channels by the recording engineer. Mixing and microphone techniques that are used to achieve a rich spatial image are being discussed in Section 2.2.

Multichannel audio modification techniques require tools that detect the locations of the audio sources from the mixture. The reverse mixing or unmixing processes demand the measurement of similarities between the audio channels and the estimates for the localization directions of the audio events from the multichannel mix. This is a strongly linking factor with multichannel audio coding techniques, which aim at reducing the redundant parts of the multichannel signals. These techniques rely significantly on measurements of interchannel relationships, which are beneficial also for the modification purposes. These measures are covered in Sections 2.3 and 2.4.

Ideally, an audio format transformation process could separate the original sound events from the mixture and then remix them for the new loudspeaker system. The spatial sound reproduction capabilities of the new system could be then maximally exploited, and the spatial leaking of the sound sources to wrong directions could be avoided as much as possible. The modern information technology has developed tools for such an unmixing process. The tools are called source separation techniques, and they have been developing rapidly since the increasing availability of computational power in the 1990's. However, the techniques are quite often computationally complex and limited. The source separation techniques are discussed in Section 2.5.

2.1 Spatial hearing

Human beings can easily detect the location of a sound event, and the size of the sound source. The sound waves reflect from boundaries, which affects our spatial sound sensation. There are considerably less boundaries causing reflections outdoors than indoors, where the reflections change the spatial properties of the sound sources and allow us to estimate also the size of the space where we are, namely the room. The spatial sound, in general, can be divided in two separate kinds of sound objects:

1. sounds that have distinct locations
2. ambient sound that is difficult to be localized and has a diffuse character.

The ambient sound gives us the sensation of an auditory scene that envelops or immerses the listener [45].

The sound event localization of the human auditory system relies mainly on the differences between the signals that are received by the two ears. These localization cues are called interaural differences. The two primary types of differences are time differences and amplitude differences. The localization cues vary as functions of the direction from where the sound arrives. The time differences are due to inequal travel path lengths from the sound source to the ears, which cause the sound waves of the same event to reach the ears at different time instants. This kind of localization cues are called either interaural time differences (ITD) or interaural phase differences (IPD). The amplitude-related cues are called interaural level differences (ILD) or interaural intensity differences (IID). The shape of the head causes different levels of attenuation to the sound on different sides of the head, which enables the use of amplitude differences as localization cues. The size of the head, however, is not enough to attenuate very low frequency signals, for which the level differences cannot be used for localization. The significancy of ILD, therefore, increases when the frequency of the arriving sound increases. At low frequencies, more importance in localization has to be given to interaural time differences. It has been suggested that the ITD is the main cue at low frequencies and ILD dominates at high frequencies. The crossover frequency between ITD- and ILD-based localization is at around 1600 Hz. The direction detection around the crossover frequency somewhat relies on both of the two cues [4].

The shape of pinna, which is the visible part of the outer ear, causes reflections and resonances, which modify the spectral content of the sound arriving to the ears. The changes in the spectrum are again direction-dependent, and this dependency allows using them for localization purposes as well, in addition to previously-mentioned ITD and ILD. The spectral changes can be summarized to head-related transfer functions (HRTF), which are unique to all listeners [45]. The elevation of sound sources is determined using HRTFs, which are particularly important in binaural simulation [23] and virtual reality [37].

Measuring localization with ITD and ILD parameters stays rather simple as long as there is only one active sound source and there are no reflections from boundaries. Normally this is not the case. In rooms, there are reflections, which affect the localization. The regular two-channel stereo listening is also an example of more than one sound source. Localization for this kind of cases is more complicated. The similarity of the many sounds arriving to the ears has to be taken in account. The level of similarity between sound waveforms can be measured with the normalized cross-correlation function and its maximum absolute value. If the maximum absolute value of the normalized cross-correlation function is 1, the sound waves are coherent [4], which means that they are identical. Although called identical, the sound waves can have possible level differences and delays, or they can be even phase-inverted versions of each others. When the normalized cross-correlation is measured from the ear input signals, it is called interaural cross-correlation (IACC) or interaural coherence. The calculation of the similarity measures will be discussed in more detail in Section 2.3.

The human auditory bandwidth is divided in several frequency bands called the critical bands. The sound events that overlap at the same critical band are often perceived as coming from the same source. The dominant source on that band therefore also dominates the localization cues. This makes it feasible for the signal processing algorithms to process the signal individually on each critical band. Indeed, this property is used in various audio coding methods to reduce the redundant non-audible information from the audio signals [2], [7], [13], [24], [52]. A popular replacement for the actual critical bandwidths are equivalent rectangular bandwidths (ERB), which can be calculated from

$$ERB(f) = 0.108f + 24.7, \tag{2.1}$$

where f is the center frequency in hertz. ERBs are closely related to critical bandwidths, but they have been measured using different methods [26].

As mentioned earlier, the sound reflections from boundaries affect our perception of the auditory environment. The localization is usually done according to the first wavefront reaching the listener. This is called the precedence effect. Besides the localization, there are other things that the auditory system perceives for the sound event. For example, we can estimate the size of the sound source. Auditory source width (ASW) is a term relating to the perceived width of the sound source, and it is particularly important in concert hall acoustics. Our auditory system perceives the apparent source width using the lateral reflections of the sound arriving during the first 5-80 ms after the direct sound [30]. Interaural cross-correlations that have been measured with time windows of up to 80 ms have shown correspondence with ASWs perceived in concert halls [45]. Listener envelopment (LEV) is another term that is closely related to concert hall acoustics and room acoustics. It is used to describe the spaciousness of the room, and depicts how much the listener feels like being surrounded by the sound. LEV has been applied also to reveal the spatial properties of sound reproduction systems [45], [46]. Similar to auditory source width, the listener envelopment is achieved by the lateral reflections. LEV can be measured as the energy fraction of the lateral reflections arriving after 80 ms and the energy of the direct sound. Roughly speaking, the lateral reflections arriving before 80 ms contribute to ASW and those arriving after 80 ms increase LEV. [30]

The localization of amplitude-panned virtual sources, or phantom sources, is under special interest when spatial audio reproduction is considered. Interaural time and level differences are the main localization cues for the stereophonic loudspeaker playback as well. Pulkki and Karjalainen [44] studied the behaviour of the localization cues in stereo listening. They reported that the low frequency ITD and high frequency ILD cues behave consistently for the same direction of the virtual source. There is a region between 1100 and 2600 Hz where the cues deviate from each other. It has been discussed that narrow-band virtual sources on this region might have spread out localization [44]. The previous results apply mainly for sources on the median plane, for which the ITD and ILD are the main localization cues. Additional localization cues are needed for virtual sources that have been elevated from the median plane [40]. In this work, only sources on the median plane are considered, however.

2.2 Production methods for multichannel audio

It is essential to have preliminary knowledge about the production process of audio signals, when the audio unmixing process is considered. The audio production can be divided roughly in two approaches: One is to capture the sound field directly with microphones and reproduce it using the loudspeakers. The other involves capturing or synthetizing the sound sources individually, and then panning them to appropriate locations in the multichannel mix. These two approaches can be of course combined in the audio recording and mixing.

The pure microphone recording techniques aim at capturing the sound sources in their original surround environments. The “dry” source signals, the direct sound waves from the sound source, and the “wet” signals that contain the echos and the reverberation are merged in the recorded audio tracks from the very beginning. Several kinds of different microphone placements can be used for both stereo and multichannel recording. The microphone signals are not fed directly to the loudspeakers, but processing or at least matrixing operations need to be done in forming the loudspeaker signals. [45]

The mixing process of single sources is as follows: In the beginning the mixing engineer has several individual recorded or synthetized audio tracks which can be instruments, singing or sound effects for example. The main problem is how to mix these sources to the available number of output channels so that the mixture has rich and enveloping spatial characteristics and the levels of the sources are in correct balance. One of the simplest forms of building a mixture is amplitude panning, which generally means giving each source signal output channel -dependent coefficients. This can be depicted in a matrix form as

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t), \tag{2.2}$$

where m output channels $\mathbf{x}(t)$ are derived from n source signals $\mathbf{s}(t)$ using $m \times n$ mixing matrix $\mathbf{A}$ that consists of real values. The amplitude panning in multichannel audio production is often pair-wise panning between loudspeaker pairs. If there are more than two output channels m, and if the n source signals $\mathbf{s}(t)$ are present only in separate loudspeaker pairs, the mixing matrix $\mathbf{A}$ will be sparse in the sense that it will contain significant amount of zero values. The energy-preserving amplitude panning rule ensures that the energy of the source signal that is present in many loudspeaker channels must remain equal to the original. Thus, the energy-preserving panning coefficients $a_{1,l}$-$a_{m,l}$ of the l:th source signal of Equation (2.2) must satisfy

$$\sum_{k=1}^{m} a_{k,l}^2 = 1. \tag{2.3}$$

There are two main panning laws that have been derived for the amplitude panning of the conventional stereo playback. These laws give the relationships between the angles ϕ

of the stereo layout, apparent azimuth θ of the virtual source and the amplitude panning coefficients. The law of sines gives the relation

$$\frac{\sin\theta}{\sin\phi} = \frac{a_2 - a_1}{a_1 + a_2}, \tag{2.4}$$

where a_1 and a_2 are the amplitude panning coefficients for the two loudspeakers. They can be easily determined for the desired direction angle θ of the virtual source by using the energy-preserving rule of Equation (2.3), for example. The other rule, the law of tangents has been said to represent better the situation when the listener is allowed to move his or her head [3]. The formula of the law of tangents differs from the law of sines in that tangents are used instead of sines:

$$\frac{\tan\theta}{\tan\phi} = \frac{a_2 - a_1}{a_1 + a_2} \tag{2.5}$$

There is not much difference between the values given by the two laws. If the amplitude panning coefficients remain the same, and the loudspeaker angles of $\phi = 30°$, the difference of the apparent direction angle θ given by the two laws is $1.7°$ at maximum [44]. Vector base amplitude panning (VBAP) generalizes the law of tangents for any two- or three-dimensional loudspeaker setup [39]. Other, non-pair-wise types of panning laws have been also proposed for multichannel loudspeaker setups. For example, Gerzon [18] derived optimal pan-pot laws for a four-channel surround setup. Later, the 5-channel panning laws have gained more interest [11], [31], [51]. Complex panning laws that use phase shifts and delays may have weaknesses in comparison to simple pair-wise amplitude panning, however. They are more sensitive to the head movements of the listener, and the moving phantom images might be unstable. The pair-wise amplitude panning, by contrast, is a simple but robust solution for many cases [45]. Pair-wise panning has remained popular in multichannel audio, because it is simple but stable, although it has limitations too.

2.3 Channel similarity measures

One of the first things in the unmixing process is to determine what is the level of similarity between the audio channels. Channel similarity measures are specially important in multichannel audio coding techniques, which aim at decreasing the amount of bits required to represent the audio data by detecting the similarities between the channels. Many of the parameters used in audio coding are perceptually motivated in the sense that they are near to localization parameters that were presented for spatial hearing in Section 2.1. MPEG surround [24], for example, uses interchannel time differences (ICTD), interchannel level differences (ICLD) and interchannel cross-correlation (ICC) or coherence (IC), which all have their corresponding interaural measures. Usually, the interchannel relationships are calculated individually for frequency bands which mimic the critical bands of the auditory system [2], [7], [13], [24] [52].

There are several measures that can be calculated to express the level of similarity between two or more audio signals. Some of these methods can be calculated in the time domain, some in the frequency domain and some in the both domains. Selecting the calculation domain depends of the application that the calculation is intended for. Different measures require different amount of computational power in different domains. In addition, some parameters of calculation can be changed in one domain more flexibly than in another domain. The time-related parameters are easy to change in the time domain, while the frequency-related parameteres are more flexible in the frequency domain.

The cross-correlation function is a very common similarity measure between two signals. The similarity is measured as a function of a time shift, which is applied to one of the signals. The cross-correlation function of two discrete-time signals, $x_1(k)$ and $x_2(k)$, is

$$\phi_{12}(\tau) = \sum_k x_1(k)x_2(k+\tau), \tag{2.6}$$

where τ is the time lag between the signals. The larger the value of the function gets, the more similar the signals are when that particular time shift is applied to one of them. If there is a negative peak in the function, it means that the signals are similar but phase inverted.

The sum in the cross-correlation function is calculated over the whole signals from time indices $-\infty$ to ∞ in a strict mathematical notation. In signal processing, the instantaneous similarities are more interesting, and the complete signal is not necessarily known. Therefore, short-term rather than long-term signals are usually analyzed. The short-term similarities between the signals can be measured using the short-time cross-correlation functions, which can be calculated by applying the appropriate window functions to the parts of the signals that are wanted to be analyzed. The short-time cross-correlation function can be written as

$$\phi_{12}(\tau) = \sum_k x_1(k)\omega_1(k)x_2(k+\tau)\omega_2(k+\tau), \tag{2.7}$$

where $\omega_1(k)$ and $\omega_2(k)$ are the window functions for the signals $x_1(k)$ and $x_2(k)$, respectively. Another reason for the use of the short-time cross-correlation can be the restrictions of computational power. The sum in Equation (2.6) can be also described as an inner product of two infinitely long vectors, which requires calculating infinite number of multiplications. It takes less multiplications to calculate the cross-correlation over shorter signals.

Let $X_1(i)$ and $X_2(i)$ be the discrete Fourier transforms (DFT) of the signals $x_1(k)$ and $x_2(k)$. The cross-correlation theorem of the Fourier transform [50] states that the Fourier transform of the cross-correlation function of two signals is equal to the cross-spectrum of the signals, that is

$$\mathcal{F}\left\{\sum_k x_1(k)x_2(k+\tau)\right\} = X_1(i)X_2^*(i), \tag{2.8}$$

where $\mathcal{F}\{\}$ denotes the Fourier transform and taking the complex conjugate is marked with *. Here again, the Fourier transforms $X_1(i)$ and $X_2(i)$ are calculated over the entire signals. The cross-correlation theorem of Equation (2.8) applies also to the short-time cross-correlation function of Equation (2.7) in the case that $X_1(i)$ and $X_2(i)$ are the Fourier transforms of the signals $x_1(k)$ and $x_2(k)$, and they have been windowed using the window functions $\omega_1(k)$ and $\omega_2(k)$.

The use of the cross-correlation theorem together with the fast Fourier transform (FFT) [10] can greatly reduce the amount of computation time required for the calculation of the cross-correlation function. The computational complexity of the direct calculation of the cross-correlation function for two N-length signals requires $O(N^2)$ operations. The FFT, by contrast, requires $O(N\ \log(N))$ operations [10]. This benefit, therefore, becomes more and more significant when longer signals are concerned. Another advantage of computation in the frequency domain is that the cross-correlation can be easily calculated for sub-bands. If the cross-correlations are wanted to be calculated for frequency bands in the time domain, these bands must be first filtered from the original signal and then individual cross-correlation functions need to be calculated for each pair of bandpass-filtered signals in the time domain. In the Fourier domain, by contrast, it is enough to calculate the cross-spectrum only for the respective Fourier bins of each band. There is not an additional computational cost for the division in sub-bands. If the cross-correlation functions are desired to be transformed back to the time domain, however, separate inverse FFT (IFFT) operation is needed for each band, and the number of required inverse transforms increases in comparison to only one inverse tranform needed for the whole band.

The computational analysis of the whole cross-correlation function would be time-consuming. Often, the only interesting data is the amount of correlation between signals rather than the values of the cross-correlation function at different time lags. The correlation can be measured with simple correlation coefficients, which are usually normalized to vary from -1 to 1 or from 0 to 1. Division by the square root of the signal powers or the standard deviations of the signals is a common normalization method. A popular correlation measure is the Pearson's product-moment coefficient [49], which is calculated by dividing the covariance between the two signals by the standard deviations of the signals. Let these two signals be called x and y, and the coefficient can be calculated using

$$r_{x,y} = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x\sigma_y}, \tag{2.9}$$

where $E[\ ]$ is the expected value operator and μ and σ denote the means and the standard deviations of the signals, respectively.

There are also a number of other correlation coefficients that can be used for different situations. The maximum value of the cross-correlation function is an example of a simple coefficient measure. The negative peaks can be taken in account by using the maximum absolute value of the correlation function. In some applications, the maximum value or the maximum absolute value is looked for only from a specific interval of time lags. This can especially be the case when interaural cross-correlations are measured because of the restrictions on how the human auditory system detects the correlation. For example, an interval of lags from -1 to 1 ms is used in binaural cue coding (BCC) [2]. The zero-lag value of the cross-correlation function can be used as a correlation measure too, but this does not necessarily detect the correlation when a time-shifted signal is compared to the original. One can notice that Pearson's correlation coefficient does not detect correlation between time-shifted signals, by taking a look at the Equation (2.9).

In the frequency domain, there are less possibilities to choose how the correlation coefficient will be calculated. One formula for the correlation coefficient calculates the normalized sum over the cross-spectrum:

$$\rho_{12} = \frac{\sum_i X_1(i) X_2^*(i)}{\sqrt{\sum_i X_1(i) X_1^*(i)}\sqrt{\sum_i X_2(i) X_2^*(i)}}. \tag{2.10}$$

This kind of cross-correlation coefficient has a few variations. The sums can be calculated over different frequency bins, or the absolute value can be taken from the sum. The analysis of these options will be started from the formula that gives the inverse discrete Fourier transform (IDFT)

$$x(k) = \frac{1}{N}\sum_{i=0}^{N-1} X(i) e^{\frac{2\pi j}{N} ik}, \tag{2.11}$$

where $x(k)$ is the signal in the time domain, $X(i)$ is the Fourier transform of the signal, N is the length of the DFT and j denotes the imaginary unit. If only the signal value at the time instant $k = 0$ is under interest, Equation (2.11) gets the form

$$x(0) = \frac{1}{N}\sum_{i=0}^{N-1} X(i). \tag{2.12}$$

Correspondingly, if the value of the cross-corelation function $\phi_{12}(\tau)$ at zero-lag $\tau = 0$ is needed, summing over the whole cross-spectrum $X_1(i)X_2^*(i)$ of the two signals $x_1(k)$ and $x_2(k)$ gives the desired result

$$\phi_{12}(0) = \frac{1}{N}\sum_{i=0}^{N-1} X_1(i) X_2^*(i). \tag{2.13}$$

This value can be then used as a correlation coefficient.

If the values of the signals are real numbers, the cross-correlation function between them is also real. The Fourier transforms of real signals are complex, but have conjugate symmetry between the negative and positive frequencies. This causes that the sum over the cross-spectrum of real signals has a real value, because the imaginary parts of the complex conjugates vanish when summed. In audio literature, however, the real part is sometimes taken from the sum over the numerator in Equation (2.10) [7]. This is unnecessary as the sum is real for real signals by definition, and the notation can be confusing for this reason. In numerical computing, however, a small imaginary part might appear because of round-off errors. Taking just the real part is justified in this sense. The conjugate symmetry also allows the other half of the spectrum to be ignored for the sake of the computational efficiency. Since the result is then complex, the imaginary part of the result must be omitted by taking only the real part.

As mentioned earlier, the correlation coefficient that is measured at the zero lag does not notice phase shifts. Calculating the sum over the cross-spectrum, thus, does not necessarily give a good measure of correlation. Summing only over the non-negative frequencies, which include the Nyquist and zero frequencies, gives a complex result and thus preserves some phase information. This can be exploited, and the absolute value of the sum can be used as a better correlation measure, which also caters for the peaks of the correlation function near the zero lag. This is given in mathematical form by Breebaart and Faller [7]:

$$\rho_{1,2} = \frac{\left|\sum_{i=0}^{N/2} X_1(i) X_2^*(i)\right|}{\sqrt{\sum_{i=0}^{N/2} X_1(i) X_1^*(i)} \sqrt{\sum_{i=0}^{N/2} X_2(i) X_2^*(i)}}. \tag{2.14}$$

Notably, this gives only correlation values between 0 and 1, and thus does not preserve information about inverse phases. The measure can be called interchannel coherence similarly to interaural coherence, which was mentioned in Section 2.1.

The geometrical mean of signal powers, which was used in Equations (2.10) and (2.14), is a common normalization term for correlation coefficients and functions. Other kinds of divisors can be used as well. An alternative normalization term that is based on signal powers was presented as an audio signal similarity measure by Avendano and Jot [1]. They use the regular arithmetic mean for normalization:

$$\rho_{1,2} = \frac{\left|\sum_{i=0}^{N/2} X_1(i) X_2^*(i)\right|}{\left[\sum_{i=0}^{N/2} X_1(i) X_1^*(i) + \sum_{i=0}^{N/2} X_2(i) X_2^*(i)\right] / 2}. \tag{2.15}$$

The value given by Equation (2.15) is also normalized between 0 and 1 as well, but it will depend more on the relative powers between the signals in comparison to the normalization by geometrical mean. For example, if the signal powers are 1000 and 1, their geometrical mean will be around 32, but their arithmetic mean will be around 500. Even though the signals would be very correlated, the correlation coefficient normalized by the arithmetic mean would show the correlation to be neglible in comparison to the coefficient that uses geometrical mean, which would indicate much stronger similarity between the signals.

2.4 Directional analysis of audio signals

The localization of sound events in multichannel audio listening is different than in listening to a natural sound source. The direction of a phantom source that has been panned between two loudspeakers on the median plane is perceived using the summing localization mechanism, and the localization is based on ITD and ILD cues, which were described in Section 2.1. In the analysis of multichannel audio mixes, it is hard to estimate the localization directions of different sound events, because there are several concurrent source signals present in the signal channels. The ability to localize sound events from the mixtures is an important part of the spatial analysis of multichannel signals, however. There are some ways to approximate the directions.

Many localization approximation methods estimate how the directions of the sound events are perceived at the sweet spot of the loudspeaker system. The apparent location of the phantom source can be represented with localization vectors, which are usually calculated by weighting the format vectors of the loudspeaker system. The format vectors are direction vectors that have their initial point at the sweet spot and the terminal point at the standard locations of the loudspeakers. The lengths of the vectors are determined by the distances between the loudspeakers and the optimal listening position. When two-channel stereo and five-channel surround signals are analyzed, unit vectors pointing to the directions of the standard loudspeaker locations can be used, because the loudspeakers are equidistant from the sweet spot. The format vectors for these reproduction systems are illustrated in Figure 2.1. The format vectors $\vec{c}_1$-$\vec{c}_m$ of a m-channel system form a format matrix $\mathbf{C}$, which is given by

$$\begin{aligned}\mathbf{C} &= \begin{bmatrix} \vec{c}_1 & \vec{c}_2 & \dots & \vec{c}_m \end{bmatrix} \\ &= \begin{bmatrix} \sin\phi_1 & \sin\phi_2 & \dots & \sin\phi_m \\ \cos\phi_1 & \cos\phi_2 & \dots & \cos\phi_m \end{bmatrix},\end{aligned} \tag{2.16}$$

where the angles ϕ_1-ϕ_m represent the directions of the loudspeakers.

Gerzon [17] presented localization vectors that estimate the apparent direction of the arriving sound. Two simplest types of these localization vectors are called the velocity-based localization vector and the energy-based localization vector. They assume that the sound is perceived at the sweet spot. The velocity vector models the particle velocity of the

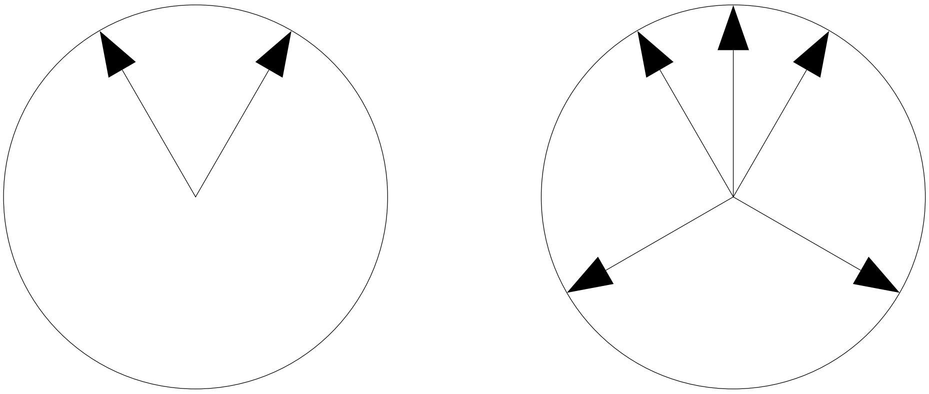


Figure 2.1: The format vectors of two-channel stereo and five-channel surround.

sound waves. Ideally, when there are appropriate measurements of the sound field available, it is very straightforward to calculate the particle velocity vector. The particle velocities can be estimated using the signal values of the loudspeaker channels. The format vectors $\vec{c}_1$-$\vec{c}_m$ of the loudspeaker layout $\mathbf{C}$ are multiplied by the signal values x_1-x_m

$$\vec{u} = -\mathbf{C} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}, \tag{2.17}$$

and the direction of the velocity-based localization vector $\vec{u}$ estimates the perceived arrival direction of the sound. Gerzon refers to this kind of localization approximation as the Makita localization. The calculation of the energy vector is very similar to that of the velocity vector. However, this time the format vectors are weighted by the signal energies rather than the signal values themselves:

$$\vec{g} = \mathbf{C} \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_m^2 \end{bmatrix} \tag{2.18}$$

The velocity vector uses a first degree signal magnitude to estimate the localization direction, while the energy vector is based on second degree values. They are thus called first degree and second degree localization models by Gerzon [17]. In later audio literature, the energy vector has been also called the Gerzon vector [19].

Usually, it is not feasible to calculate the localization vectors for every time instant. The directions and magnitudes of the vectors might change significantly in fast phase. Averaging over time should be considered to avoid rapid changes. The localization vectors can be also calculated for frequency-domain signals. Calculations in the Fourier domain include already averaging in the time domain because of the time-frequency resolution characteristics of the Fourier transform. The localization vectors could be calculated for each Fourier bin, but averaging over frequency bands can be beneficial. It is also supported by the frequency selectivity of the human auditory system. The frequency-band averaging has been used in present day methods [19], [20], [21], [43], [41], [42].

2.5 Blind source separation techniques

Recovering original source signals from a signal mixture is one of the fundamental problems of signal processing. This ability could be beneficial in removing audible signal artefacts or noise. Different kind of signal processing could be applied only to the desired part of the signal mixture. A real-life source separation process is the ability of human beings to follow one speaker in a situation where several persons are speaking on top of each others. This classical example of source separation is called the “cocktail party problem”. Information technology has developed a number of blind source separation (BSS), or blind signal separation, techniques. The term “blind” indicates that there is no information or very limited amount of information available about the original source signals or the underlying mixing process. These techniques aim ideally to pure extraction of the original signals. Assumptions of mutual independence, uncorrelation or orthogonality between signals need to be made usually.

Principal component analysis (PCA) or Karhunen-Loève transform (KLT) is a statistical method that can be used for splitting the input signals in orthogonal signal components. Its objective is to reduce the dimensionality of a data set consisting of variables or signals that are related to each others. The reduction operation is wanted to preserve the variance of the original data set as much as possible. PCA looks for the linear combination of the original variables having maximum variance. Next, it finds the variance-maximizing linear combination that is orthogonal to the first linear combination. The variance of the second linear combination is, thus, smaller than the variance of the first one. This procedure can be continued until the number of orthogonal linear combinations equals to the dimensionality of the original data set. The new signals achieved by the linear combinations are called the principal components (PCs). In dimensionality reduction, it is desirable that first few principal components already contain the most of the variance that is present in the original data set. [29]

As it finds the dominant signal component amongst two or more signals, PCA has its applications in source separation problems. In audio processing, it can be used, for example, to separate the primary source signal and the ambience signals from a multichannel audio signal [8], [22], to remove interchannel redundancy from multichannel signals for better audio coding performance [52] or to upmix stereo signals [27].

PCA is calculated by forming the covariance matrix $\mathbf{R}$ of the data samples in matrix $\mathbf{X}$. The entries of the covariance matrix are the covariances between the data vectors $\mathbf{x}_n$. The values of the covariance matrix are given by

$$R_{i,j} = E[(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)]. \tag{2.19}$$

The principal components are calculated from the covariance matrix using the eigenvalue decomposition. The eigenvector $\mathbf{v}$ must satisfy the linear equation

$$\mathbf{R}\mathbf{v} = \lambda\mathbf{v}, \tag{2.20}$$

where λ is the eigenvalue that corresponds the eigenvector $\mathbf{v}$. The eigenvectors form an orthogonal vector base for the the principal components. The eigenvalues correspond to the variances of the principal components. The eigenvector corresponding to the largest eigenvalue is therefore the linear combination that forms the first principal component. Similarly, the eigenvector with the second largest eigenvalue can be used to form the second principal component and so on. PCA is a reversible transform. The principal components can be transformed back to the original signals by using the same eigenvectors, although scaling by the eigenvalue will be needed to preserve the original signal energy. No further scaling is required if the eigenvectors are initially scaled so that their second vector norm is one. The scaling changes also the eigenvalues, but their order is not changed, and the actual eigenvalues are not needed in the transform process anyway.

Independent component analysis (ICA) [9] is a group of blind source separation techniques, which find independent nongaussian signals from the mixture. The number of observations, or channels in the mixture, is required to be greater than or equal to the number of signals to be separated. The latter requirement is common with principal component analysis, but independence and nongaussianity requirements of the source signals are different from PCA, which is based on orthogonality of the sources. ICA forms the independent components as linear combinations of the original data, similar to PCA. Besides uncorrelatedness, ICA requires nonlinear independence between the original source signals. Using the covariance matrix to form the linear combinations would allow decorrelating the signals, which is not enough. The ICA methods, thus, use some sort of "higher order statistics". [25]

The PCA and ICA methods estimate the original source signals as linear combinations of the mixture signal. The number of source signals cannot be larger than the number of signal channels or variables. This can be a bad restriction for the analysis of stereo or multichannel audio signals. It is very probable that a stereo signal has more than two underlying sources in the mixture or that a 5-channel signal has been mixed using more than five sources. Additinal source separation methods are required for recovering signals from this kind of underdetermined cases. Further presumptions about the signals need to be introduced then. Sparse component analysis (SCA) [15], [32], [33], [53] or sparse principal component analysis (SPCA) [54] assume that signals have sparse representations. For audio signals, the sparseness means that there are discrete sources that are the only active sources at some time instances or frequency bands. An example of a sparse representation is a stereo recording that contains a conversation of more than two persons, where the speech of different persons does not overlap significantly. If the mix has been recorded with a system of two directional microphones, location information of the speakers can be estimated from the instants where only one participant of the discussion is active.

The sparseness property makes it possible to estimate the underlying mixing coefficients for underdetermined cases, where the number of mixtures is less than the number of sources. Knowing the mixing coefficients, however, does not directly give a solution for the source separation problem, because there are an infinite number of solutions in underdetermined cases. Figure 2.2 illustrates how the mixing coefficients can be obtained from two signals that are mutually sparse in the frequency domain. The signal mixtures $X_1(f)$ and $X_2(f)$ contain two common amplitude-panned source components. Both mixture signals have also independent signal components. All source signals are harmonic tones which don't overlap with each others in the frequency domain. The four different signal components can be seen as four linear relationships between the Fourier transform bins of the mixture signals $X_1(f)$ and $X_2(f)$. This case is of course a very ideal example of sparse components and the sparseness is seldom at this level in reality. However, even partial sparseness can be exploited in estimating the mixing coefficients [33].

2.6 Conclusion

This chapter discussed the spatial audio analysis. The spatial properties of multichannel audio need to be analyzed first in order to preserve them as much as possible after the format conversion, which is the objective of this study. The spatial hearing of the human auditory system was chosen as the initial topic because our hearing determines the directions where we localize sound events and how we sense the spaciousness from sound. These are very relevant sensations when the spatial audio reproduction is considered. The perceptual cues are derived from the sound field that our ears receive. This sound field is initiated by the loudspeaker signals in loudspeaker listening. Correspondingly, the spatial analysis for the modification of multichannel audio is more signal-oriented. The common production methods of such signals were briefly discussed to give the reader the crucial basic background

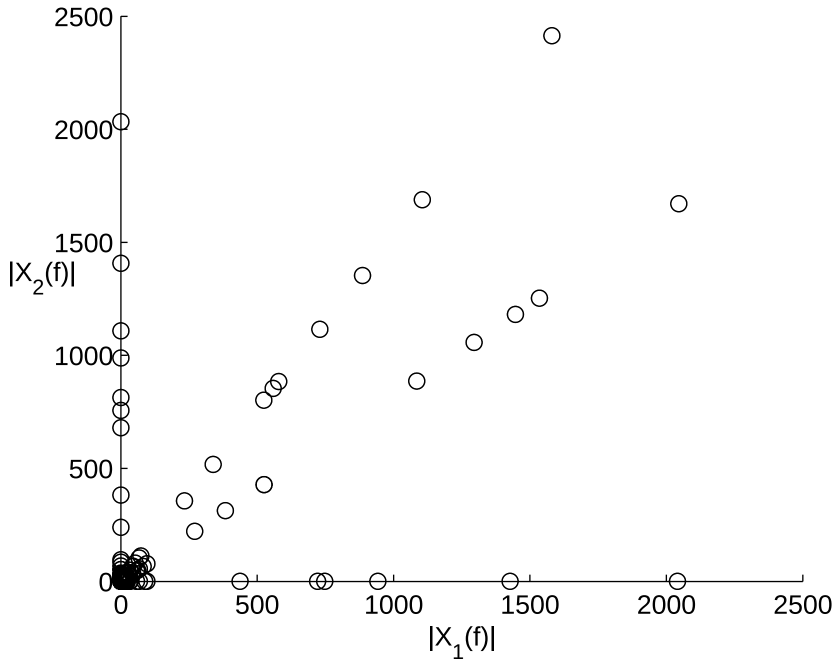


Figure 2.2: Fourier transform bins of two mutually sparse signals plotted as points $(|X_1(f)|, |X_2(f)|)$. The four different sparse signals can be observed as four linear relationships of the points.

knowledge about the information that the signals contain before moving on to the actual analysis of the audio signals.

Multichannel audio coding techniques were mentioned as a specifically developed topic that benefits from the spatial analysis of the audio signals. The coding techniques aim at reducing the amount of overlapping information from the audio signal channels. The interchannel similarity measures are essential in such techniques, and they are important in multichannel audio transformations as well. Different similarity measures were described, and specially their computational aspects were opened thoroughly. Knowing the significant properties and the computational efficiency of the measures is indeed important in later examination of spatial transform methods. Signal-based estimates for the apparent directions of audio sources were then presented. As is for the interchannel similarity measures, these estimates are used in spatial audio coding techniques.

Finally, blind source separation techniques of information technology were discussed as tools that aim at ideal decomposition of signal mixtures to the original source signals. A common limitation in the BSS techniques is that the amount of mixture signals is required to be equal to or more than the amount of underlying source signals. Other limitations are the heavy amount of computational power needed, and to some extent, the idealist approach of perfect separation of the original sources. The restriction of perfect separation could be relaxed for spatial audio transformations, and hopefully it could allow separation of more source signals than the number of mixture signals, although the restriction can be somewhat avoided in audio processing by dividing the signals in sub-bands and analyzing them one by one. Lately, research interest has risen on the source separation techniques that rely on sparse representations. The sparseness means that the original signals are non-overlapping in some domain. Even though the sparseness assumption would not be always valid for average audio signals, the SCA techniques are a prominent future research topic for the purpose of audio decomposition. The focus on this study, however, was directed to the analysis and decomposition approaches that have been developed for the audio coding techniques, as they are already at a quite mature stage.

Chapter 3

Multichannel audio transform techniques

This chapter describes four different techniques and algorithms that are intended for converting multichannel audio from a format to another. The algorithms aim at preserving the perceivable spatial properties of the sound image. The block diagram in Figure 1.1 (page 2) showed a general form for this type of techniques: the audio channels are analyzed in the analysis block, and the extracted parameters are used to control the tranformation, which exploits information about the new reproduction system.

In multichannel audio coding techniques, it is important to remove redundant information from the audio channels. Many coding techniques calculate the spatial analysis of the original signals and then transmit only reduced amount of information, such as the compressed sum signal of the original signals, and the spatial side information yielded by the analysis. In the synthesis phase, the spatial properties of the signal can be reconstructed using the transmitted analysis information [7]. The compression of signal data is not an objective in multichannel transform techniques, but the spatial analysis information is very beneficial for this purpose. Multichannel audio coding is, therefore, a strongly linking factor between the techniques that will be discussed here. The first method uses correlation analysis similar to that of MPEG audio coding. Two of the other techniques are directly called as coding techniques by their authors, whereas the last one is often referred to as a preprocessing step for multichannel audio coding.

The organization of the chapter is as follows: First, the general introduction to the techniques will be given. Then the techniques are compared in functional blocks, which are similar to those shown in Figure 1.1. Finally, the applicability of the techniques for multichannel audio transforming is discussed.

3.1 General introduction

A potential transform technique for stereo signals was proposed by Faller [14] (the text was later reproduced with some additions by Breebaart and Faller [7]). The model used in this technique is inspired by a situation where there is a single sound source in the room causing reflections from the side walls. Therefore, the stereo signal is modeled by a single amplitude panned signal and lateral channel-specific signals:

$$\begin{aligned} x_1(k) &= s(k) + n_1(k) \\ x_2(k) &= as(k) + n_2(k) \end{aligned} \tag{3.1}$$

The source signal $s(k)$ corresponds to the dry sound source and the side signals $n_1(k)$ and $n_2(k)$ estimate the reflections. All three signals are defined to be independent. The source signal $s(k)$ is multiplied by the amplitude panning coefficients 1 and a, which determine the direction of the source between the two stereo channels. The model assumes that the lateral signals have the same power. This is the critical assumption that enables estimation of the decomposition signal components and amplitude panning coefficients in Equation (3.1). Multiple concurrent sources are allowed by dividing the stereo signal in frequency bands and applying the model individually to each band. The estimated signal components can be used to convert the stereo signal to be played from other reproduction systems, for example a loudspeaker line array or a traditional 5.1 surround sound system.

Goodwin and Jot [19], [20], [21] presented a multichannel audio format conversion technique that estimates the perceived direction of the sound at the sweet spot. The technique can be exploited in multichannel audio coding, in which context the authors call it universal spatial audio coding. The estimation of directions is done separately for different sub-bands. The directions are calculated using a localization vector similar to the signal power-based vector of Equation (2.18) on page 15. In this technique, however, the signal power values have been normalized so that the sum of the energies over all the signal channels is one on each sub-band. The normalized power-based localization vector is, thus, given by

$$\vec{g} = \mathbf{C} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix} \frac{1}{\sum_{l=1}^{m} p_l}, \tag{3.2}$$

where $\mathbf{C}$ is the channel format matrix as the one in Equation (2.16). The subscripts of the powers p are the channel indices and m is the number of signal channels.

Normally, when the locations of the loudspeakers are equidistant from the listener, the format matrix consists of unit vectors. If normalized power values are used in the calculations, the length of the power vector is never greater than one. The length will be one if and only if all of the signal power is concentrated on a single channel, which means that all the sound is played from one loudspeaker. The direction of the power vector will be pointing towards the corresponding loudspeaker in this case. The length of the power vector is always less than one if there are more than two active signal channels. Let us consider a two-channel stereo system with channel directions of $\pm 30^\circ$ as an example. If there is equal amount of signal power in both channels, their relational powers are 0.5 each. The localization vector will be then

$$\begin{aligned}\vec{g} &= \begin{bmatrix} \vec{c}_1 & \vec{c}_2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \\ &= \begin{bmatrix} \sin(-30^\circ) & \sin(30^\circ) \\ \cos(-30^\circ) & \cos(30^\circ) \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0.866 \end{bmatrix}. \end{aligned} \tag{3.3}$$

There is only y-component in the direction vector, and it points exactly to the center point between the stereo channels. The length of the vector is less than one. If the power proportions of the channels are changed, the vector will again point to the center line between the terminals of the format vectors, but its direction will change. The line connecting the terminals is illustrated in the left side part of Figure 3.1. The normalized power vector of the two-channel stereo will always point to this line. More generally, if there are only two active signal channels, the power vector will always point to the line connecting the terminals of the format vectors of these channels. The right-side part of Figure 3.1 shows the lines connecting the adjacent format vectors of a 5-channel surround system. Goodwin and Jot note that these lines give the maximum lengths for each direction of the normalized power vector. These lengths can be reached only when there are exactly two active signal channels, and the channels have adjacent format vectors. In Goodwin and Jot's method, the format vectors with maximum lengths depict the case of completely directional sound, which is modeled as coming from a single point-like source. The model interprets that the sound is at least partially non-directional, when the localization vector cannot reach the connecting lines. In these cases, there are more than two active channels or exactly two active channels, that are non-adjacent. Finally, null vector as a localization vector represents completely non-directional sound. The above-mentioned properties allow the lengths of the power vectors to be used for measuring the proportions of the directional and non-directional sound. The ratio between the actual vector length and the maximum vector length depicts the power relation between the directional sound and overall sound. The sound direction and directionality information can be then exploited to render the multichannel audio to a new reproduction format.

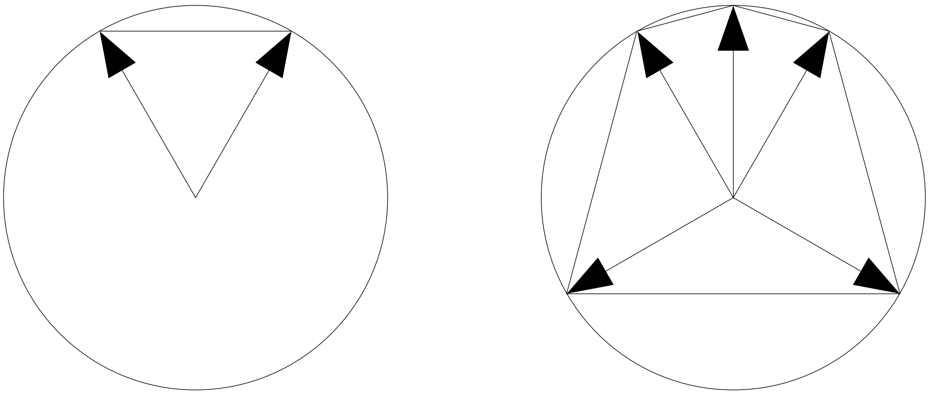


Figure 3.1: The power vector maximum lengths and format vectors of the two-channel stereo and the five-channel surround.

Directional audio coding (DirAC) [42] is a technique intended for spatial audio coding and flexible audio reproduction with different playback systems. It can be also applied to stereo upmixing. DirAC divides the sound signals in frequency bands and then analyzes the diffuseness of the sound field and the direction of arriving sound on each band at different time instances. Both STFT and filterbank-based time-frequency decomposition can be used. DirAC is based on Spatial Impulse Response Rendering (SIRR) [35], which calculates the sound direction information and the diffuseness of the sound from the physical magnitudes of the sound field. The primary magnitude is the instantaneous sound velocity, which is further used to calculate the sound intensity and the sound energy density at the locations of the microphones or at the listening point. The directionality of the arriving sound can be obtained from the intensity and energy density measures. The input sound can be then divided to non-diffuse (directional) and diffuse (non-directional) parts. Finally the sound is resynthesized as a combination of diffuse and non-diffuse streams for the selected reproduction system. The main focus in DirAC has been presently on using Ambisonic [16] B-format microphone recordings as input signals, but the analysis of stereo signals has been considered as well. [43], [41], [42]

Principal component analysis (see Section 2.5) is not exclusively an audio signal format conversion method, but it can be used to extract primary and ambience signals from stereo [34] and multichannel [22] recordings. PCA is also used to remove interchannel redundancy for audio coding purposes [52]. The original multichannel signal $\mathbf{x}(k)$ can be multiplied by the principal component vector $\mathbf{v}$, which gives the mono signal of the primary beam:

$$\begin{aligned} s(k) &= \mathbf{v}^T\mathbf{x}(k) && (3.4) \\ &= [v_1, v_2, \ldots, v_m] \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_m(k) \end{bmatrix} && (3.5) \end{aligned}$$

The calculation of the primary component vector itself was being discussed in Section 2.5. The primary beam can be then mixed for the new reproduction system, and extracted from the original signals using simple subtraction. The one-dimensional primary component is subtracted from each original signal using the primary component vector as a weight.

$$\mathbf{n}(k) = \mathbf{x}(k) - \mathbf{v}\mathbf{s}(k) \tag{3.6}$$

The results of the subtraction operation are the ambience signals $\mathbf{n}(k)$. The analysis of principal components and the extraction of the primary signal can be done separately on sub-bands either in the time-domain or after the Fourier transform. Goodwin and Jot [22] have also proposed that PCA-based primary-ambient decomposition could be used as a preprocessing method for their multichannel format conversion technique.

3.2 Analysis phase

All four techniques that were mentioned previously in this chapter divide the signal in sub-bands and analyse each sub-band separately. It has been assumed in the techniques that the human auditory system can distinguish only the direction of one dominant sound source on each critical band. The sub-bands mimic the frequency resolution of the human ear in one way or the other. Faller's method uses frequency bands equal to auditory critical bands [14], whereas the frequency bands of Pulkki's DirAC implementation are two ERBs each [41]. Short-time Fourier transform (STFT) or filterbank based time-frequency decompositions are both possible in all the methods. STFT is usually preferred for real-time applications because of its lower computational complexity [14], [42] and flexible possibilities to change the frequency bands adaptively [20]. The strength of the filterbank-based design is that it allows different bands to be analyzed with different time-frequency resolutions. In STFT designs, all the bands share same time-frequency resolutions, which is one of their weaknesses. It is often desirable to have better frequency resolution at lower bands than at higher bands, because the resolution of the human auditory system is also better at the low frequencies.

Faller's method estimates the common source signal and its amplitude panning coefficient exploiting the signal powers and interchannel coherence. First, equations of the signal powers and ICC are formed using Equation (3.1):

$$\begin{aligned} p_{x_1} &= p_s + p_{n_1} && (3.7)\\ p_{x_2} &= a^2 p_s + p_{n_2} && (3.8)\\ \Phi &= \frac{a p_s}{\sqrt{p_{x_1} p_{x_2}}} && (3.9) \end{aligned}$$

Above, the signal powers and ICC are denoted by p and Φ, respectively. An important definition in the decomposition model of Equation (3.1) was that the decomposed signals $s(k)$, $n_1(k)$ and $n_2(k)$ are independent and thus uncorrelated. This definition makes the powers of the stereo signals $x_1(k)$ and $x_2(k)$ equal to the summed powers of the independent signal components in Equations (3.7) and (3.8). Equation (3.9) exploits the uncorrelatedness as well, because the lateral signals $n_1(k)$ and $n_2(k)$ contribute to the correlation coefficient only through their powers, which are present in the normalization term.

Equations (3.7), (3.8) and (3.9) have more unknown than known variables, and therefore it is assumed that the powers of the independent lateral signals (p_{n_1} and p_{n_2}) are equal. They can be marked by $p_n = p_{n_1} = p_{n_2}$. Now, the three remaining unknown variables (p_s, p_n and a) can be solved. This was already done by Faller in [14] and [7], but without any intermediate results. For the sake of clarity, a thorough solution is presented below. It starts from writing Eqs. (3.7), (3.8) and (3.9) in new forms:

$$\begin{aligned} p_n &= p_{x_1} - p_s && (3.10)\\ p_s &= \frac{p_{x_2} - p_n}{a^2} && (3.11)\\ a &= \frac{\Phi\sqrt{p_{x_1} p_{x_2}}}{p_s} && (3.12) \end{aligned}$$

The power of the common signal p_s can be solved by substituting p_n in Eq. (3.11) with the right hand side of Eq. (3.10):

$$\begin{aligned} p_s &= \frac{p_{x_2} - p_{x_1} + p_s}{a^2} && (3.13)\\ p_s &= \frac{p_{x_2} - p_{x_1}}{a^2 - 1} && (3.14) \end{aligned}$$

Then a can be substituted from Eq. (3.14) by the right hand side of Eq. (3.12):

$$\begin{aligned}
p_s &= \frac{p_{x_2} - p_{x_1}}{\left(\frac{\Phi\sqrt{p_{x_1}p_{x_2}}}{p_s}\right)^2 - 1} \\
p_s\left(\frac{\Phi^2 p_{x_1}p_{x_2}}{p_s^2} - 1\right) &= p_{x_2} - p_{x_1} \\
\frac{\Phi^2 p_{x_1}p_{x_2}}{p_s} - p_s &= p_{x_2} - p_{x_1} \\
\Phi^2 p_{x_1}p_{x_2} - p_s^2 &= p_s(p_{x_2} - p_{x_1}) \\
p_s^2 + (p_{x_2} - p_{x_1})p_s - \Phi^2 p_{x_1}p_{x_2} &= 0
\end{aligned} \tag{3.15}$$

The quadratic Equation (3.15) can be solved using the well-known quadratic formula:

$$p_s = \frac{-(p_{x_2} - p_{x_1}) \pm \sqrt{(p_{x_2} - p_{x_1})^2 + 4\Phi^2 p_{x_1}p_{x_2}}}{2} \tag{3.16}$$

All the signal powers are positive by definition. Power of the common source, p_s, has to be positive as well, and therefore

$$|p_{x_2} - p_{x_1}| \leq \sqrt{(p_{x_2} - p_{x_1})^2 + 4\Phi^2 p_{x_1}p_{x_2}}. \tag{3.17}$$

Using the property of Equation (3.17), the minus sign before the square root in the numerator of Equation (3.16)can be omitted:

$$\begin{aligned}
p_s &= \frac{p_{x_1} - p_{x_2} + \sqrt{(p_{x_2} - p_{x_1})^2 + 4\Phi^2 p_{x_1}p_{x_2}}}{2} \\
&= \frac{2\Phi^2 p_{x_1}p_{x_2}}{p_{x_2} - p_{x_1} + \sqrt{(p_{x_1} - p_{x_2})^2 + 4\Phi^2 p_{x_1}p_{x_2}}}
\end{aligned} \tag{3.18}$$

The final form of Eq. (3.18) is the same as the one given by Faller in [14] and [7]. Now the amplitude panning coefficient a can be solved using Equations (3.12) and (3.18):

$$\begin{aligned}
a &= \frac{\Phi\sqrt{p_{x_1}p_{x_2}}}{\frac{2\Phi^2 p_{x_1}p_{x_2}}{p_{x_2} - p_{x_1} + \sqrt{(p_{x_1} - p_{x_2})^2 + 4\Phi^2 p_{x_1}p_{x_2}}}} \\
&= \frac{p_{x_2} - p_{x_1} + \sqrt{(p_{x_1} - p_{x_2})^2 + 4\Phi^2 p_{x_1}p_{x_2}}}{2\Phi p_{x_1}p_{x_2}}
\end{aligned} \tag{3.19}$$

Finally p_n can be calculated from Eq. (3.10) using the measured value of p_{x_1} and the solved p_s.

The estimated power values p_s and p_n, and the amplitude panning coefficient a can be further used to estimate the modeled signals $s(k)$, $n_1(k)$ and $n_2(k)$ from the original signals

$x_1(k)$ and $x_2(k)$. The signal estimates are simply linear combinations of the original stereo signals. For example, the estimate $\hat{s}(k)$ of the amplitude panned signal $s(k)$ is given by

$$\hat{s}(k) = w_1 x_1(k) + w_2 x_2(k), \tag{3.20}$$

where w_1 and w_2 are real-valued weights. Similar equations can be written for the estimates of the independent lateral signals:

$$\hat{n}_1(k) = w_3 x_1(k) + w_4 x_2(k) \tag{3.21}$$
$$\hat{n}_2(k) = w_5 x_1(k) + w_6 x_2(k) \tag{3.22}$$

The weights are solved by minimizing the error signals

$$e_s(k) = \hat{s}(k) - s(k) \tag{3.23}$$
$$e_{n_1}(k) = \hat{n}_1(k) - n_1(k) \tag{3.24}$$
$$e_{n_2}(k) = \hat{n}_2(k) - n_2(k) \tag{3.25}$$

in least squares sense. The minimization will be shown here for $e_s(k)$ as an example. First, Equation (3.20) is written in a new form using Equation (3.1)

$$\hat{s} = w_1\big(s(k) + n_1(k)\big) + w_2\big(as(k) + n_2(k)\big), \tag{3.26}$$

which yields in combination with Equation (3.23)

$$e_s(k) = \big(w_1 + aw_2 - 1\big)s(k) + w_1 n_1(k) + w_2 n_2(k). \tag{3.27}$$

The optimal weights have been found when $e_s(k)$ is orthogonal to $x_1(k)$ and $x_2(k)$ [14]. Orthogonality holds true statistically when the signals are uncorrelated [38]:

$$E[x_1(k)e_s(k)] = 0 \tag{3.28}$$
$$E[x_2(k)e_s(k)] = 0 \tag{3.29}$$

Using Equations (3.1) and (3.27) and the assumed independence of signals $s(k)$, $n_1(k)$ and $n_2(k)$, Equations (3.28) and (3.29) yield

$$\big(w_1 + aw_2 - 1\big)P_s + w_1 P_{n_1} = 0 \tag{3.30}$$
$$a\big(w_1 + aw_2 - 1\big)P_s + w_2 P_{n_2} = 0, \tag{3.31}$$

from which the weights are solved

$$w_1 = \frac{P_s P_n}{(a^2+1)\,P_s P_n + P_n^2} \tag{3.32}$$
$$w_2 = \frac{aP_s P_n}{(a^2+1)\,P_s P_n + P_n^2}. \tag{3.33}$$

The rest of the weights are estimated similarly. Finally, post-scaling is applied to the estimated signals to ensure that the powers of the estimated signals equal to the previously solved signal powers p_s and $p_n = p_{n_1} = p_{n_2}$. Faller [14] gives the exact formulas for the signal weights w_3-w_6 and the post-scaling factors.

The spatial signal analysis of Faller's method is based on signal power measures and interchannel coherence, whereas Goodwin and Jot's method exploits solely the signal power measures in the analysis of the localization directions. Goodwin and Jot's method calculates the relative powers of each signal channel and uses them as weights for the format vectors of the loudspeaker channels. The sum of the weighted format vectors gives the power-based localization vector. Its angle gives the apparent localization direction of the signal, if the audio is played back from the appropriate reproduction system given by the channel format vectors. This kind of localization vector is also called the Gerzon vector (see Sec. 2.4).

Goodwin and Jot's model calculates the power proportions between the directional and non-directional parts of the multichannel audio signal. The directional sound is modeled as a single point-like source, which has been amplitude panned between two signal channels that have neighboring format vectors. The length of the power-based localization vector can be used to measure the relational signal powers of the directional and non-directional sound. Figure 3.1 shows the maximum lengths of the power vector in the case when only two neighboring signal channels are active. If the length of the vector does not reach the maximum value of the respective direction, there is non-directional sound present in the signals. The sound directivity ratio r can be obtained from

$$r = \left\| \begin{bmatrix} \vec{c}_\alpha & \vec{c}_\beta \end{bmatrix}^{-1} \vec{g} \right\|_1 , \tag{3.34}$$

where $\vec{c}_\alpha$ and $\vec{c}_\beta$ are the format vectors that are directly neighbouring the localization vector $\vec{g}$, and $\| \, \|_1$ is the vector 1-norm [20]. The directionality ratio gives the power ratio of the directional sound in comparison to the overall sound. The signal power ratio of the non-directional sound is then $1-r$. Unlike in Faller's method, the directional and non-directional audio signals themselves are not extracted from the original signals in the analysis phase of Goodwin and Jot's method.

Goodwin and Jot's method approximates the power-based localization vetor at the optimal listening position. Directional audio coding estimates the properties of sound field in the same position. The analysis of DirAC was originally applied to B-format recordings, which ideally capture particle velocities and sound pressure in the location of the microphone [35]. The DirAC analysis is based on these physical magnitudes, but it can be applied to stereo or multichannel recordings if particle velocity and sound pressure are first estimated. In this kind of signal-based approach, the particle velocity vector can be approximated by taking the same format vectors that were used in Goodwin and Jot's method and scaling them with the signal values. This kind of velocity vector is given by Equation (2.17). It should be noted that the format vectors are weighted by plain signal values in DirAC, while signal power values are used in Goodwin and Jot's model. The sound pressure p can be approximated by summing the signal channels

$$p = \sum_{n=1}^{N_1} x_n. \tag{3.35}$$

DirAC calculates two important magnitudes from the velocity vector: the instantaneous sound intensity vector and the instantaneous sound energy density. The sound intensity vector can be calculated from the sound pressure and the particle velocity vector $\vec{u}$ using the formula

$$\vec{I} = p\vec{u}. \tag{3.36}$$

The formula for the sound energy density is

$$E = p^2 + \|\vec{u}\|^2, \tag{3.37}$$

where $\| \, \|$ is the Euclidean norm. Equation (3.37) differs a bit from the real physical formula. It omits the fluid density and wave impedance values, because they vanish when B-format microphone signals are used [35]. The direction of the intensity vector can be used to indicate the direction of the arriving sound. The main feature behind DirAC is that the diffuseness of the sound field can be measured by comparing the sound intensity and the energy density. If the length of the intensity vector is small, the sound energy arrives from many directions and is almost diffuse. If the intensity is closer to the energy density value, the sound is non-diffuse. The diffuseness ratio can be calculated from the instantaneous intensity and the sound energy density. It is given by

$$\begin{aligned} \psi &= 1 - \frac{2\|\vec{I}\|}{E} \\ &= 1 - \frac{2|p|\|\vec{u}\|}{p^2 + \|\vec{u}\|^2}. \end{aligned} \tag{3.38}$$

Equation (3.38) omits again some physical magnitudes that are not apparent in audio signals. The normalization factor 2 in the numerator makes sure that the diffuseness ratio is between 0 and 1. This is easily proven:

$$(|a| - |b|)^2 \geq 0 \tag{3.39}$$

$$|a|^2 + |b|^2 - 2|a||b| \geq 0 \tag{3.40}$$

$$\frac{2|a||b|}{|a|^2 + |b|^2} \leq 1 \tag{3.41}$$

It is important to point out that Equation (3.38) is in a little different form in [42], where the omnidirectional B-format signal W is used directly as the pressure value, In DirAC, it has been presumed that the signal W has been scaled down by 3 dB, thus divided by $\sqrt{2}$. This scaling is used in Ambisonics B-format signals [12].

The PCA-based primary-ambient decomposition starts from calculating the eigenvector decomposition of the covariance matrix of the signal. Noteworthy, the covariance matrix consists of signal correlations and power values. The eigenvector corresponding to the largest eigenvalue is called the primary component vector. It is used as a weight for the signal channels. Multiplication by the primary component vector forms the signal beam with the most energy. This is the primary beam of the signals. The primary beam can be subtracted from the original multichannel signal. After the subtraction, the resulting multichannel signal is the ambience signal. The primary component vector contains the direction information of the primary beam. This information is given for the signal space, which has as many dimensions as there are signal channels. Each signal represents one dimension. If needed, the direction information can be derived using the format vectors to be used in the context of a multichannel reproduction system. As the values of the primary component vector are simply used as the signal channel weights, the direction vector of the primary beam can be calculated by weighting the format vectors with the values of the principal component vector. This kind of direction information may be important in the transform phase.

The four above-mentioned methods can be used to transform stereo or multichannel signals to a different reproduction format. The only method that is mentioned only applicable to stereo signals is Faller's method. That is because the important hypothesis of the method is that the independent lateral signals, or the ambience signals, contain equal amount of energy. This hypothesis is difficult to expand to multichannel systems. Meanwhile, the method by Goodwin and Jot is quite fruitless to be used for directionality analysis of two-channel stereo signals, since it cannot distinguish any diffuseness if there are only two signal channels. It merely calculates the power-based localization vector for each sub-band and assumes that all the signal power is arriving from point-like sound sources in the directions indicated by the vectors of the sub-bands. PCA and DirAC can analyze audio content with two or more channels.

The analysis phases of the methods are based on various signal measures. The measures used by Faller's method and the PCA method are the most similar. Both calculate the signal powers and the correlations and then extract primary and ambience signals from the original signals by using these measures. Faller does this by making the assumption of the equal powers of the ambience signals, and then estimating the signals. The signal estimates are simply linear comibinations of the original signals, and the weights are calculated with the minimum squared error criterion. The PCA method does not make further assumptions about the primary or the ambience signals. It calculates the eigenvalue decomposition and uses the primary component vector to form the primary beam for the multichannel signal. The ambience signal is formed in both methods by subtracting the primary signal component from the original signal. This makes them very closely related, and the signal power assumption of Faller's method seems to be the biggest basic difference between the two methods.

All the four methods presented in this chapter can be considered psychoacoustically motivated in the sense that they divide the signal in frequency bands that mimic the frequency selectivity of the human auditory system. They also assume that only one primary direction of arrival can be heard for the sound of each band. Except for this assumption, the analysis approaches of Faller's method and the PCA method seem totally signal-oriented, because they concentrate on examining the signal channels, but not the auditory scene perceived by the listener. Goodwin and Jot's method and DirAC, by contrast, exploit the information of the loudspeaker system to analyze the directions and the directionality of the arriving sound at the listening point. Goodwin and Jot's method calculates the direction of arrival as well as the directionality from the powers of the signals. DirAC even forgets the original signal space by approximating the physical magnitudes of the sound field at the optimal listening point. Goodwin and Jot's method and DirAC analyze only the apparent directions of sound and the diffuseness on each sub-band, but do not separate the actual diffuse sound signal from the original signals. Goodwin and Jot have proposed also applying their method to the signals given by the PCA-based the primary-ambient signal decomposition [22]. This could give good results with DirAC analysis as well, and it is a prominent topic for future studies.

3.3 Transform phase

The previous section presented the sound direction and directionality analysis of four audio format conversion techniques. All the techniques evaluate the portions of the directional and non-directional sound from total sound energy, although in DirAC these are called non-diffuse and diffuse sound, respectively. Faller's method and the PCA method both extract the primary signal component and the ambience signal from the multichannel mix, whereas Goodwin and Jot's method and DirAC estimate only the direction of the primary source and the diffuseness of the sound. If diffuse sound is required to be played back in the latter two methods, a special sound diffusor block will be needed. This can be a decorrelating filter, which can be realized by convolving the sound with certain type of noise in the time domain or filtering with a random phase all-pass filter in the frequency domain, for example.

Before the transform can be conducted, it is important to decide how the direction values of the original reproduction format project to the new format. This includes choosing the panning mechanisms. Similar decisions need to be made for how to play the diffuse sound in the new system.

Faller considers in detail some specific reproduction formats for his method. The first one has a loudspeaker array in front of the listener, while the second format has the same loudspeaker array in front of the listener and additional side loudspeakers exactly on the both sides of the listener. The aim of the loudspeaker array is to widen the sound image while having more precise localization selectivity for the sound events arrivign from different directions. The third reproduction system is the conventional 5-channel surround system. For the loudspeaker arrays he proposes that the sound directions of the primary sources are scaled linearly so that the maximum angles ($\pm 30^\circ$) of the stereo system map to the loudspeakers that have widest direction angles from the view of the listener. Then the primary source is simply amplitude panned to the correct location between the nearest two loudspeakers. In the first case, where there are no side loudspeakers, the lateral independent sound should be played back from the loudspeakers which are furthest away from the listener. When the system contains two additional speakers on the sides of the listener, the lateral sound should be played from these loudspeakers, and the loudspeaker array in front of the listener should be used solely for playing the primary sources. For the upmixing process of converting the original stereo sound to the 5-channel format, Faller proposes that the lateral signals are played from the rear loudspeakers and the primary sources are panned between the two of three frontal loudspeakers, which are chosen according to the source location [14] [7]. The PCA method extracts specific primary and ambience signals, which are similar to those of Faller's method. The transform phase can therefore greatly resemble that described for Faller's method. The PCA method, however, is also applicable to multichannel signals, unlike Faller's method, which was intended only for transforming stereo signals.

Goodwin and Jot's method, as described in the original papers, does not extract specific non-directional and directional sound signals. The model bases on the localization vectors calculated from the channel-specific signal powers. A question raises concerning what is the actual sound content that will be played from the new playback system. As there are not any additional signals that are extracted from the original sound signals, the original signals have to be used in some way. Goodwin and Jot suggest various ways to fill the new loudspeaker channels with different signals derived from the original multichannel signal. Then the directionality information should be used to calculate channel-specific gain factors. For audio coding purposes, they propose that every loudspeaker plays a mono downmix of the original signal to reduce the transmitting bitrate. The spatial cues of the original multichannel signal will be preserved better if the the new signal channels contain only the audio signals of the original signal channels that are located closest to them. This can be used for the audio format conversion since bitrates are not an issue.

The gain factors of the new signal channels are calculated separately for the directional and non-directional sound and then summed. For the directional sound, only the two channels that are directly neighboring the localization vector, will have non-zero weights. This originates from the fact that the directional sound is modeled as coming from a single point-like and amplitude panned source. The weights can be calculated from

$$\begin{bmatrix} \sigma_\alpha \\ \sigma_\beta \end{bmatrix} = \begin{bmatrix} \vec{c}_\alpha & \vec{c}_\beta \end{bmatrix}^{-1} \vec{g}, \tag{3.42}$$

where $\vec{c}_\alpha$ and $\vec{c}_\beta$ are the format vectors and σ_α and σ_β are the power gain factors of the directional sound for the closest channels α and β, respectively. The gain factors should be normalized so that $\sigma_\alpha + \sigma_\beta = 1$, which is the condition for energy-preserving. A different approach is used for the weights of the non-directional part. The completely non-directional sound was described to have a null vector as the localization vector. This property should remain also for the non-directional part of the transformed signal. For this purpose, a special null-weight vector $\vec{\delta}$ is calculated. The term null-weight vector comes from its property that the channel format matrix $\mathbf{C}$ multiplied by it produces a null vector

$$\mathbf{C}\vec{\delta} = \vec{0}. \tag{3.43}$$

It is desirable that the values of the null-weights for different signal channels are non-zero and equal each others as much as possible. This way the situation that the non-directional sound is played back only from two loudspeakers which are opposite to each others can be avoided, for example. This specification leads to the best synthesis results as the non-directional sound is played back evenly from all the possible directions. Proper null-weights can be derived by minimizing a cost function and using the method of Lagrange multipliers [48]. This method was chosen by Goodwin and Jot [20]. The null-weight approach is valid as is only in reproduction systems where the loudspeakers are facing directly to the same listening position and are equidistant from this position.

DirAC has different application-specific approaches for implementing the transform. The choice of the realization depends on the type of the original signals as well as the loudspeaker system that is used for the reproduction. The DirAC implementation for B-format signals extracts the loudspeaker signals from the sound field information by using virtual cardioid microphones pointing at the loudspeaker directions. In DirAC for 2-to-5 upmixing, which is more interesting application considering this work, a sum signal of the original stereo signal can be played from a direction given by the directionality analysis for the non-diffuse sound. The directions need some temporal averaging to avoid audible artefacts. This can be done by calculating gain factors for the new directions and then adapting the gain factors, but not the actual directions, as proposed by Pulkki. The left stereo signal is used for the diffuse reproduction of the left hemisphere, and, in proportion, the right stereo signal is used for the diffuse sound of the right hemisphere. To make the sound really diffuse, the different loudspeaker signals can be decorrelated. Pulkki proposes convolving the diffuse loudspeaker signals with short bursts of exponentially decaying white noise, although also other methods can be used. [41] [42]

STFT-based stereo-to-three-channel implementations were made for all four transform techniques. The implementations had an interval of 23 ms between the FFT frames, and there was 50% overlap for consecutive frames. The transform phase proved to be generally difficult in informal listening experiments. A common problem was that the analyzed directions changed a lot from frame to frame, which caused audible artefacts, because the direction of the primary point-like sound component was changing rapidly. The spatial properties of the sound image deteriorated, if the direction vectors were averaged in time to reduce the amount of artefacts, however. The averaged directions remained near the center direction almost all the time. Decreasing the time constant of the averaging caused again variations in the directions of the point-like sources, which were heard as clicking, as the signal powers of some bands of the signal channels were changing rapidly. A specific problem of Goodwin and Jot's method was that the stereo material which had been panned to the center was played solely from the center channel of the three-channel system. This made the sound image narrower than what it was when it was played from the stereo loudspeakers.

3.4 Conclusion

Four methods for transforming stereo or multichannel audio signals from a format to another were presented in this chapter. The methods are called Faller's method, Goodwin and Jot's method, DirAC and the PCA method. First, a general introduction to the methods was given. The spatial analysis of the audio signals was then presented in detail for all the methods. Finally, the transform modules of the methods were described. All the methods divide the audio signals in frequency bands that mimic the critical bands of the human auditory system. The directions of the primary directional components of the sound are analyzed for all the bands. The power ratios between the directional and non-directional sound are also estimated. In DirAC the directional and non-directional components were called the non-diffuse and diffuse sound, respectively. Two of the methods, Faller's method and the PCA method, also extracted the primary component as well as the remaining ambience signal from the mixture. Goodwin and Jot's method or DirAC do not extract specific signals from the mixture. Diffuse sound streams are synthetized by means of decorrelation in DirAC.

The methods were intended for different kind of signals. Faller's method was only applicable for two-channel stereo signals, because it made an assumption that the ambience signal of the left and right sides have equal frequency band-wise power levels. This assumption is difficult to generalize for multichannel content. Goodwin and Jot's method, by contrast, is not very usable for the transformation of the two-channel stereo, because it cannot perceive non-directionality from this kind of signals. DirAC and the PCA method can be used to conduct the spatial transform on any signal that has two or more channels. More studies will be needed on how to apply the methods for transforming 5.1-surround to other formats, for example. Faller's method appeared to be very promising for the modification of stereo signals, and it would be highly desirable to be able to generalize it for audio content with more signals. PCA could be applied directly to multichannel content, but its drawback is that it forms the primary component as a combination of all the signal channels. This is not necessarily desirable, since the pair-wise panning of sound sources is still a popular mixing technique in multichannel audio (see Section 2.2). Therefore a pair-wise analysis approach similar to Faller's method is under interest for multichannel signals. This, however, would require more sophisticated models that could estimate the powers of channel-specific independent signal components.

Chapter 4

Analysis of multichannel audio mixes

Models for amplitude panning were presented in the earlier sections. In these models, a multichannel mixture could be interpreted as consisting of signals that were amplitude-panned between two signal channels or independent signal components that were unique for a signal channel. These models could be used in the analysis of multichannel mixtures. One of the simplest realizations of these models is having one amplitude-panned common signal between each channel pair and one independent signal on each channel. The model has less sensor signals than source signals, which makes it underdetermined according to Section 2.5. Therefore, some restricting assumptions about the model have to be considered, or the equations will have infinite number of solutions. The Faller model, which was presented in Section 3, supposes that in a two-channel stereo signal there is one common amplitude panned signal between the stereo pair and an independent signal in both channels. In this model, the necessary assumption was that the independent signals of the stereo channels have equal signal powers. The degree of freedom for making this kind of assumptions is rather small for two-channel stereo. When the number of channels is increased, it gets more difficult to make intuitive hypotheses about the powers of the signal components and about the interchannel relationships in general. Further investigation is needed to study which signal channel pairs contain common signals more often than the others.

Knowing the mixing techniques that are used in the production of multichannel mixtures would help making the necessary assumptions about the content. This, in turn, would help the process of transforming the audio to a different loudspeaker layout. A database of 5.1 multichannel audio was extracted from DVD records and analyzed in order to gain better understanding of the mixing techniques used to produce contemporary multichannel audio. The aim was to measure which of the channel pairs in 5.1 audio content usually contain common amplitude panned sound events. Knowing this, it would be possible to allocate, for example, more computational resources for analyzing one signal pair than for some other signals pairs. The signal powers of the audio channels were also measured to support later development of the multichannel audio analysis algorithm.

4.1 Test setup

A database of five-channel audio signals was constructed by extracting the audio tracks from commercially available 5.1 DVD recordings. The low frequency effect (LFE) channels of the recordings were omitted. There were 14 DVDs in total. The main focus was in analyzing the audio tracks of DVD movies, hence 12 of the audio entries in the database were from movies. The audio tracks from two concert DVDs were also analyzed for comparison. The DVD recordings are listed in Table 4.1. The audio tracks were resampled from 48 kHz to 44.1 kHz. This was done to maintain compatibility with the analysis-synthesis algorithm in development, which was working at the sampling frequency of 44.1 kHz.

Table 4.1: The DVD records in the database.

DVD name	Type	Year
Buena Vista Social Club	Concert	1999
Cars	Movie	2006
Chigago	Movie	2002
Dreamgirls	Movie	2006
Ice Age 2: The Meltdown	Movie	2006
In the Line of Fire	Movie	1993
Jumanji	Movie	1995
The Lord of the Rings: The Return of the King	Movie	2003
Ray	Movie	2004
Santana: Supernatural Live	Concert	2000
Spiderman	Movie	2002
Spiderman 2	Movie	2004
Spongebob Squarepants Movie	Movie	2005
Twister	Movie	1996

The DVD audio signals were analyzed in chunks of 2^{15} samples, which corresponds to 0.743 ms at the 44.1 kHz sampling rate. The chunks of this particular size were used to enable later comparability with the analysis results of a proprietary speech activity detector, which required long analysis windows for proper detection. Using smaller chunks would have enabled better temporal resolution for the measurements, on the contrary.

The main target of the measurements was in investigating how the amplitude panning is used in the mixing process of the DVD audio tracks. This was measured by calculating cross-correlation coefficients between the audio channels. The chosen correlation coefficient was Pearson's product-moment correlation coefficient, which was given in Equation (2.9). Pearson's coefficient was considered good for measuring pure amplitude panning relationships because it does not detect time shifts, which occur in convolutive mixtures.

To lighten the computation power of the calculations, an assumption of symmetry between the left and right half planes was made. The assumption states that over a long time period the spatial properties of the audio channels on the left side of the listener are the same as those of the audio channels on the right side of the listener. This can be refered to as a concept of spatial balance: in the long run, there is equal amount of content on the left side and on the right side of the listener. The instantaneous focus of the audio events can, however, be on either side. Figure 4.1 illustrates this assumption, which states that the correlations between the front right (FR) channel and the center (C), back left (BL) and back right (BR) channels are approximately the same as the correlations between the front left (FL) channel and the C, BL and BR channels. Similarly the back right - center (BR-C) cross-correlation is near that of the BL-C channel pair.

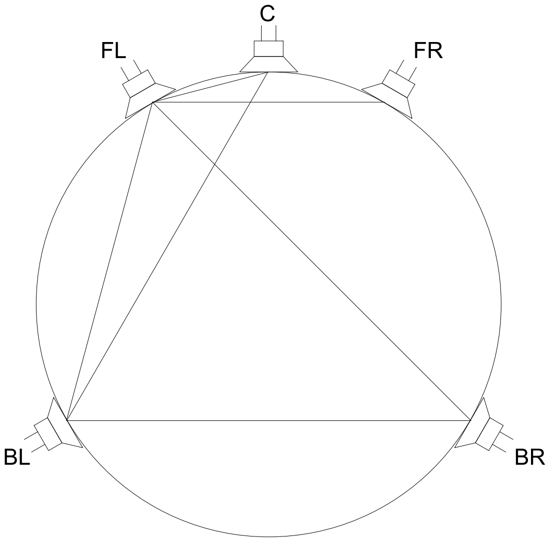


Figure 4.1: Channel pairs for which the cross-correlation coefficients were calculated

The signal powers of each signal chunk were calculated to study where the sound is usually concentrated in the multichannel mixes. The power measurements could be also used to verify the initial assumption of balance between the left and right half planes, and the results of the power measurements supported this assumption, indeed. Further results are being discussed in the next section.

4.2 Results

The first measures calculated from the correlation data were the average correlations between the channel pairs. Table 4.2 shows the average correlation values of the movie DVDs. There are six values, one for each channel pair. The channel pairs are clearly divided in two groups of three. The other group has the average correlation values near zero while the other has values from 0.16 to 0.26. The strongest correlation is between the front stereo pair (FL-FR). The two next pairs with almost the same average correlation are the back stereo pair (BL-BR) and the front left - center channel pair (FL-C). Three channel pairs having least correlation are the left side (FL-BL), the back left - center (BL-C) and the front left - back right (FL-BR). This means that the back left (BL) channel is largely independent from the three front channels.

Table 4.2: Average correlations between each of the analyzed channel pairs for the movie DVDs in the database.

Channel pair	FL-FR	FL-C	FL-BL	BL-C	FL-BR	BL-BR
Mean value	0.26	0.16	0.02	0.01	0.01	0.19

The results in Table 4.2 give an indication of the channel pairs that have significant common signals. Following from the left-right symmetry assumption, the front right - center pair (FR-C) is one of these as it is the counterpart of the FL-C pair. It can be deduced using the symmetry, correspondingly, that the back right channel contains only little in common with the three front channels. A summary of the measured interchannel relationships is illustrated in Figure 4.2, where the signal channel pairs that contain the most significant correlation have been connected to each others using lines. The figure shows, that the common five channel system is usually divided in two channel groups, the frontal channels and the back channels. It is more probable to find amplitude panned signals between the members of the same channel group than between the members of separate groups. This means that most effort in analyzing the interchannel relations for detecting amplitude panning should be put into the relations between the members of these groups, and the front channels and the back channels can be analyzed separately.

The results that were given above are time averages for the analyzed movie DVDs, and do not tell about the instantaneous relationships between the channels. The differences of the time-averaged correlation coefficients between the movies are not shown either. One movie, for example, had the average correlation values exceptionally below 0.02 for all the measured channel pairs except for the BL-BR pair, which had the average correlation of 0.06. Two other movies had 0.05 as the average correlation value between the FL-C pair, which was not common behaviour either. The FL-FR channel pair has the largest average correlation value in Table 4.2. This channel pair had the most significant correlation only in half of the movies. For four movies the most significant pair was the BL-BR. The FL-C and FL-BR pairs were the most significantly correlated pairs of one movie each.

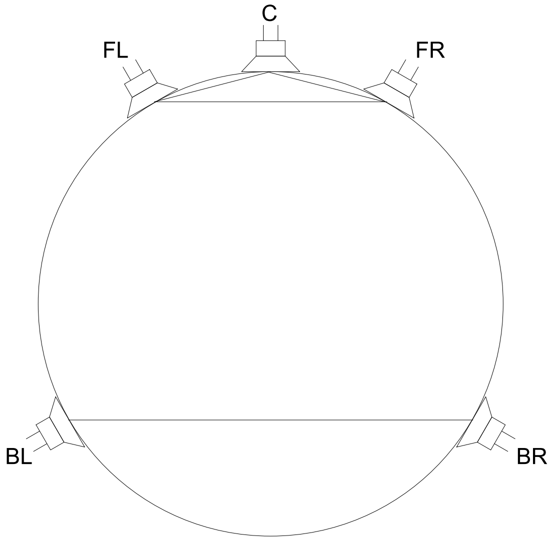


Figure 4.2: The channel pairs that have the most common signals. The signal channels which have significant common signals have been connected to each others with lines.

The two concert recordings in the database had correlation values that differed significantly from the values of the movies. The average correlations for the both concert DVDs have been presented in Table 4.3. Comparison to Table 4.2 reveals that both, the movies and the concerts, have significant correlation between the FL-FR and BL-BR pairs, albeit these correlation values are higher for the concert audio. The four other measured channel pairs (FL-C, FL-BL, BL-C and FL-BR) have completely different correlation values between the concert DVDs. For the first concert DVD the FL-BL, BL-C and FL-BR have correlation values near zero, which is similar to the average behaviour of the movies. Surprisingly, the FL-C pair has zero average correlation in the first concert. For the second concert DVD all the correlations are above 0.16 and there is also correlation between non-adjacent audio channels. The results suggest that the first concert DVD has been produced by mixing single instrument tracks, and that the second concert DVD has been recorded using a microphone configuration to capture the sound field (see Section 2.2). No general conclusion, however, can be made from these concert DVDs because there were only two concert recordings in the database. Still, these results indicate that the concert DVDs are mixed in a different manner than an average movie DVD.

The instantaneous powers of the signal channels in the movie DVDs were averaged over time. The average powers of all the movies show that the biggest portion of the signal

Table 4.3: Average correlations between each of the analyzed channel pairs for the two concert DVDs in the database.

Channel pair	FL-FR	FL-C	FL-BL	BL-C	FL-BR	BL-BR
Concert 1	0.59	-0.05	-0.02	-0.01	-0.01	0.40
Concert 2	0.64	0.79	0.33	0.27	0.16	0.42

energies is concentrated on the center channel. For all the movies, the ratio between average power of the center channel and the average power of the loudest remainging signal channel was between 1.9 and 5.8. The comparison of the average channel powers are given in Table 4.4, which gives the ratios in desibels between the powers of all channels to the power of the center channel. There are noticeable differences between the average power levels of the center channel, the front stereo channels and the rear stereo channels.

Table 4.4: Ratios between the channel powers and the power of the center channel.

Channel	FL	FR	BL	BR	C
Average power (dB)	-4.7	-4.5	-10	-10	0

The average powers of the front left and front right channels are usually approximately the same. The relative difference between the average powers is given by

$$\Delta = 2\frac{|P_1 - P_2|}{P_1 + P_2}, \tag{4.1}$$

which is the absolute difference between the powers P_1 and P_2 of the signals 1 and 2 divided by their average. The relational difference between the average left and right channel powers was calculated for all the movies. The values revealed that the relational difference between the front stereo channels is 0.07 by average and mostly below 0.10. There was also one clear outlier, *Twister*, which is a catastrophe movie from the year 1996 (see Table 4.1), and which had 0.36 relational difference between the average powers of the left and right channels. A similar balance can be found between the rear stereo channels as well. For the BL and BR channels, the relational difference is below 0.17 in all the movies, while the mean difference is 0.10. Thus, the balance between the left and right is slightly smaller at the back than at the front of the listener.

The center channel had the highest mean power value of all the channels. Intuitively, it should be very probable that the center channel has the most powerful channel signal at a given time instant. Similarly, the front stereo channels should be instantaneously the most powerful channels with equal probabilities. The same applies for the back stereo channels. The probabilities for a channel to have the highest instantaneous power are plotted in Figure 4.3. The plot is a box plot, and the boxes have lines at the lower quartile, median and upper quartile of the values of each channel over all the movies. Whiskers extending from the boxes show the extent of the rest of the data. There is one outlier for the front left channel that has been marked with a cross. There is a noticeable difference between the distribution of the FL and FR channels. The front right channel has the probabilities mostly near the median but the probabilities of FL spread out more. The outlier case, movie *Ray*, had 0.03 probability for FL channel but 0.40 probability for FR channel. This, however, does not tell about the overall balance between the channels because the difference in decibels might be really small or vice versa. The mean values of the probabilities that the channel is the instantaneously most powerful channel are shown in Table 4.5. The median values in Figure 4.3 seem to match with the mean values in Table 4.5. The center channel has the largest power for 2/3 of the time, which was excepted. The back channels are most dominant only rarely.

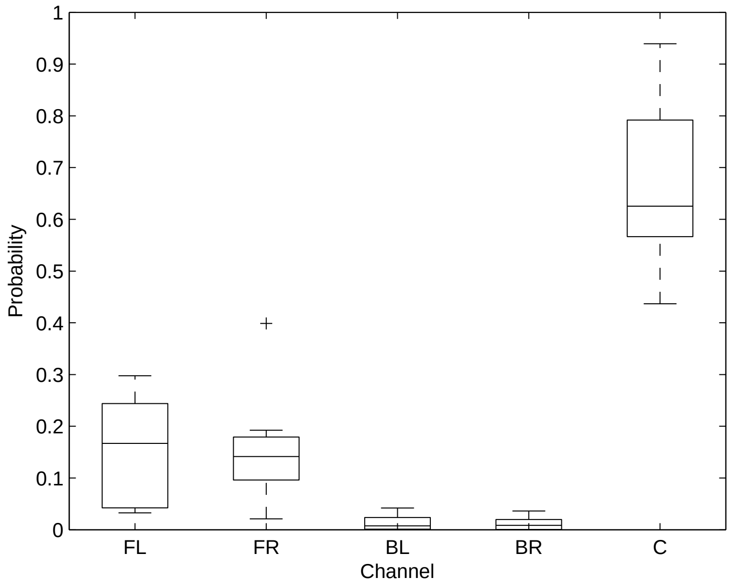


Figure 4.3: Probability of the channels to be the instantaneously most powerful channel of the movie. The distributions have been plotted over all the movies.

Table 4.5: Probability for a channel to have the highest instantaneous power in a movie.

Channel	FL	FR	BL	BR	C
Probability (%)	0.16	0.15	0.01	0.01	0.67

4.3 Conclusion

Five-channel audio recordings were analyzed to investigate the properties of the audio channels and to detect which channel pairs contain common amplitude panned signal components more often than the others. The analyzed material was extracted from 12 DVD movies and 2 multichannel concert recordings.

The results revealed that there are amplitude panned signals most often between the front stereo (FL-FR), the front left - center (FL-C), the front right - center (FR-C) and the rear stereo (BL-BR) pairs in the movies. Amplitude panning between the three frontal channels (FL, FR, C) and the rear channels (BL-BR) is rare. These results can be used to allocate analysis effort to these interchannel relationships. The nature of the results is general, however, and apply to the most of the movies but not all of them. Similarly, the instantaneous interchannel behaviour can differ from the results given above, and include amplitude panning between the front and rear channels, for example. The channel pairs that had most significant correlation values changed from movie to movie.

The correlation measures were calculated using Pearson's correlation coefficient, which is little influenced by time shifts between the amplitude panned signals. Thus, the results of the measurements do not describe the interchannel relationships that have been achieved by using convolutive mixing, for example. The movie *Lord of the Kings: Return of the King* had significantly low correlation values for all channels, but it is improbable that the channels are completely independent. The movie is a big budget production from the year 2003, and most likely that its audio content has not been mixed using the conventional pair-wise amplitude panning methods, but recorded using microphone techniques.

There are weaknesses in the use of correlation coefficient as a measure of amplitude panning. When there are independent signals present in both of the channels that are being measured, correlation coefficient varies as a function of the amplitude panning coefficients, even though they conform with the energy-preserving rule (see Eq. (2.3) on page 8). The correlation coefficient is largest when the common signal has been panned equally to the both channels. The correlation is low if the other amplitude panning coefficient has a small value. This can be shown in the form of equation, and Faller's stereo signal model in Equation (3.1) is good for the purpose, but now written in the form of energy-preserving amplitude panning:

$$\begin{aligned} x_1(k) &= a_1 s(k) + n_1(k) \\ x_2(k) &= a_2 s(k) + n_2(k) \end{aligned} \tag{4.2}$$

Here $x_1(k)$ and $x_2(k)$ are the signal channels from which the correlation is measured. The common signal $s(k)$ is amplitude-panned using energy-preserving coefficients a_1 and a_2. Signals $n_1(k)$ and $n_2(k)$ are independent, and they do not contribute to the correlation coefficient. Similarly to what was done in Equation (3.9), the correlation coefficient Φ can be characterized by

$$\Phi = \frac{a_1 a_2 p_s}{\sqrt{p_{x_1} p_{x_2}}}, \tag{4.3}$$

where p_s is the power of the common signal, and p_{x_1} and p_{x_2} are the signal powers of the known signal channels, which have the following dependency on the power of the common signal, the panning coefficients and the powers of the independent signals:

$$p_{x_1} = a_1^2 p_s + p_{n_1} \tag{4.4}$$

$$p_{x_2} = a_2^2 p_s + p_{n_2} \tag{4.5}$$

Figure 4.4 shows how Pearson's correlation coefficient changes as a function of amplitude panning coefficients. White noise signals were used for $s(k)$, $n_1(k)$ and $n_2(k)$. The common signal $s(k)$ is 9 dB more powerful than the channel-specific independent signals $n_1(n)$ and $n_2(k)$, which have equal powers. The way how the correlation coefficients measure amplitude panning stresses the fact that the results of this DVD analysis experiment do not reveal accurately the amplitude panning relationships between the channels. Thus, the results give indications of the relationships but more complex similarity measures are required to exactly measure the use of amplitude panning.

The instantaneous powers of the five-channel audio signals were calculated in addition to the investigation of the interchannel amplitude panning relationships. The power measurements revealed that the center channel was the most powerful channel in the analyzed DVD movies 67% of the time and it was generally 4.5 dB more powerful than the other frontal channels and 10 dB more powerful than the rear channels. The powers of the front left and the front right channel were approximately equal in overall. Same observation can be made from the equal powers of the rear channels. There was a preliminary assumption of the symmetry between the signal channels of the left and right sides in the test. The long run balance between the power levels of the left and right supports this assumption.

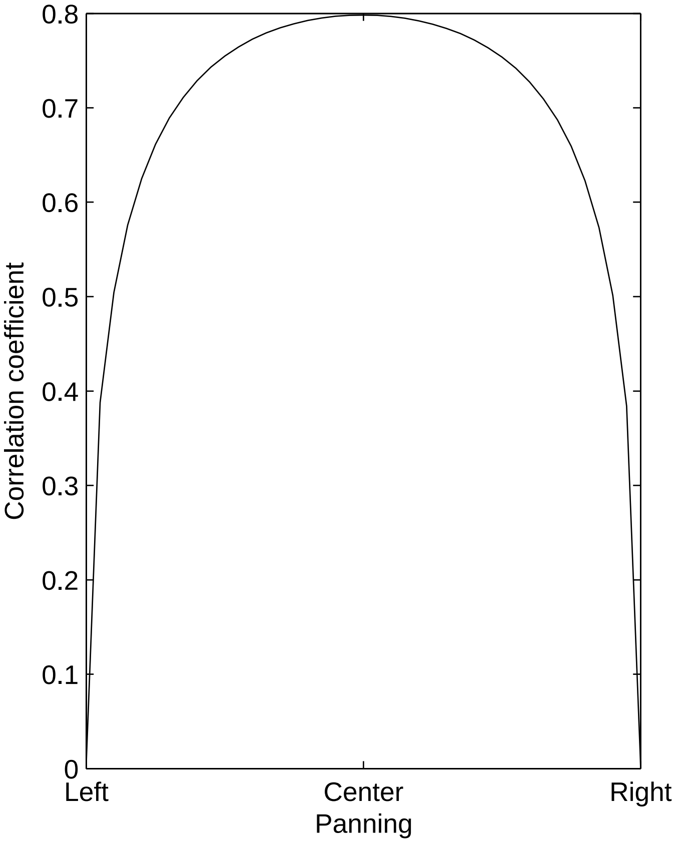


Figure 4.4: The correlation coefficients between two channels which consist of a common amplitude panned signal and independent signals. The common signal is 9 dB louder than the independent signals. The signals consisted of white noise.

Chapter 5

Experiments

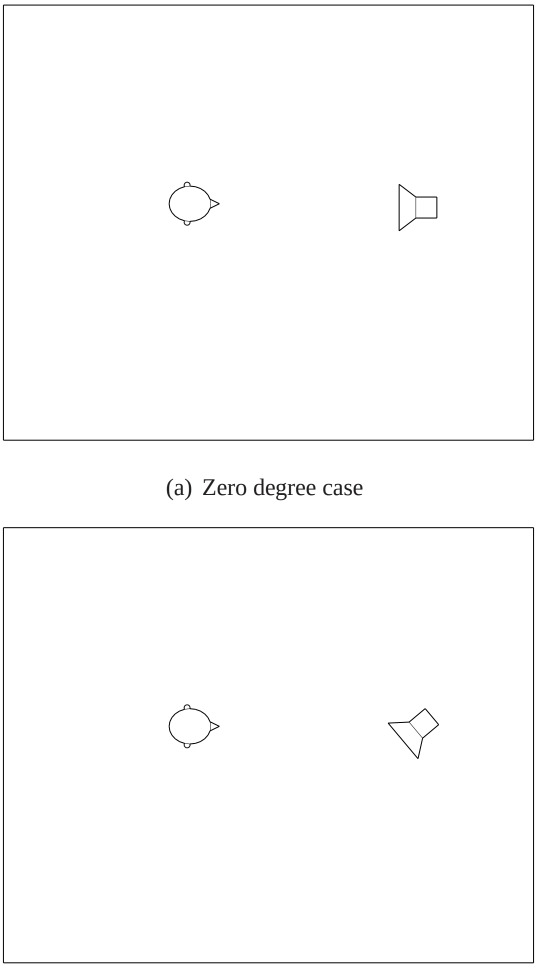
5.1 Motivational listening test

A particular topic of interest in this work was finding modification methods for playback of audio signals from compact, non-standard, loudspeaker configurations. A specific listening experiment was designed in order to study what are the differences in various physical arrangements of loudspeaker drivers in such a compact loudspeaker system. However, because of practical reasons the test was conducted using headphones as sound reproduction devices. One of these reasons was the requirement of keeping the experiment as a blind test, in other words not giving the subjects any clues about the loudspeaker layouts that they were listening to. It would have been also a time-consuming task to change the layout after each test case. Finally, the loudspeaker virtualization allowed flexibility in the positioning of the loudspeaker.

Binaural impulse responses of a loudspeaker were measured using an artificial head. The measurements were repeated for different loudspeaker orientations. The impulse responses were used to create virtual loudspeaker systems that could be listened using headphones. Monophonic and multichannel versions of test audio signals were played from these virtual loudspeaker systems and the subjects were asked to detect differences between mono and multichannel cases. This was expected to give information about the different loudspeaker layout and their suitability for compact but spatially rich playback of audio.

5.1.1 Test setup

The first step in setting up the test was to measure binaural impulse responses of M-Audio StudioPro 3 loudspeaker with different orientations. The floor plan of the listening room was rectangular with one 5.2 m long wall and another 6.3 m long wall. The height of the room was three meters. Georg Neumann GmbH KU81i artifial head was used for recording the impulse responses. The location of the artificial head remained constant during the measurements whereas the angle of orientation of the loudspeaker changed. The room and the locations of the loudspeaker and the artificial head are illustrated in Figure 5.1. The loudspeaker was intentionally placed to a non-symmetrical position in the room. Non-symmetrical in this case means that the distances from the side walls were not equal. The loudspeaker was also placed 1.5 meters away from the back wall. There was a distance of 2.5 meters between the artificial head and the loudspeaker. The head and the loudspeaker were at a same height, 80 cm. The room layout was wanted to follow a regular rather small living room, where a compact loudspeaker system and the listener are more towards to the center of the room than to the walls.



(a) Zero degree case

(b) -40 degree case

Figure 5.1: The layout of the room while recording the binaural impulse responses of the loudspeaker with various angles of direction.

In the initial setup, the loudspeaker driver was facing the ears of the artificial head. This case shown in Figure 5.1(a) is referred to as zero degree case. The binaural impulse response of the loudspeaker was measured in this location. Before the next measurement the loudspeaker was rotated 10 degrees towards the wall on the left side of the artificial head. The rotation of 10 degrees was repeated after every measurement until the loudspeaker was directly facing the wall on the left side of the head, thus reaching the azimuth of 90 degrees. Then same measurements were repeated so that the loudspeaker driver was facing the wall on the right side of the artificial head, covering the rotation angles from $-10°$ to $-90°$. Figure 5.1(b) shows the $-40°$ case as an example. The measurements yielded 19 binaural impulse responses in total.

The measured binaural loudspeaker impulse responses were used to construct 10 different virtual loudspeaker layouts. All the layouts consisted of three virtual loudspeakers named left side, center and right side speaker according to the directions where their drivers were pointing to. In all cases, the center loudspeaker was pointing directly at the listener while the direction angles of the side loudspeakers changed between different layouts so that they had different signs but same absolute values. The side loudspeakers were, that is to say, mirror images of each others in respect to the center axis.

The virtualization of the loudspeaker layouts was achieved by convolving three mono channels with respective binaural impulse responses and summing up the resulting stereophonic signals. Two different tri-channel signals were used for the experiment. The first signal was composed of four separate instrument tracks of a same song. Two of the instruments were played back from different side channels and the last two from the center channel. The second signal consisted of white noise bands with equivalent rectangular bandwidths (ERB). None of the bands were overlapping and they were thus uncorrelated with each others. The noise bands had equal sound power levels. Both side loudspeakers played four bands and the center speaker played 8 bands. For both signal types all 10 virtual loudspeaker layouts were calculated. For all the layouts and signal types also comparison signal was made by summing the three audio channels as one monochannel which was then convolved with the three loudspeaker impulse responses.

The headphone model used in the test was Beyerdynamic DT-990 Pro. There was not any individual headphone equalization used. During the experiment, the subjects had to compare a reference sound X to samples A and B and tell which one of the samples A or B had been chosen as the reference. The A samples represented always the situation where all three virtual loudspeakers played uncorrelated channels and the B samples the situation where all three virtual loudspeakers played the same mono down-mix. The user interface of the test program can be seen in Figure 5.2. There were 20 different ABX test cases: 10 loudspeaker layouts for both two signal types. The subjects answered to each test case only once, and the order of the test cases was randomized for all the subjects. In the beginning of the test the subjects went through a short training period during which they learnt how to use the test program to listen to the samples, and practiced hearing the possible differences between them.

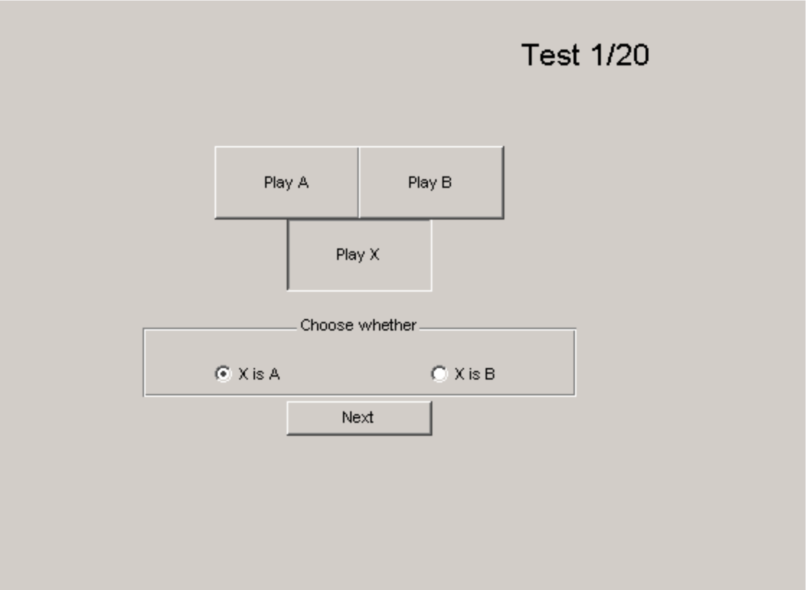


Figure 5.2: The user interface of the listening test. The listeners had to choose whether the audio sample X was equal to the sample A or sample B.

5.1.2 Results

In total 16 subjects performed the listening test. The results are shown in figure 5.3. The Y-axis denotes the ratio of correct answers and the direction angles of the virtual side loudspeakers are on the X-axis. For each of the 10 side loudspeaker azimuths there are two bars: black bars for the music cases and white bars for the noise cases.

With both signals types, a raising tendency can be found from the results as the azimuths of the side loudspeakers are increased. For the zero degree case the proportion of correct answers stays below 50%. It should be remembered that in the zero degree case the samples A and B are the same, although a slightly different processing procedure was used. The difference between the samples is that all four signal channels have been first convolved separately with the binaural impulse response of the zero degree case and then summed in A. In B, the signal channels were first summed as four equal mono channels and then convolved separately with the same impulse response.

The number of correct answers at $0°$ was 7 for the music and 5 for the noise, which are 47% and 31% of all the answers. The latter ratio seems low when one keeps in mind that

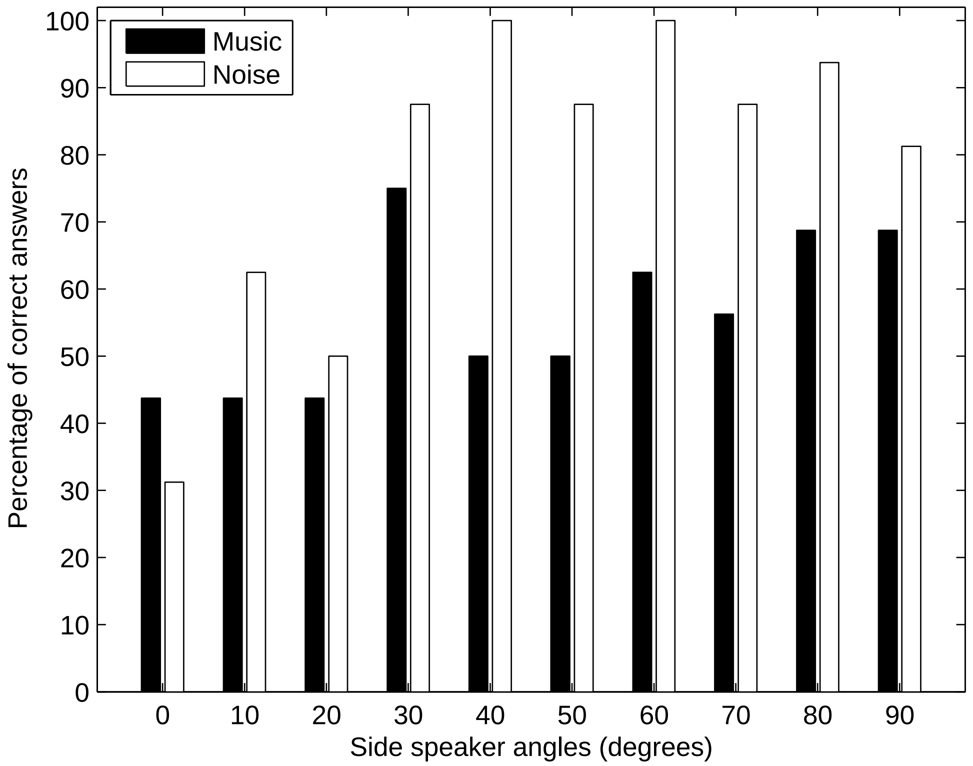


Figure 5.3: The results of the listening test. For both, the music and the noise samples, the results are reported as percentages of correct answers for different direction angles of the side loudspeakers.

there were equal probabilites for correct and incorrect answers, and the expectation value is 50%. According to the probability theory, the probability of getting k correct answers from n listeners is given by the probability mass function of the binomial distribution

$$P(K = k) = \binom{n}{k} p^k (1 - p)^{(n-k)}, \tag{5.1}$$

where p is the probability of a correct answer. This probability is 0.5 for the zero-degree case. Figure 5.4 shows the binomial distribution for $n = 16$ listeners. The probabilities of getting 7 and 5 correct answers in the case of $p = 0.5$ are 17% and 7%, respectively. The probability that half of the answers are correct is 20%. These probabilities can be compared to the case when there are $n = 100$ listeners. In our $0°$ case for noise, there are three correct answers less than the expectation value. If $n = 100$, the probability that the number of correct answers is three less than the expected amount of the correct answers is 7%, which is near to the probability of getting 5 correct answers when $n = 16$. However, if $n = 100$ and $k = \frac{n}{2} - 3 = 47$, the ratio of correct answers is 47%. Thus, a quantitatively small and probable deviation from the expection value of the correct answers can lead to a big drop in the ratio of the correct answers if the number of test subjects is as low as in this experiment.

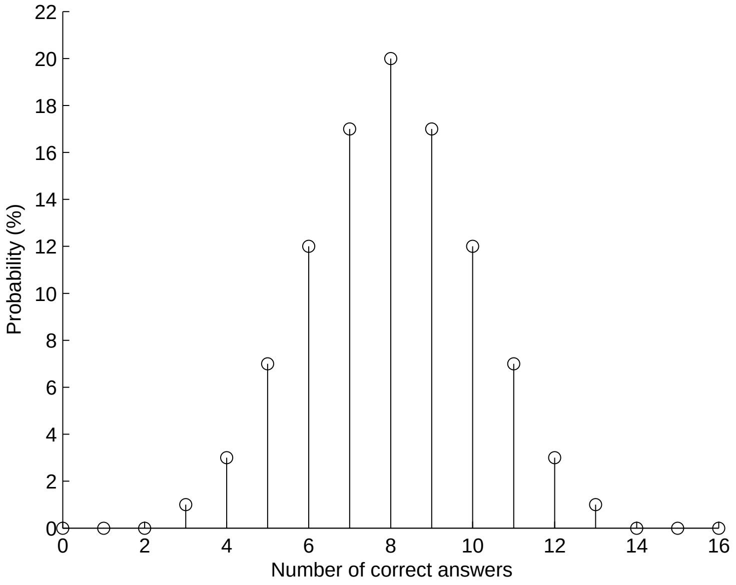


Figure 5.4: Binomial probability distribution for $n = 16$ tests.

As the direction angle increases, the ratio of the music signals increases only slightly. At 90 degrees the ratio has risen to around 70%. At this azimuth the difference between the samples A and B should be most evident. This is caused by the directivity properties of the loudspeakers, which attenuate the high frequencies that are radiated to the sides of the loudspeakers significantly more than the low frequencies coming to the same direction. The directivity patterns of the loudspeakers are shown in Figure 5.5. When the side loudspeakers have been rotated, the directivity pattern of the loudspeaker attenuates and colorizes the high frequency sound that arrives to the ears of the listener via the shortest path possible. This in turn should increase the perceptivity of the high frequency reflections from the walls because they do not suffer from the colorization or the attenuation that much. This means that the increase of listener envelopment achieved by the reflections should be most noticeable at the azimuth of 90 degrees. With the noise the increase of the right choices at higher direction angles is much more evident. After the azimuth of 30 degrees the proportion of correct answers in the noise cases fluctuates between 80% and 100%.

One more thing to notice from Figure 5.3 is an exceptional peak with the music cases at the azimuth of 30 degrees, where the amount of correct answers is at highest, at 75%. This

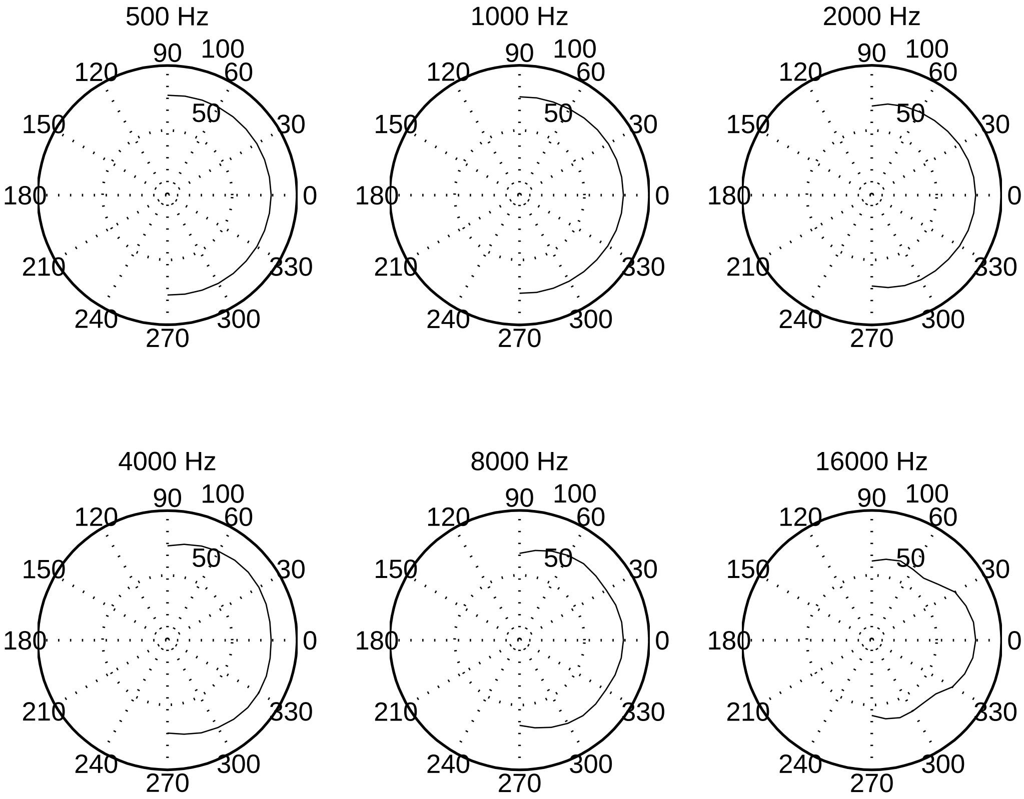


Figure 5.5: The directivity patterns of the loudspeaker used in the test. The responses have been normalized to 80 dB for on-axis (0°) position.

may indicate that there is a clear difference between the samples at this azimuth. However, the listeners did the selection for each case only once, which means that the test results cannot give a statistically significant measures about the just noticeable differences. Some listeners that could not distinguish between the samples A and B could have still chosen the correct answer. This problem of a possibly excessively high peak is closely related to the low proportion of the correct answers for the noise samples at zero azimuth.

5.1.3 Conclusion

Most subjects reported that choosing the sample equal to the reference was easy for the noise samples and very difficult for the music. This observation is consistent with the results in Figure 5.3, which show that the probability to choose a correct answer with noise increases significantly as the rotation angle increases. The same increase is much slighter with music.

The better ratio of correct answers for the noise samples at higher direction angles can be explained by coloration of high frequencies. Sound is radiated also from the sides of the loudspeaker, and the high frequencies of this leakage sound are much more attenuated than the low frequencies. When the loudspeakers are rotated the sides of the loudspeakers are facing the listener more and more directly, and thus more of the leakage sound reaches the ears of the listener. The majority of the sound energy radiated from the loudspeaker will arrive to the ears of the listener after a wall reflection, and it has a longer path to the listener than the leakage sound, which is the first sound arriving at the ears.

The noise signals had much more powerful high frequency content than the music signals. When the rotation angle of the side loudspeakers was large, the spectral properties of the sound coming from the center loudspeaker and the side loudspeakers were different. The test subjects were comparing the differences between multichannel version and mono mix version of the sound, and could hear more easily the difference when some signal components moved partially from a loudspeaker to another. In the case of the music samples, the coloration of the high frequencies was not evident. Therefore the test subjects could not usually perceive the change of the spatial image or virtually any difference between the mono mix and the multichannel mix. The difference between the A and B music sample was probably too small compared to the corresponding difference between the A and B noise samples. The test could have been modified by advising the test subjects to listen to the drums of the music samples. The drums were played back from the right side loudspeaker in the multichannel mix and the drum track contained cymbals, which are very loud at high frequencies. Knowing this, the listeners could have been able to distinguish between music samples A and B more easily.

The experiment could not show clear differences between the virtual loudspeaker configurations. A palpable reason for this is that listening to virtualized loudspeakers using headphones is not a natural listening situation. The sound coming from the headphones is not changed if the test subject moves his or her head. The head movements, however, play big role in sound localization. A possibility of using a head tracker to change the sound according to head movements would be beneficial to make the virtual loudspeakers more realistic [5]. When the loudspeaker configurations of the experiment were listened to subjectively in the real environment and in various listening positions, clear differences could be perceived between the multichannel and mono mix cases.

Another issue in the listening test setup was the compactness of the virtual loudspeaker system. The binaural impulse responses were recorded from the loudspeaker which always remained in pretty much the same location of the room. Only the orientation of the loudspeaker was changed between the measurements. The system was almost unrealizably compact. Although this was specifically wanted from the system, some later non-formal experiments were made with a system, in which the side loudspeakers had been moved 20 cm off from the center and more towards the walls. These tests suggested that moving the side loudspeakers could help in discriminating between the multichannel and the mono mix, when listening to the virtual loudspeaker system played back from headphones.

5.2 Enhancing spatial dimensions of a compact loudspeaker playback system

A listening test was conducted to experiment with the playback of stereo audio signals from an experimental compact loudspeaker system. The objective of the test was to convert two-channel signals to three-channel signals, and play the resulting signals from three loudspeakers. Using a three-channel system was believed to enhance the spatial dimensions of the perceived sound image in comparison to a two-channel reproduction system. Unlike the test that was described in Section 5.1, this experiment was conducted using a real compact audio system rather than a virtualized loudspeaker system. A compact two-channel system was used as a reference playback system for the unprocessed stereo audio.

Three of the audio transform techniques presented in Section 3 were chosen as processing methods for transforming the two-channel stereo audio to three-channel form. The aim of the processing was to get a separated center channel and two side channels that would be reflected from the boundaries of the listening room, and could produce sensation of a diffuse sound field. The participants of the test were asked to evaluate the spatial dimensions of the sound image in comparison to the reference stereo. All the listeners performed the test in two listening positions.

5.2.1 Test setup

The experimental sound reproduction system used in the test was a three-channel compact loudspeaker system, which consisted of a center loudspeaker and two side loudspeakers. The model of the loudspeakers was M-Audio StudioPro 3, and their directivity patterns are shown in Figure 5.5. The side loudspeakers were in angles of $\pm 90^\circ$ while the center loudspeaker had its driver pointing to the zero degree rotation angle, thus directly to the listener sitting at the sweet spot. A compact stereo system was placed on top of the experimental three-channel system as a reference system. The test setup is illustrated in Figure 5.6. The three-channel playback system was designed to be situated in a rather small room, where the room dimensions would allow exploiting the wall reflections in producing diffuse sound that would reach the ears of the listener from the sides. This was expected to enhance the spatial properties of the perceived sound image.

The objective of the test was to measure changes in perceived spatial dimensions of the sound image. The comparison was made between the different processing methods and unprocessed stereo. Unprocessed reference stereo was played back using the reference stereo system which consisted of two loudspeakers in angles of $\pm 45°$. These rotation angles were conceived as a compromise between the $\pm 90°$ and $0°$ azimuths, at which the loudspeaker drivers would be facing to completely opposite directions or to exactly same directions. The stereo loudspeakers in $\pm 90°$ azimuths would sound unnatural to the listener at the center axis because all the loudspeaker drivers would be pointing off from the listener, and the direct sound reaching the listener directly without wall reflections would be coming through the side panels of the loudspeakers. The frequency-dependent directivity patterns (see Figure 5.5) of the loudspeakers would attenuate considerably the higher frequencies, which would cause coloration of the sound. The zero-degree system, on the contrary, would not make good use of the wall reflections and would be a bad reference system for the purpose of this experiment.

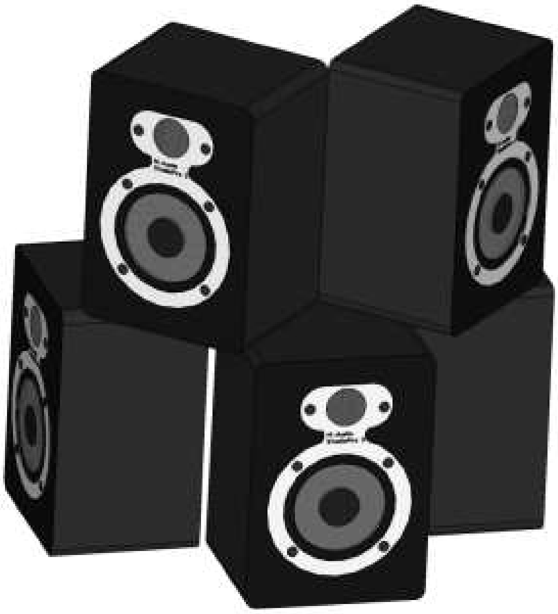


Figure 5.6: The loudspeaker layout used in the listening test. Three loudspeakers used for playing processed audio are on the bottom. Upper two loudspeaker were used to play the stereo reference.

The side loudspeakers of the three-channel compact loudspeaker system had two functions. First of all, they were used to play diffuse side signals, which would be reflected from the walls of the room. Having diffuse sounds coming from the sides of the listener was supposed to give a beneficial boost to the size of the perceived sound image. In the second place, the side loudspeakers were used to pan the primary center signals to be located between the center and the side channels.

Three of the different processing techniques described in Section 3 were applied to modify stereo audio signals to a suitable format for the experimental loudspeaker system. These were the stereo decomposition method by Faller, a primary-ambient signal extraction method based on principal component analysis, and DirAC. The transforming method by Goodwin and Jot was not considered to be very suitable for processing stereo signals. The analysis phase of the Goodwin and Jot method would merely calculate the energy-based localization vector for each sub-band, and interpret that point-like amplitude panned sources in the respective locations play the sub-band signals. The method would not extract any diffuse sound from stereo signals, and playing diffuse sound from the side loudspeakers was the primary starting point for the experiment.

The Faller and PCA methods both extract a primary signal and two side signals from the original stereo signal on each sub-band. The methods also analyze how the primary signals were amplitude panned between the original stereo pairs. The location information will be needed later in the synthesis process. In the experiment, it was assumed that the original stereo signals were intended for playback from standard stereo system that has the loudspeakers in $\pm 30^\circ$ angles. This assumption limited also the directions of the analyzed primary sources. In the synthesis phase, not much processing was applied to the primary signals given by the two methods. They were simply panned either between the left side and the center channels or the right side and the center channels. The panning method used was regular energy-preserving amplitude panning. The diffuse side signals, or the ambience signals in other words, were played back from the side loudspeakers without further processing.

DirAC, in comparison with the other two algorithms, has a different kind of analysis and synthesis scheme. First of all, the directional and diffuseness analysis is performed separately for each Fourier bin instead of critical bands [43]. Furthermore the primary sources are not separated from the diffuse sound. DirAC measures the directions of particle velocity at the listening position, while the other methods model the original arrival directions of the sound sources. The difference is that the particle velocity has a wave nature unlike the directions of the sound sources. Therefore the direction of particle velocity changes even in the case of single sound source. For the analysis of stereo audio this means that the directions are detected from an angular range that is greater than the assumed $\pm 30^\circ$ range of stereo playback. The azimuth angles resulting from the stereo analysis are limited to $\pm 90^\circ$ as reported in [41].

The three-channel signals synthetized by DirAC consisted of non-diffuse and diffuse part. The non-diffuse part was played from all three loudspeakers and the diffuse part was played from the side loudspeakers. The non-diffuse synthesis was based on the sum and difference signals calculated from the original stereo signal channels. The center channel played the sum signal, and the side channels played the difference signals. The left side loudspeaker played the phase inverted version of the signal played from the right side loudspeaker. The non-diffuse signals were multiplied by channel specific and time-adaptive amplitude gains. The gains were calculated according to the azimuth analysis information. Vector base amplitude panning (VBAP) [39] was used to calculate the instantenous gains, which were then adapted using channel energies in the same way as in [43]. The diffuse signals were filtered from the original stereo signals by using an uncorrelating filter. This experimental implementation used an all pass filter that randomized the phases of the Fourier bins. Finally, the non-diffuse and diffuse signals were summed with the appropriate diffusiness weights applied.

Eight stereo audio excerpts were processed using the aforementioned methods and played in the test. Five of the audio samples were taken from pop and rock albums released in the years 2006-2008. These excerpts were representing present day mixing techniques. For comparison, an additional pop song from the year 1998 was used as a test sample. The seventh sample was taken from a radio broadcast and it consisted of the speech of a sports reporter over a singing audience at a hockey stadion. The last sample was a recording of a symphony orchestra.

There were 12 participants in the listening test. The participants were research students or researchers in the field of acoustics and audio engineering. They were asked to compare the processed samples to the reference and give their preference for the spatial width and depth of the perceived sound image. The abstract term sound image was told to be comparable to the size of the sound stage on which the sources in the sound mix were heard to be located. The test subjects were advised to ignore all the other possibly changed charasteristics in the modified sound and sound quality issues, for example distortion and audible artefacts.

The evaluation scale of the test was from 0 to 100. The score 50 meant that the processed sound had equally good spatial dimensions as the reference stereo. Lower score denoted that the spatial dimensions of the processed sound were worse than those of the reference stereo. Similarly scores greater than 50 were given to samples which had better required properties than the reference. The participants were also asked to give comparable scores to the three processed samples of each audio excerpt so that the one with the best spatial dimensions had the highest score and the worst one had the lowest score. The user interface of the listening test program is shown in Figure 5.7.

All the test subjects had a short supervised training period before the test. In the training

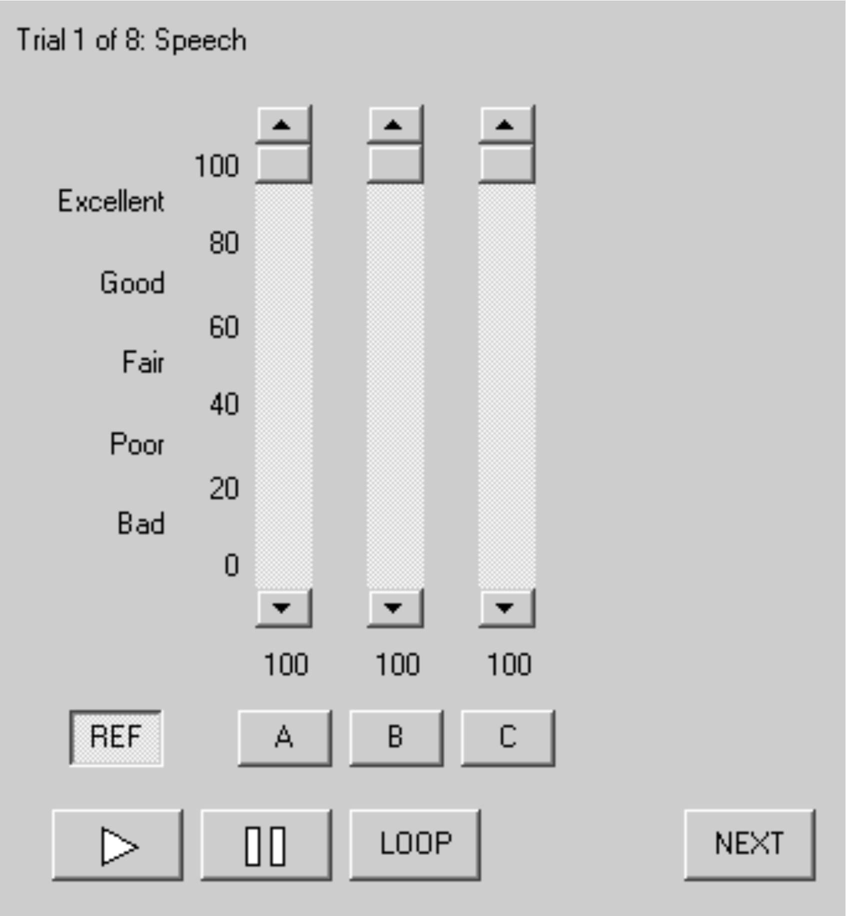


Figure 5.7: The user interface of the listening test program.

they learned how to use the test program and which properties of the signals they should compare between the processed samples. During the test, each of the participants listened to the reference stereo and all the processed samples of all eight audio excerpts. The orders of the excerpts as well as the orders of the processed samples were randomized for each participant. Everyone conducted the test in two listening positions. The first listening position was on the central axis of the loudspeaker system and the second one was an off-axis position closer to the wall of the room. Figure 5.8 illustrates the two listening locations in the room as well as the location of the loudspeaker system. It should be noted that the center axis of the loudspeaker system is not exactly in the midway between the side walls. This placement was intentional and it aimed to avoid the symmetrical special case of having the loudspeaker system on the center axis of the room. The dimensions of the room were: width 5.7 meters, depth 7.4 meters and height 3.0 meters. The room was intended to simulate a regular living room.

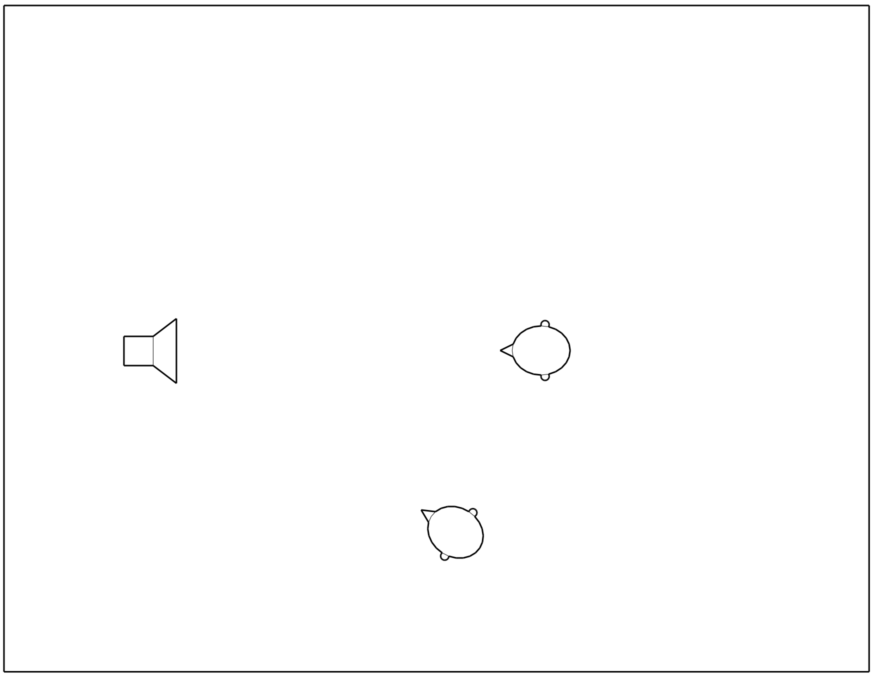


Figure 5.8: The locations of the listening positions and the loudspeaker system in the room.

5.2.2 Results

The results of the listening test were analyzed using a simple “win, lose or draw” system. The basic idea was that the processing methods were competing against each other in pairs and in every comparison the one with the better evaluation score got one point. If the listener gave same score to two methods, both methods got half a point. This system can be considered equal to a sports tournament of three competitors, which then get points according to won or drawn games. This comparison or “tournament” was made separately for the answers of each excerpt by each listener. Finally, the winning ratios of the method over the others could be calculated from the scores. This enabled producing simple key ratios of the performances that the methods have over all the subjects, all the excerpts, in both listening positions, or any combination of them.

The general results of the listening test are shown in Table 5.1. The results can be characterized by saying that the spatial dimensions of the method by Faller were usually rated the best from those of the chosen methods. It was rated better than DirAC in 86% of the cases and better than PCA method in 67 % of the cases. The PCA method had the second best spatial properties in general as it was rated better than DirAC in 75% of the cases while it was better than the Faller method only in one third of the cases. Hence, in overall scores, DirAC was the worst.

Table 5.1: The overall results of the listening test. On every row there is a percentage of what was the probability that the method in the specific row was better than the method named in the column.

Methods	Faller	PCA	DirAC
Faller	-	67	86
PCA	33	-	75
DirAC	14	25	-

Listing the results was not all that simple, however. The preferences between the three methods varied in some extent from excerpt to excerpt and specially between the listening points one and two. The difference between the listening points can be seen by comparing the Tables 5.2 and 5.3. The overall scores between the Faller method and DirAC are nearly the same for both the listening positions, whereas the scores of the PCA method decrease at the listening position two from those of the first listening position. The preference between the methods, however, stays the same for the both listening positions.

Table 5.2: The overall results in the listening position one.

Methods	Faller	PCA	DirAC
Faller	-	62	85
PCA	38	-	81
DirAC	15	19	-

Table 5.3: The overall results in the listening position two.

Methods	Faller	PCA	DirAC
Faller	-	71	87
PCA	29	-	70
DirAC	13	30	-

Interesting observations can be made by looking at the average values of the actual evaluation scores. Taking a look at the means per audio excerpt, for example, one can see how successful a processing technique was in comparison to the stereo reference. Table 5.4 lists the mean evaluation values of each processing technique for each song in both listening positions. It should be remembered that values over 50 mean that the processed sound had better spatial dimensions than the reference stereo whereas the values below 50 mean worse spatial dimensions. The Faller method has always means which are over 60. The means of the PCA method are generally lower than those of the Faller method but still more than 50. The excerpt number five is an exception: the Faller method has clearly lower mean score than the PCA method. By average, the two first methods are rated to improve the width and depth of the spatial sound in comparison to the reference stereo. On the contrary, the mean values of DirAC reach the reference score 50 only with three excerpts and exceed it merely with the eighth excerpt.

Table 5.4: The average evaluation scores given by the listeners. Each row presents the processing method specific values for one sound excerpt. Values over 50 mean that the processed sound had better spatial dimensions than the reference stereo whereas the values below 50 mean worse spatiality.

	Faller	PCA	Dirac
1	72	63	50
2	69	64	42
3	73	51	50
4	62	57	41
5	61	68	43
6	62	60	45
7	72	63	46
8	74	65	55

The dispersion of the evaluation scores was relatively large. The answers for each song and processing method over all the listeners had standard deviations which were nearly always something between 10 and 20. The best mean value in Table 5.4 was 74, which is 24 more the reference but the lowest average value stayed above 40. For most of the sound excerpts in the test, there were scores both greater and less than 50 given. There was not a simple suitable method found for normalizing the scores given by the test subjects and it should be remembered that the average values are easily affected by outliers. Therefore the results should be interpreted only as indicative of what kind of processing methods should be used in further testing and development for this kind of sound reproduction systems.

5.2.3 Conclusion

The experiment measured the performance of three processing methods that were used to transform two-channel stereo audio to a three-channel form that is suitable for the three-channel compact loudspeaker system under the experimentation. A compact two-channel stereo system was used as a reference sound source. The experiment evaluated the listening experiences of 12 test subjects. They were asked to score their preferences of perceived spatial dimensions, namely width and depth, of the sound image played from the compact three-channel loudspeaker system. The test showed that two of the processing methods, the Faller method and the PCA method succeeded to improve the sound image in comparison to the stereo reference. The Faller method was usually rated the best. The third method, DirAC, did not perform as well as the two other methods. DirAC was scored, by average, worse than the reference stereo.

There were two different listening positions under examination. The first listening position was the sweet spot, which was located three meters directly in front of the loudspeaker system. The second position was an off-axis listening position near the side wall of the room. The PCA method was rated slightly worse in the second listening position. The preference order of the three methods remained the same, however.

One reason for the poor performance of DirAC is that it has been developed for analyzing the three-dimensional B-format audio signals, which are meant to capture the sound field in a given location. DirAC is meant to calculate physical measures from the sound field information, although the author of the method has tested it also on two-channel and multichannel audio signals with success. However, the stereo signals contain only little spatial information in comparison to sound field measures. The more signal-based methods used in the test, therefore, had a better evaluated performance. DirAC cannot extract diffuse signals from stereo signals either, but the diffuse signals must be filtered from the original signals. Possibly the use of a more powerful uncorrelating filter could have improved the performance of DirAC.

The test showed that the spatiality of the sound image produced by a compact two-loudspeaker stereo system can be improved by adding a third loudspeaker so that there are a center loudspeaker and two side loudspeakers. Signal quality factors like the existence of audible processing artifacts or distortion were not measured in the experiment, and these should be tested separately. The experiment did not compare the compact loudspeaker system to the standard stereo playback either. Further investigation needs to be done on suitable processing methods for the compact audio systems and on the spatial improvements achieved by these systems. The number of test subjects needs to be increased. The methods that performed well in the test form a good basis for further development.

Chapter 6

Conclusion

This work was a research on transform techniques that modify multichannel audio content for non-standard loudspeaker configurations. The modification process was desired to preserve the spatial properties that the audio reproduction has in the original loudspeaker configuration. The initial objective of the work was to develop a technique of this kind. The type of loudspeaker systems that was under special interest was the compact loudspeaker systems. In such systems, the loudspeakers are located close to each others in a single spot. The research started from studying methods for spatial audio analysis. These included the human auditory system, which is indeed one of the most delicate audio analysis systems. The other described spatial analysis systems were more of a signal-based type, and included channel similarity measures and source separation techniques. The latter are specially used in the information theory. Various techniques for the spatial transformation of audio signals from a format to another were reviewed. The development work of the transformation technique was supported by an analysis of commercial multichannel audio recordings, which measured interchannel relationships and power values from the signal channels. Finally, two listening experiments were conducted to study the actual compact loudspeaker systems and suitable processing methods that adapt audio content for these systems.

Multichannel audio coding techniques rely also on the spatial analysis of multichannel signals. The signal similarity measures that are used in the coding techniques seemed like a promising starting point for the research. In these techniques, the similarity measures are calculated separately on frequency bands that mimic the auditory critical bands. They can be thereby described to be psychoacoustically motivated. The audio coding methods aim to remove similarities from the signals and then later re-synthetize them. This could be beneficial in the audio transformation techniques. Indeed, one of the studied multichannel format conversion techniques exploited the cross-correlation measures, which are used in multichannel audio coding, to successfully extract primary signal components from stereo audio signals. This technique seemed promising to be generalized for the modification of multichannel content with more channels as well. The technique made an initial assumption about equal power levels of channel-specific independent signals, but the assumption did not appear to scale well for audio content with more than two channels, however.

Preliminary knowledge about the actual multichannel content was gathered by analyzing DVD movies and concerts. The means of analysis were cross-correlation coefficients and instantaneous power measures. The correlation values showed that generally the frontal channels of five-channel surround systems had common amplitude panned signals. The two rear channels had common amplitude panned signals between each others, but rarely shared signals with the frontal channels. The power measurements clarified that there is a long-term balance between the power levels of left and right half-planes. The investigation of more delicate interchannel relationships was left for future work. These relationships can be for example panning using time-shifts or other convolutive mixing methods. Solving these relationships requires also more complex channel similarity measures, which would help in transformation of multichannel audio signals as well.

Developing a transformation technique for multichannel audio proved to be significantly more complex task than the transformation of stereo signals. Therefore, the main effort in the listening experiments was put into studying the applicability of the present day audio format conversion techniques for the playback of two-channel stereo signals from a compact loudspeaker system. The first listening test experimented with different compact loudspeaker systems. These were realized virtually by first recording binaural impulse responses and then constructing various loudspeaker layouts by the means of convolving the input signals with the impulse responses and then summing the convolved signals. This test did not show much differences between the virtual loudspeaker systems although it revealed that the differences in how the signals were played were much easier to distinguish from synthesized noise signals than from natural music signals. It was assumed that the virtualization was a factor that caused poor perceptual resolution of the differences between the input signals. Therefore, the next listening experiment was conducted using a real, albeit experimental, compact loudspeaker system. Three processing methods were tested as audio transform tools. It turned out that two of the systems could improve the spatial dimensions of the audio material played from a three-channel setup in comparison to the original signals played from a compact two-channel setup. The two-channel setup could be compared to a conventional beatbox also known as “ghetto blaster”. The result of the experiment denotes that existent stereo playback systems could be improved by adding a third loudspeaker and using spatial processing.

The present day blind signal separation techniques are getting effective and fast in terms of computational power, and they could be applied to multichannel audio transformation processes that are performed in real-time. The goal of the BSS techniques is to perfectly reconstruct original signals that form the mixture. This goal requires that the number of signals to be extracted is equal to or less than the number of mixture signals. This is rarely a valid assumption for the multichannel audio signals. There are less strict requirements for a new group of source separation methods that is called sparse component analysis. These can prove as promising spatial transformation tools for future multichannel audio methods.

Bibliography

[1] C. Avendano and J.-M. Jot. A frequency-domain approach to multichannel upmix. *Journal of the Audio Engineering Society*, 52(7/8):740–749, July/August 2004.

[2] F. Baumgarte and C. Faller. Binaural cue coding - part I: Psychoacoustic fundamentals and design principles. *IEEE Transactions on Speech and Audio Processing*, 11(6):509–519, November 2003.

[3] B. Bernfeld. Simple equations for multichannel stereophonic sound localization. *Journal of the Audio Engineering Society*, 23(7):553–557, September 1975.

[4] J. Blauert. *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, 1999.

[5] J. Blauert. Analysis and synthesis of auditory scenes. In J. Blauert, editor, *Communication Acoustics*. Springer, 2005.

[6] A. D. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. U.K. Patent 394.325, December 1931.

[7] J. Breebaart and C. Faller. *Spatial Audio Processing: MPEG Surround and Other Applications*. Wiley, 2007.

[8] M. Briand, D. Virette, and N. Martin. Parametric representation of multichannel audio based on principal component analysis. In *120th Convention of the Audio Engineering Society*, May 2006. Convention Paper 6812.

[9] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994. Special issue on Higher-Order Statistics.

[10] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[11] P. G. Craven. Continuous surround panning for 5-speaker reproduction. In *AES 24th International Conference: Multichannel Audio, The New Reality*, June 2003.

[12] P. G. Craven and M. Gerzon. Coincident microphone simulation covering three dimensional space and yielding various directional outputs. US Patent 4.042.779, August 1977.

[13] C. Faller. *Parametric coding of spatial audio*. PhD thesis, École Polytechnique Fédérale De Lausanne, 2004.

[14] C. Faller. Multiple-loudspeaker playback of stereo signals. *Journal of the Audio Engineering Society*, 54(11):1051–1064, November 2006.

[15] P. Georgiev, F. J. Theis, A. Cichocki, and H. Bakardjian. Sparse component analysis: a new tool for data mining. In P. M. Pardalos, V. L. Boginski, and A. Vazacopoulos, editors, *Data Mining in Biomedicine*, chapter 6, pages 91–116. Springer, 2007.

[16] M. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, January/February 1973.

[17] M. Gerzon. General metatheory of auditory localisation. In *92nd Convention of the Audio Engineering Society*, March 1992. Preprint 3306.

[18] M. Gerzon. Panpot laws for multispeaker stereo. In *92nd Convention of the Audio Engineering Society*, March 1992. Preprint 3309.

[19] M. Goodwin and J.-M. Jot. A frequency-domain framework for spatial audio coding based on universal spatial cues. In *120th Convention of the Audio Engineering Society*, May 2006. Convention Paper 6751.

[20] M. Goodwin and J.-M. Jot. Analysis and synthesis for universal spatial audio coding. In *121st Convention of the Audio Engineering Society*, October 2006. Convention Paper 6874.

[21] M. Goodwin and J.-M. Jot. Multichannel surround format conversion and generalized upmix. In *AES 30th International Conference*, March 2007.

[22] M. Goodwin and J.-M. Jot. Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2007.

[23] D. Hammershøi and Henrik Møller. Binaural technique – basic methods for recording, synthesis and reproduction. In J. Blauert, editor, *Communication Acoustics*. Springer, 2005.

[24] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, W. Oomen, K. Linzmeier, and K. S. Chong. MPEG Surround - the ISO/MPEG standard for efficient and compatible multichannel audio coding. *Journal of the Audio Engineering Society*, 56(11):932–955, November 2008.

[25] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, 2001.

[26] J. O. Smith III and J. Abel. Bark and erb bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7(6):697–708, November 1999.

[27] R. Irwan and R. M. Aarts. Two-to-five channel sound processing. *Journal of the Audio Engineering Society*, 50(11):914–926, November 2002.

[28] ITU-R BS.775-2: Multichannel stereophonic sound system with and without accompanying picture, July 2006. International Telecommunication Union Radiocommunication Assembly.

[29] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[30] H. Kuttruff. *Room Acoustics*. Taylor & Francis, 4th edition, 2000.

[31] C. Kyriakakis and R. Sadek. A novel multichannel panning method for standard and arbitrary loudspeaker configurations. In *117th Convention of the Audio Engineering Society*, October 2004. Convention Paper 6263.

[32] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, February 2000.

[33] Y. Li, A. Cichocki, and S. Amari. Sparse component analysis for blind source separation with less sensors than sources. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003.

[34] J. Merimaa, M. Goodwin, and J.-M. Jot. Correlation-based ambience extraction from stereo recordings. In *123rd Convention of the Audio Engineering Society*, October 2007. Convention Paper 7282.

[35] J. Merimaa and V. Pulkki. Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, December 2005.

[36] B. Moore. *An Introduction to the Psychology of Hearing*. Emerald Group Publishing, 5th edition, 2003.

[37] P. Novo. Auditory virtual environments. In J. Blauert, editor, *Communication Acoustics*. Springer, 2005.

[38] A. Papoulis and S. U. Pillai. *Probability, random variables and stochastic processes*. McGraw-Hill, 4th edition, 2002.

[39] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[40] V. Pulkki. Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning. *Journal of the Audio Engineering Society*, 49(9):753–767, September 2001.

[41] V. Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. In *AES 28th International Conference*. AES, June 2006.

[42] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, June 2007.

[43] V. Pulkki and C. Faller. Directional audio coding: filterbank and STFT-based design. In *120th Convention of the Audio Engineering Society*, May 2006. Convention Paper 6658.

[44] V. Pulkki and M. Karjalainen. Localization of amplitude-panned virtual sources I: Stereophonic panning. *Journal of the Audio Engineering Society*, 49(9):739–752, September 2001.

[45] F. Rumsey. *Spatial Audio*. Focal Press, 2001.

[46] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross. Objective measures of listening envelopment in multichannel surround systems. *Journal of the Audio Engineering Society*, 51(9):826–840, September 2003.

[47] M. Sugawara. Super hi-vision - research on a future ultra-HDTV. Technical report, European Broadcasting Union, 2008.

[48] I.B. Vapnyarskii. Lagrange multipliers. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics*. Springer, 2002. Available online at http://eom.springer.de.

[49] E. W. Weisstein. Correlation coefficient. From MathWorld–A Wolfram Web Resource. Wolfram Research, Inc. [online]
Available: http://mathworld.wolfram.com/CorrelationCoefficient.html (accessed 25 April 2009).

[50] E. W. Weisstein. Cross-correlation theorem. From MathWorld–A Wolfram Web Resource. Wolfram Research, Inc. [online]
Available: http://mathworld.wolfram.com/Cross-CorrelationTheorem.html (accessed 25 April 2009).

[51] J. R. West. Five-channel panning laws: An analytical and experimental comparison. Master's thesis, University of Miami, 1998.

[52] D. T. Yang, C. Kyriakakis, and C.-C. J. Kuo. *High-Fidelity Multichannel Audio Coding*. EURASIP Book Series on Signal Processing and Communications. Hindawi, 2006.

[53] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, April 2001.

[54] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, June 2006.