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

Marko Takanen

A Binaural Auditory Model for Evaluating Quality Aspects in Reproduced Sound

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

Espoo, Apr. 28, 2008

Supervisor:	Professor Matti Karjalainen
Instructor:	M.Sc. Gaëtan Lorho

HELSINKI UNIVERSITY OF TECHNOLOGY		ABSTRACT OF THE MASTER'S THESIS

Author:	Marko Takanen	
Name of the thesis:	A Binaural Auditory Model for Evaluating Quality Aspects in Reproduced Sound	
Date:	Apr. 28, 2008	**Number of pages:** 89+9
Faculty:	Electronics, Communications and Automation	
Professorship:	S-89	
Supervisor:	Prof. Matti Karjalainen	
Instructor:	Gaëtan Lorho, M.Sc.	

Binaural cues describing the differences between signals in the left and righ ears, in terms of phase and power, enable our auditory system to localize and segregate sound sources spatially even in the presence of multiple overlapping sound stimuli. Recent publications and binaural auditory models have illustrated how interaural coherence can be used to estimate these cues and thus model the capability of our auditory system to localize sounds.

In this Master's thesis this approach is developed further and a new binaural auditory model is presented. The model is built on some of the existing auditory models. The aim is to use the model to evaluate binaural recordings of reproduced sound in terms of spatial and timbral aspects.

The binaural cue estimation is based on the cross-correlation model by Jeffress and the binaural cues are estimated in this model by taking into account the frequency selectivity of the peripheral hearing. The purpose of this approach is to localize sound sources from a broadband signal and to evaluate the spatial aspects based on these localizations.

Composite loudness level spectra are also calculated in this work by modeling the transfer functions of the peripheral auditory system. These spectra enable the analysis of the frequency balance from reproduced sound. Consequently, this Master's thesis illustrates the possible application of a binaural auditory model to the analysis of reproduced sound in terms of loudness, timbral and spatial aspects.

Keywords: Spatial sound, Timbre, Binaural hearing, Auditory Model, Binaural cues, Cross-correlation model

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Marko Takanen		
Työn nimi:	Binauraalisen kuulon mallin käyttö toistetun äänen laadullisten parametrien arvoinnissa		
Päivämäärä:	28.4.2008	**Sivuja:**	89+9
Tiedekunta:	Elektroniikka, tietoliikenne ja automaatio		
Professuuri:	S-89		
Työn valvoja:	Prof. Matti Karjalainen		
Työn ohjaaja:	DI Gaëtan Lorho		

Ihmisen kuulojärjestelmän kyky paikantaa äänilähteitä perustuu korviin saapuvien äänten välisten vaihe- ja tasoerojen analysointiin. Näiden binauraalisten vihjeiden avulla voimme erotella eri ääniähteiden sijainnit myös useiden samanaikaisten äänien läsnäollessa. Viimeaikaiset tutkimukset ja auditoriset mallit ovat osoittaneet kuinka nämä erot voidaan arvioida ristikorrelaation avulla ja kuinka täten voidaan mallintaa kuulojärjestelmämmme kykyä paikantaa ääniä.

Tässä diplomityössä esitellään tähän lähestymistapaan ja nykyisiin auditorisiin malleihin pohjautuva uusi binauraalinen kuulon malli. Työn tavoitteena on pystyä arvoimaan binauraalisesti nauhoitetun äänen tilavaikutelmaan ja väriin liittyviä ominaisuuksia kehitetyn mallin avulla.

Binauraalisten vihjeiden arviointi mallissa perustuu Jeffressin ristikorrelaatiomalliin, ottaen huomioon myös basilaarikalvon taajuuserottelukyvyn vaikutuksen äänilähteiden erottelukykyyn. Työn tavoitteena on tämän lähestymistavan avulla pystyä paikantamaan äänilähteitä laajakaistaisesta signaalista ja arvioida sitten äänen tilavaikutelmaan liittyviä ominaisuuksia näiden paikannusten avulla.

Tässä työssä esitettävässä mallissa nauhoitetusta äänestä lasketaan myös osaäänekkyystiheysspektri, jossa kuulojärjestelmän eri osien vaikutukset ääneen on huomioitu. Näitä spektrejä käytetään sitten nauhoitetun äänen äänekkyyteen ja väriin liittyvien ominaisuuksien arvioinnissa. Näin ollen tämä dipltomityö esittelee mahdollisuuden käyttää binauraalista kuulon mallia äänenlaadun arvointiin äänen tilavaikutelmaan, äänekkyyteen ja väriin liittyvien ominaisuuksien avulla.

Avainsanat: Tilaääni, Äänen väri, Binauraalinen kuulo, Kuulon mallinnus, Ristikorrelaatio malli, Binauraaliset vihjeet

Acknowledgements

This Master’s thesis has been done for Nokia Oy at Helsinki, Finland, and was carried out between May 2007 and April 2008. Firstly, I want to thank my coworkers at Nokia for creating the enjoyable atmosphere and for answering to all the (irrational) questions I came up with. Things could have been a lot more difficult without you.

Most importantly I wish to thank my instructor MSc. Gaëtan Lorho. This topic has been very challenging and because of that your advices and support have been essential during this journey. You were eager to help me in this process and thus managed to keep my motivation up also when things just did not go my way.

I also want to thank my supervisor Professor Matti Karjalainen. Your feedback and suggestions about the work, and the fruitful conversations we had during this project have been a vital help in this process.

I would also like to thank Mika Hanski from Nokia, Ville Pulkki from Helsinki University of Technology and Aki Härmä from Philips for their important advices. Ole Kirkevy and Eira Seppälä are thanked for the advices on the simulated HRTF database and Toshio Irino from Wakayama University is acknowledged for providing the gammachirp filter bank software.

I would like to express my gratitude to my family, as this could not have been accomplished without your support. Finally, I want to thank my friends. Your "supportive comments" during my studies have kept me going (for all these years).

Sievi, April 4, 2008

Marko Takanen

Contents

Abbreviations

BAM	Binaural auditory model
BMLD	Binaural masking level difference
BRIR	Binaural room impulse response
CF	Characteristic frequency
CLL	Composite loudness level
DRP	Eardrum reference point
ELC	Equal loudness contour
ERB	Equivalent rectangular bandwidth
FFT	Fast Fourier transform
FFTF	Free-field transfer function
FIR	Finite impulse response
GTFB	Gammatone filter bank
HATS	Head and torso simulator
HRIR	Head-related impulse response
HRTF	Head-related transfer function
IACC	Interaural cross-correlation
IC	Interaural coherence
ILD	Interaural level difference
ITD	Interaural time difference
MRI	Magnetic resonance imaging
NaN	Not a number
PEAQ	Perceptual Evaluation of Audio Quality
SPL	Sound pressure level

List of symbols

N'	Loudness in sones
L_L	Loudness level
L_p	Sound pressure level in dB
z	Critical band number
p	Sound pressure
p_0	Sound pressure level reference (20 μPa)
$N'(z)$	Specific loudness at the given critical band
$E'(z)$	Excitation energy at the given critical band
i	Sample number
x_1, x_2	Left and right ear input signal
γ	Interaural cross-correlation (IACC)
m	Time lag in IACC calculation and the maximum allowed ITD
α	Forgetting factor
T	Time resolution of IACC calculation
f_s	Sampling frequency (48 kHz)
τ	Estimated interaural time difference (ITD)
c_{12}	Estimated interaural coherence (IC)
ΔL	Estimated interaural level difference (ILD)
itd, ild	Unwidowed binaural cue estimates
k	Time instant in the original input signal
I	Length of the time window
j	Number of the time window
θ	Estimated source direction
$ITD_{\text{ref}}, ILD_{\text{ref}}$	Reference values for the binaural cues in the lookup table.
θ_T	Perceived direction of sound source
θ_0	Direction of loudspeaker
g_1, g_2	Gain of the left and right loudspeaker
$P(\theta)$	Probability of sound presence in the given azimuth direction

List of Figures

List of Tables

Chapter 1

Introduction

Sound quality is a concept that has been the interest of audio research for years. Despite the efforts, no (precise) definition for high quality sound has been discovered. In sound reproduction over loudspeakers, perfect reproduction of the original sound would naturally lead to perfect quality, but this is not feasible. Loudspeaker manufacturers usually aim for high quality of their device by trying to get the magnitude response of their loudspeaker as flat as possible, as the flat magnitude response would indicate that the original sound is reproduced transparently. The evaluation of the magnitude responses of the loudspeaker therefore gives the first 'measure' of quality of the devices and is consequently the approach used in the media for evaluating the quality of loudspeakers.

It is however the listener who finally determines whether the device reproduces a high quality sound or not, and the human auditory system does not evaluate the quality based only on the magnitude response. Because of this, the developed devices (loudspeakers) are used in a consumer study to compare the quality of the sound that they provide in comparison to the competitors before the device is released. There are however some things one needs to take into account when organizing these studies. The quality of sound is a highly perceptual concept, which means that different people consider different aspects of sound to have more influence on the overall quality and therefore the opinions about the quality vary between people. This has been shown for instance in the studies by Staffeldt (1974), Toole (1986), Zielinski et al. (2002), Moore & Tan (2003) and by Lorho (2006, 2007). For example good spatial reproduction can be a very important aspect of quality for some listeners, whereas other listeners might not care whether the spatial aspects are reproduced well or badly. Hence one gets as many different opinions about the quality as there are listeners in the study. This makes the organizing of these consumer studies time-consuming and expensive for the companies, because of need to have a large number of test subjects in the study in order to obtain reliable results.

Consequently, auditory models provide a tempting alternative, as with them one can evaluate the quality of sound "objectively", straight from a computer simulation. Due to the interest, various models have been presented over the years and an overview of the different models can be found in the publication by Rix et al. (2006). The first approaches for this goal were presented in literature by Schoeder et al. (1979) and by Karjalainen (1985), and currently there exists a standardized auditory model by Thiede et al. (2000) for evaluating the perceived audio quality. This standardized model, known as the PEAQ (Perceptual Evaluation of Audio Quality), evaluates different attributes of sound and then returns an overall quality measure of the quality. The aim of this thesis work is however a bit different.

The goal of this thesis is to develop a binaural auditory model (BAM) that can be used to evaluate the sound in terms of characteristics relating to spatial and timbral aspects. This approach is selected to illustrate the fact that people judge the quality based on different aspects, and that different loudspeaker systems focus on different aspects of sound reproduction. Consequently, the perceived overall sound quality is not evaluated. Hence the aim is to develop a model that can allow greater variation between the evaluated devices and that would provide more information about the reproduced sound than an overall quality measure.

1.1 Structure of this thesis

This thesis begins with a literature review in Chapters 2 and 3. In those chapters the human auditory system is described in terms of the aspects that are relevant to perception of spatial and loudness aspects of the sound. The developed binaural auditory model (BAM) is presented in Chapter 4. In that chapter the peripheral structure and functionality of the BAM is described and some comparison to the existing auditory models is also presented. The Chapters 5 and 6 focus then on testing of the model with different stimuli. In Chapter 5 the model functionality is tested in anechoic conditions and in Chapter 6 the model is used in a case study to evaluate differences between mobile loudspeakers. Conclusions and some suggestions for the future work are then presented in Chapter 7.

Chapter 2

Human hearing

We need information about our environment in order to cope in our everyday life. This information helps us to react properly to different events and to receive feedback of the effects of our actions in the environment. The gathering of this information is then handled by the cooperation of out senses (vision, hearing, smell, feeling and taste), where each sense observes the environment by the skills and capacities that it possesses. These observations are then passed to the brain for post-processing. As one can easily imagine, the information that different senses can pick out from our environment differ both in nature and accuracy. So the brain's task is to form an overall view from these observations. In most cases this operation works smoothly, as the observations are supportive to each other and there is only little overlap between them, but sometimes different senses provide information that is not compatible with the others, like in the McGurk effect[1]. In these cases, the most prominent information is selected. This means that the sense that is most certain of its observation is selected as an "eyewitness" and the overall view is formed based on the information that it has provided, even if this observation is providing false information about the actual stimulus.

The visual system being the most accurate of our senses (in 3-D localization of stimuli) normally dominates these observations and the role of other senses is to provide supporting information to get the overall view as complete as possible. Vision has however its own limitations. Even if a person has "normal vision", it cannot work accurately in too dark (or too bright) ambient lighting conditions or if the distance between the stimulus and the observer is too long. Also, it cannot tell anything about the stimulus, if there are obstacles (such as walls) blocking the direct line of sight between the observer and the stimulus or if the stimulus is located behind our head.

[1] An effect first noticed by McGurk & McDonald (1976), where they discovered that conflicting audio-visual information leads to perception that differs from both audio and visual stimuli (Riederer 2005).

One can however rest assured, as we are able to detect a great deal of our environment even without our eyes, just by listening. This is easy to prove by closing your eyes. When you are sitting at your desk, you can hear and detect sounds such as the footsteps of a person walking by your desk, the humming of the air conditioner, the sound of typing coming from your colleague's computer and so on. This ability to detect objects that are not visible is very important, since we can many times hear the sound source coming towards us before we can actually see it. This gives us more time to react since we do not have to wait until the source appears in our range of vision. For other animals this extra reaction time is essential for their survival, as it allows them to react to the presence of a predator (or prey) moving in their surroundings.

For humans though, the most important task of hearing is most probably its role in our everyday communications by speech with other people. Although hearing as such is not an absolutely necessary requirement for successful communication, as one can use other senses (vision and feeling) to compensate the lack of hearing, being able to hear facilitates these communications greatly. We use hearing also for recreation purposes when we listen to music. So it is easy to come to the conclusion that hearing is one of the most important senses that we possess. This chapter explains how the sound stimuli are perceived, by starting with the presentation of the structure and functionality of the human auditory system in Section 2.1. Some properties of human hearing are presented in Section 2.2 and Section 2.3 presents some of the developed auditory models for human auditory system. A short summary of the chapter is also given in Section 2.4.

2.1 Structure and functionality of human auditory system

"The peripheral part of the auditory system of most mammals is rather similar" (Moore 1997) but they differ in size of different parts, in accuracy of hearing in different frequencies and in the range of frequencies they hear. Although the human auditory system is not the most skilled of them, it is still equipped with stunning capabilities, since we can hear and distinguish sounds with frequencies ranging from approximately 20 to 20 000 Hz and sound pressure differences from 20 μPa to 63 Pa (Rossing, Moore & Wheeler 2002). These are quite remarkable differences, when one takes into consideration that the human eye, despite its superior accuracy in intensity and localization, is only capable to detect light waves whose wavelength range from about 400 to 700 nm (Goldstein 2002).

The reason for this remarkable ability of hearing lies in the seamless cooperation between the anatomy (Figure 2.1) of the ear and the neural processing in the brain. Based on the way the sound travels in the different parts, the peripheral hearing can be divided into three parts (external ear, middle ear and outer ear). The following sections explain the structure and functions of the three different parts.

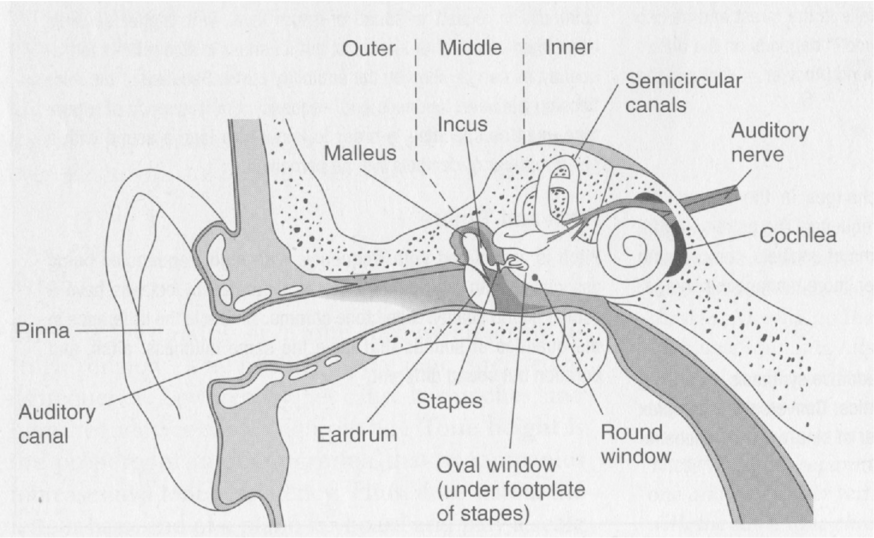


Figure 2.1: Cross-section of structure of human ear (Adapted from Goldstein (2002)).

2.1.1 External ear

The external ear consists of two parts, the pinna and the auditory canal extending from the pinna to the tympanic membrane (eardrum), which is the 'border' between the external and the middle ear. The pinna, being the only visible part of the human ear, is what we mean when we talk about the ear in our everyday communications. Traditionally the effect of the pinna on hearing has been considered to be rather insignificant, since its main task is to collect the incident sound waves from different angles to the auditory canal. Recent studies have however shown that its role in spatial hearing is quite remarkable as we will see later in Chapter 3.

From the pinna, the sound waves travel to the auditory canal, which is a tube-like structure (length of about 22.5 mm and diameter of about 7.5 mm in average (Goldstein 2002). Besides transmitting sound waves to the eardrum, the auditory canal has also an effect on the sound, as it amplifies some frequencies by the means of acoustical resonance. The physical structure of the auditory canal allows however only longitudinal waves to move in the canal, so the direction of coming sound wave does not influence the transfer function of the auditory canal (Karjalainen 1999). Hence the auditory canal has no or only little influence on sound source localization and the localization cues are based mostly on the pinna, head and rest of the torso. More information about the effect of the auditory canal can be found in Hammersøi & Møller (1996).

2.1.2 Middle ear

The difference in densities of media between the external ear and the cochlea in the internal ear causes problems. In the external ear the sound waves travel in air as air pressure variations and as these pressure variations "collide" with the tympanic membrane, they set it into vibration. Inside the cochlea the material is however liquid, which has much higher density. Hence the small differences in air pressure have to be amplified, in order for them to be able to pass the oval window into the cochlea (Moore 1997). This task is handled by the middle ear.

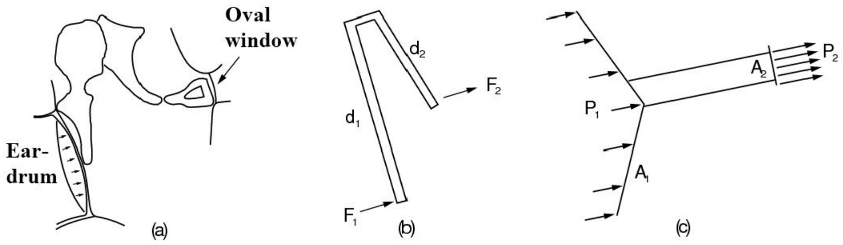


Figure 2.2: Structure of middle ear (a) and basis of its amplification effect explained with lever difference in ossicles (b) and difference in effective area between eardrum and oval window (c) (Adapted from Karjalainen (1999)).

The middle ear itself consists of ossicles (malleus, incus and stapes, the smallest bones in human anatomy) and of minute muscles (smallest muscles in human anatomy), which are attached to the ossicles. The bones are in contact with each other and the footplate of the malleus is attached to the tympanic membrane. The footplate of the stapes is attached to the oval window so when the sound wave sets the tympanic membrane into vibration, the vibration is transmitted through the bones to the oval window and amplified on the way (the amount of reflections is also reduced in the process). "This is accomplished mainly by the difference in effective areas of the eardrum and the oval window and to a small extent by the lever action of the ossicles" (Moore 1997) (Figure 2.2). This amplification is not however linear throughout the whole hearing range, as the transfer function is most effective around 500 Hz to 1200 Hz (Figure 2.3).

The middle ear has also a protective function, as in the presence of too intensive sounds, the minute muscles (stapedius muscle) suppress the transfer capacity of the ossicles by making them stiffer. Thus this effect protects the ear from damaging sounds. This function is called the stapedius reflex (named after the muscle) and more information about it can be found in Goldstein (2002).

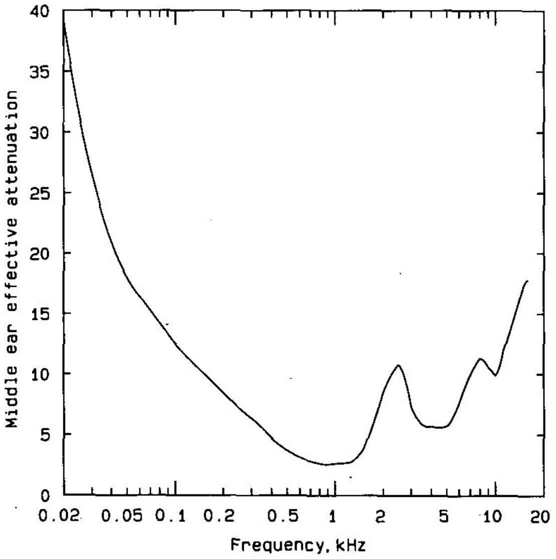


Figure 2.3: Transfer function of middle ear, plotted as effective attenuation versus frequency (Adapted from Moore et al. (1998)).

2.1.3 Inner ear

The main parts of the inner ear are the semicircular canals and the cochlea. From these two parts, the semicircular canals do not contribute to hearing as such, as its main task is to provide balance information to the brain by detecting changes in the horizontal-vertical position of the body (Rossing et al. (2002)). The cochlea on the other hand has an important role in hearing perception, and understanding its functionality tells a lot about the capabilities and limitations of our hearing system. The bony, snail-like structure of the cochlea makes it however hard to visualize. So we get a better idea on how the sound waves propagate inside it by looking at its cross-section (Figure 2.4) and picturing the cochlea as a straightened tube.

The cochlea itself is divided into three separate liquid-filled chambers throughout its length by three membranes, the Reissner's membrane, the tectorial membrane and the basilar membrane (Figure 2.4). The chambers are known as the scala media, the scala vestibuli and the scala tympani, the latter ones being connected through the helicoterma in the apex (Karjalainen 1999) (Figure 2.4). From the hearing point of view, the basilar membrane is the most interesting one, since its vibration triggers the neural activity leading to hearing sensation.

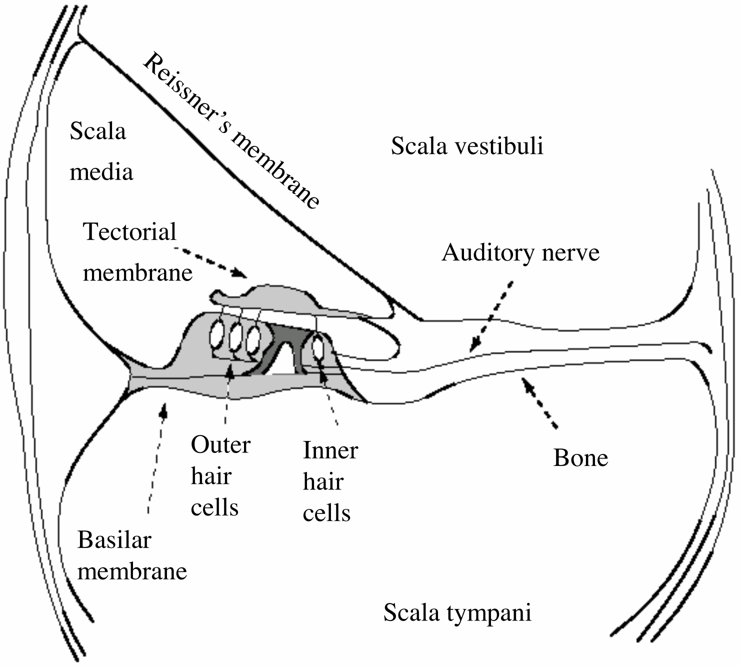


Figure 2.4: Cross-section of cochlea (Edited from Karjalainen (1999)).

The basilar membrane is elastic by nature and its mass and width vary with position (Karjalainen 1999). On top of the basilar membrane lays the delegate organ of Corti. The organ of Corti itself consists of two different types of receptors (the inner and outer hair cells), which are connected to the auditory nerve fibres by their roots and to the tectorial membrane by the fine cilia (Figure 2.5) (Goldstein 2002).

What this all means in practice is that when the stapes vibrates against the oval window, it creates pressure waves to the liquid in the scala vestibuli. These pressure waves then start to move towards the apex and through the helicoterma to the scala tympani and back to the middle ear through the round window. On the way, these pressure waves create ripples in the basilar membrane. This so called traveling wave was first discovered by Hungarian scientist Békésy[2]. He consequently confirmed the idea presented earlier by von Helmholtz that high frequency sounds generate a peak in the basilar membrane oscillation near the oval window and that low frequency sounds on the other hand generate theirs near the helicoterma in the apex. This place principle thus states that the frequency separation of hearing is made in the inner ear as different frequencies generate the greatest energy concentration (peak of oscillation) on different places along the basilar membrane (Zwicker & Fastl 1999).

[2] Békésy, György (1899-1972), a Hungarian scientist who received the Nobel Prize in Physiology or Medicine in 1961 for his research on the function of the cochlea (Goldstein 2002).

Békésy's research results were not however totally accurate, since in his research he could only observe the functionality of the cochlea taken from human cadavers and some of its functions stop working at death. This leads to poorer accuracy of the basilar membrane selectivity in Békésy's studies compared to more recent studies, which are made by observing the functionality of the cochlea in living animals. More detailed information about these studies can be found in Moore (1997).

This frequency-dependent movement of the basilar membrane does not explain how we are able to recognize and distinguish different sound events based on the frequency content of the sound stimulus, even if the place of the highest peak in the basilar membrane movement would be the same in two different sounds. As in perception of all senses, the post-processing (and the recognition) of the stimulus is done in different regions of the brain. This means that the basilar membrane movement has to be converted to electrical signals and then transferred to the brain in order to the sound event to be "registered". This encoding and transmitting is made in the organ of Corti (Figure 2.5), which stops working at the time of death (also exposure to damaging sound can cause this).

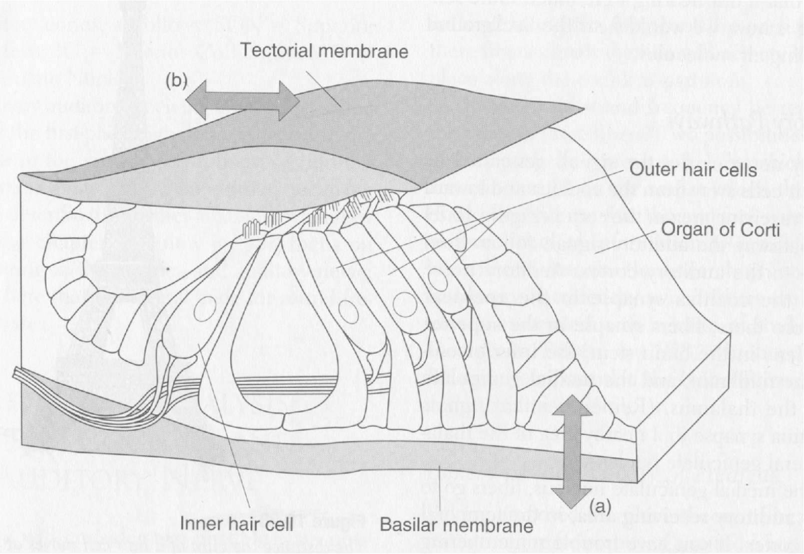


Figure 2.5: Structure of organ of Corti (Adapted from Goldstein (2002)).

When the basilar membrane moves up and down in different positions along its length, it causes the organ of Corti also to move up and down. At the same time, this movement causes the tectorial membrane to move back and forth relative to the cilia of the hair cells (Figure 2.5). As a result of these movements, the cilia of the inner hair cell bends, generating electrical signals in the hair cell. These signals are then transmitted via the auditory nerve fibres through different nuclei to the primary auditory cortex in the brain. This way the brain receives information about both the frequency (position of hair cell in the basilar membrane) and the power (amplitude of the electrical energy) of the sound, as the firing rate of the inner hair cell is dependent on the amount of bending of the cilia, which on the other hand is dependent on the amplitude of the movement in the basilar membrane. (Goldstein 2002, Karjalainen 1999).

Recent studies have shown that the signal is then passed from the primary auditory cortex to different regions in the brain for higher level processing (such as pattern recognition, word recognition, etc.) of the sound event. These processes are however still mostly unknown, due to the difficulty of knowing what different regions of brain analyze from the signal and what is their role in the whole hearing perception.

2.2 Some aspects of hearing

Due to the physical structure of the peripheral hearing, the transfer functions of different parts of it are not the same for all frequencies, as we noticed in the amplification of the middle ear and in the elasticity variation in the basilar membrane. Thus the limits and accuracies of hearing are not equal throughout the whole hearing range. The transfer functions for instance cause the required sound pressure for a sound to be heard to be different between different frequencies. These sound pressure limits are called hearing thresholds (Figure 3.8) and are measured in anechoic[3] conditions, with a test subject giving feedback whether he or she can hear the sound event played at given intensity.

2.2.1 Hearing selectivity

As the previously explained place principle states, the inner hair cells at different places along the basilar membrane all have their own characteristic frequency (CF), to which they are most sensitive to respond. A look at the envelope patterns (Figure 2.6) of the basilar membrane movement caused by different pulse sounds however reveals that the shape of the peak is not sharp. This would indicate that the adjacent inner hair cells would also fire, making the "frequency identification" impossible, since there would be a lot of misinformation given to the brain from the hair cells. Luckily our hearing is equipped with methods to overcome this problem.

The first of these methods is the motile response of the outer hair cells to the vibration of the basilar membrane. This is initiated by the movement of the organ of Corti, to which the outer hair cells respond by pushing the basilar membrane upward and by so create the maximum of the basilar membrane vibration to that position (Goldstein 2002). As this so

[3] No echoes of the sound reflecting from surrounding materials in the environment as the structures absorb all the sound waves colliding with them.

called motile response is frequency-dependent, the peak becomes sharper, resulting in a better frequency selectivity.

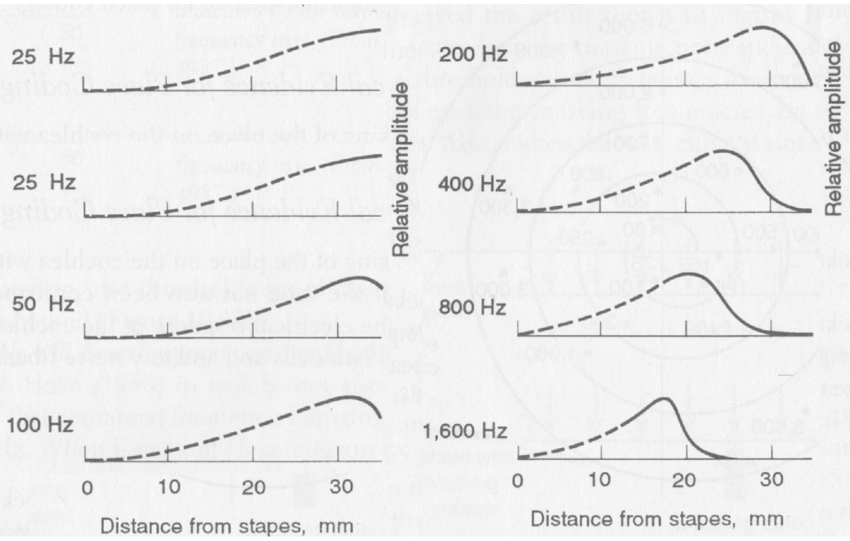


Figure 2.6: Examples of envelope patterns of basilar membrane's vibration with different frequency impulses as measured by Békésy (1960) (Goldstein 2002).

Another important aspect influencing the hearing selectivity is the phase locking of inner hair cells. This principle states that the inner hair cells fire only at a certain phase of the waveform of the basilar membrane vibration at its position (Moore 1997). This is due to the fact that the time instants when the inner hair cell fire are stochastically distributed and therefore different hair cells fire at different phases (time instants) of the basilar membrane movement. Consequenty, the waveform information of the basilar membrane vibration is also transfered to the brain, as well as the information about the main frequency content at given time. Both of these information pieces are necessary in sound localization, as we will see later.

There have also been studies on the effect of lateral inhibition on hearing perception, which can be described "as a suppression of neural activity at one place at the receptor field as a consequence of the stimulation of adjacent places in this field" (Houtgast 1971). In hearing this means that the firing of the inner hair cells at one position initiates feedback information from the higher levels of the auditory system to the hair cells at another position. This feedback information then prevents the inner hair cells at that position from firing. As the higher auditory levels also combine and analyze the information from the hair cells, they also have an effect on the accuracy and selectivity of the hearing. These functions and their impacts on hearing are however yet mostly unknown and disputed among researchers, so this work will focus on the lower level processes of hearing.

2.2.2 Masking

One very interesting feature of hearing concerns the masking of other sounds by a masker sound so that the target sound is not audible. This is a very common case in our everyday life, where the environmental sounds (noise) often increase the hearing threshold level for other sounds (such as speech). This can be easily proved by trying to talk with a friend in a library and in a bar using the same sound volume in both places. Besides the obvious problems this effect causes, it is also used beneficially for instance in music where a louder instrument (such as leading singer) masks the other instruments (such as a guitar) to be inaudible. When this instrument pauses, the other ones become audible again (Zwicker & Fastl 1999).

Since masking can occur when the masker and the masked sound are occurring either simultaneously or nonsimultaneously, it can be divided into two categories. In simultaneous masking, the masking effect level is mostly dependent on the frequencies of the sounds, hence the name frequency masking. This is at its maximum when the frequencies (energies) of the sounds are within the same critical band (Section 2.3.2), but can also occur when the frequencies differ more than that.

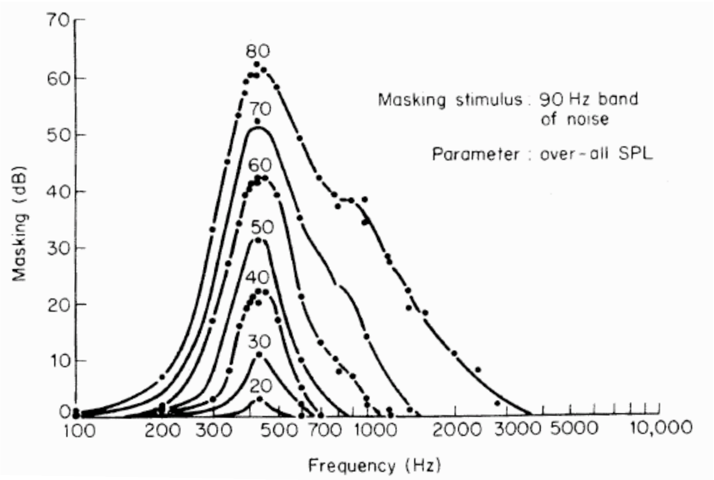


Figure 2.7: Effective frequency area and masking effect of 400 Hz masker tone (bandwidth 90 Hz) at different intensities (Adapted from Moore (1997)).

A look at the frequency masking pattern of a 400 Hz masker tone on different intensities (Figure 2.7) reveals that the tail of the envelope of a low frequency masker sound spreads wider and with more amplitude to the higher frequency areas than to the lower frequency areas. This means that low frequency sounds have more masking power towards high frequency sounds. Because of this, they can mask higher frequency sounds with lower level than what the high frequency sounds need to mask lower frequency sounds.

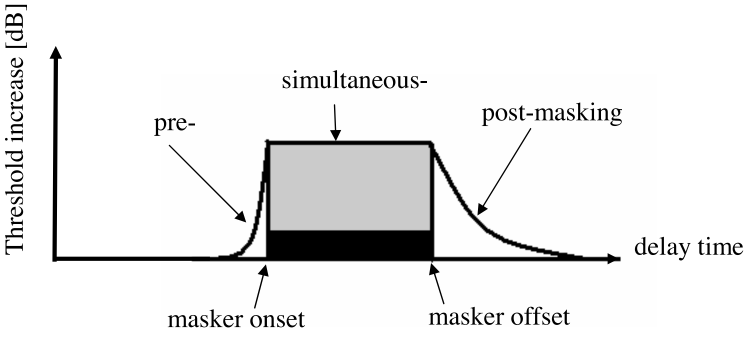


Figure 2.8: Temporal masking patterns and their effective ranges.

Nonsimultaneous masking on the other hand occurs, when a louder masking tone is played either just before (post-masking) or after (pre-masking) the quieter sound. In post-masking a masker tone raises the hearing threshold temporarily for following sounds and in pre-masking a masker tone masks sounds that were played just before the beginning of the masking tone (Zwicker & Fastl 1999). Of these two, post-masking has longer effective range in time (Figure 2.8). More detailed information about the measurements and literature references of masking levels at different cases can be found in Zwicker & Fastl (1999).

2.3 Models of human auditory system

Due to the interest of many different instances (such as structural architecture, music and movie industry, etc.) to know the capabilities and limits of human hearing, a need to model the human hearing as accurately as accurately as possible has always existed. In order to meet this demand, audio researchers have developed various models over the years to describe the functionality of the human auditory system. The first functional models concentrated on describing the whole auditory system as a frequency analyzer, in order to model the capabilities the hearing possesses in perceiving differences in for instance loudness and distortion in monaural hearing (Karjalainen 1985).

This kind of approach is still usable in some cases, but in order to model some properties of binaural hearing (Chapter 3) one cannot discard the phase information of the two inputs. So the modelling of transfer functions of the auditory system's different parts and the analysis of the sound has to be made in the time domain. This chapter will now look at the ways the different parts of the auditory system are modelled and used in recent auditory models.

2.3.1 External and middle ear modelling

Usually the functionality of the external and the middle ear is modelled by cascaded fixed filters, which shape the incoming sound by amplifying and attenuating the frequencies according to the amplification patterns in the external ear (effect of the pinna and the resonance of the auditory canal) and the impedance matching of the ossicles in the middle-ear (Figure 2.3). In these models, the effect of the pinna on the incident sound is estimated to be independent of sound source localization and is therefore modelled as a scalar transfer function. In these cases the transfer function of the pinna is usually measured in cases, where the sound source is located at the same level as the ear canal at zero elevation (Figure 3.2). More information about this kind of approach can be found in Moore et al. (1998).

As we will see in Chapter 3, this is not however the case in reality. This is why in some of the modern models (especially the ones addressing sound localization aspects) the effect of the pinna is modelled by head-related transfer function (HRTF) or binaural room impulse response (BRIR) (Section 3.3) measurements. The auditory system modelling starts then either at the beginning of the auditory canal or at the eardrum. The reason why this approach is not being used in all of the current auditory models, is the fact that the effect of the pinna can be estimated to be insignificant for example in ambient noise measurements.

2.3.2 Cochlear modelling

Psychophysical studies have shown that our hearing sums the energies of two different sounds (narrowband noise) together in the evaluation of loudness, if the frequencies of the two sounds are close to each other, and considers the energies separately if the frequency difference gets larger (Moore 1997). This phenomenon is a direct result of the physical properties of the basilar membrane, which also explains why these limits for frequency similarities known as critical bands are not same throughout the audible frequency range. At low frequencies the band has a width of around 100 Hz and at the highest frequencies it is more than 1 kHz wide (Karjalainen 1999). To take this property of hearing into account, the hearing range in the auditory models is usually divided into critical bands in Bark scale or into logarithmically divided Equivalent Rectangular Bandwidths (ERBs). Each of these bands is then modelled with at least one bandpass filter.

This is however the place where the similarity of the different cochlear models stops, as different models use different types of filters. The most commonly used approach is a gammatone filterbank (GTFB) developed by (Glasberg & Moore 1990) (Figure 2.9) where the filters[4] overlap a bit with each other, in order to describe the frequency masking in the basilar membrane (Glasberg & Moore 1990). This type of approach is however linear in the time domain, so it cannot describe the dynamic characteristics of the basilar membrane motion (Irino & Patterson 1997). That is why audio researchers have developed also some adaptive approaches, where the previous output power of the filter bank affects the shape of the filters. Hence the temporal masking effect is also taken into account. More information about this type of models can be found in the papers by Slaney (1988) and Irino & Patterson (1997). A comparison between different filters can be found in the publication by Unoki et al. (2006).

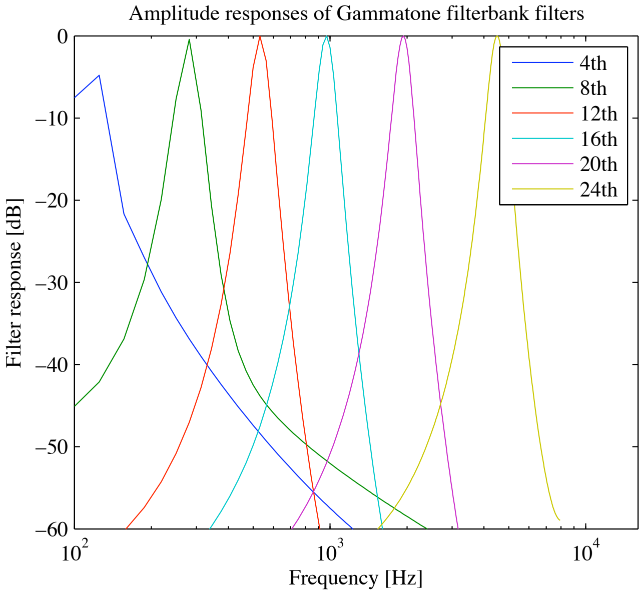


Figure 2.9: Amplitude response of gammatone filterbank filters in different critical bands.

The inner hair cell (and auditory nerve) activity is modelled in the auditory models by using either a physical or a functional approach. In the physical models the goal is to model each individual impulse accurately. The functional models on the other hand focus on modelling the pattern (envelope) of fired impulses. The impulse pattern is then calculated with different temporal windowing functions, which can also take care of the adaptivity of the model. The process in the functional models begins by half-wave rectification of the filterbank output. Then it is compressed by a given factor and convoluted with a temporal window function (and possibly passed also through a low pass filter) in order to produce a smoothed envelope of the filter bank output (Härmä 1998). In some functional models the

[4] Roex filters developed by Patterson et al. (1982).

temporal window is constant over time, but in the other models the temporal window has an onset time, during which it rises to the maximum value (saturation point), and an offset time, where it gets back to zero (Figure 2.10).

With this type of shape, the neural model is able to model forward masking and describe how the hair cell activity first rises in the presence of a new stimulus but descends afterwards, if the type and the level of the stimulus remain the same (Karjalainen 1999). The shape of the temporal window however differs between the different temporal models because the exact functionality of the hair cells is still to be resolved. Interested readers can find more information about the different approaches used in the models in the publications by Bernstein et al. (1999), Karjalainen (1996), Dau et al. (1996), Plack & Oxenham (1998) and by Härmä (1998).

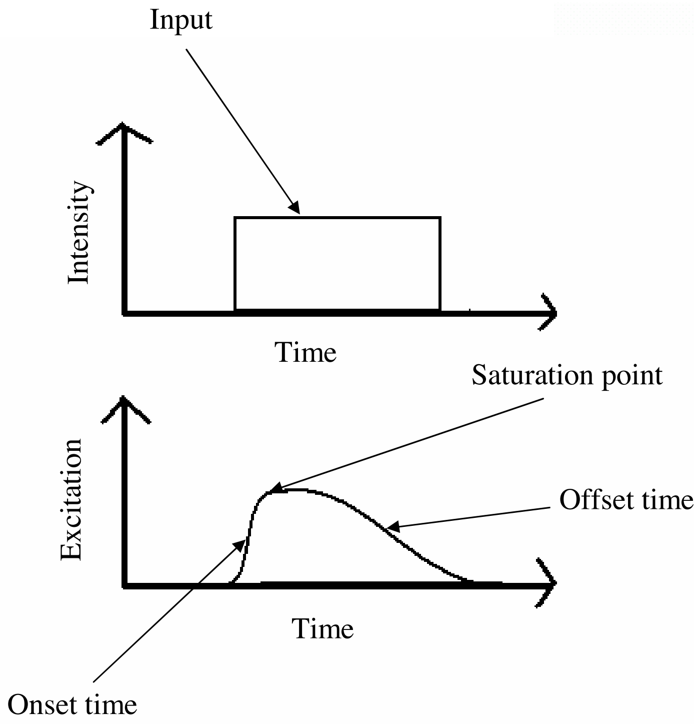


Figure 2.10: Output of functional neural activity model (with variable temporal window) (bottom) to imaginary input signal (above).

2.4 Summary

This chapter started by describing the structure of the human auditory system and showed how the sound waves are transferred through it and transformed in different phases before the sound stimulus becomes a perceived event in the brain. The chapter also presented some more detailed properties of human hearing capabilities and limitations. Some of the functional models of human auditory system's parts were also presented as well as the fundamental ideas behind these models.

When considering the models of hearing, one must however remember that the physical structure of the auditory system differs between people. For instance, the size and shape of the pinna and the length and the sangle of the auditory canal vary between people, both of which have an effect on the hearing perception. Besides this, the hearing thresholds for high frequency sounds rise when the person gets older or gets exposed to damaging sounds.

The higher levels of hearing perception, such as word recognition, are still largely unknown. Although MRI (Magnetic Resonance Imaging) scanning allows one to visualize how the neural pulses from the primal auditory cortex create activity also in various parts of the brain, the more precise role of these parts of the brain in the hearing process is yet to be resolved. This combined to the fact that the sensitivity of hearing (like other senses also) is greatly dependent both on the physical (age, health) and psychological (mood, alertness) condition of the person, one can see how complex task the modelling of the auditory system actually is.

One should not however give up, since by discarding some aspects of hearing (such as bone transduction[5]) and focusing on the aspects, which are relevant to the current case in the modelling, one still gets a much better idea about how people will perceive sound stimuli (or are they by any chance harmful to our health) than by simply recording the signal and applying frequency and amplitude analysis on it. There is still however a lot of work to be done.

[5] Low-frequency sounds can travel to the middle ear also by vibration through the bones in the skull.

Chapter 3

Binaural hearing

The fact that we have two ears makes our hearing more reliable in case of an accident since if one of them is "disabled" by some reason, we are still able to use the other one and hear almost as well as before. We do not have two ears just for backup reasons. It is beneficial in many ways to be able to use the information gathered by both ears, when we are to perceive a sound or different things about it. Studies on hearing threshold have for instance shown that the ability to use two ears to listen lowers the threshold for perceiving the presence of a target sound (speech, click tone) in presence of a masking tone (noise)[1]. The more difference there is in direction between the target sound and noise sound sources in the horizontal plane (Figure 3.2) the greater the effect.

Most importantly, binaural hearing is very useful and almost necessary in spatial sound localization. In vision, a person can use just one eye to perceive the relative positions of two or more stimuli accurately, since the images are plotted on different places along the retina (Figure 3.1). Also depth and distance estimation is relatively easy even with just one eye, because the person can use the relative heights of the stimuli with respect to height of the surroundings, and compare this information to the ones in his or her memory to estimate the distance. So the binocular (two-eyed) vision just makes these estimations even more accurate by using a triangulation method.

In hearing, however, the sounds from two or more sources are combined and mixed together before they enter the auditory canal (Figure 3.1), which makes the localization a lot more difficult and considerably less accurate in monaural hearing. By having the ability to rely on two ears (binaural hearing), our auditory system is able to overcome this mixing of information using the information from the two ears based on different cues.

[1] Binaural Masking Level Difference (BMLD) indicates that whenever the signals to the left and right ears are not the same, the target sound is easier to perceive in the presence of a masker tone than in the cases where the two signals are the same or when the signal enters only one of the ears. More information about this phenomenon can be found in Moore (1997).

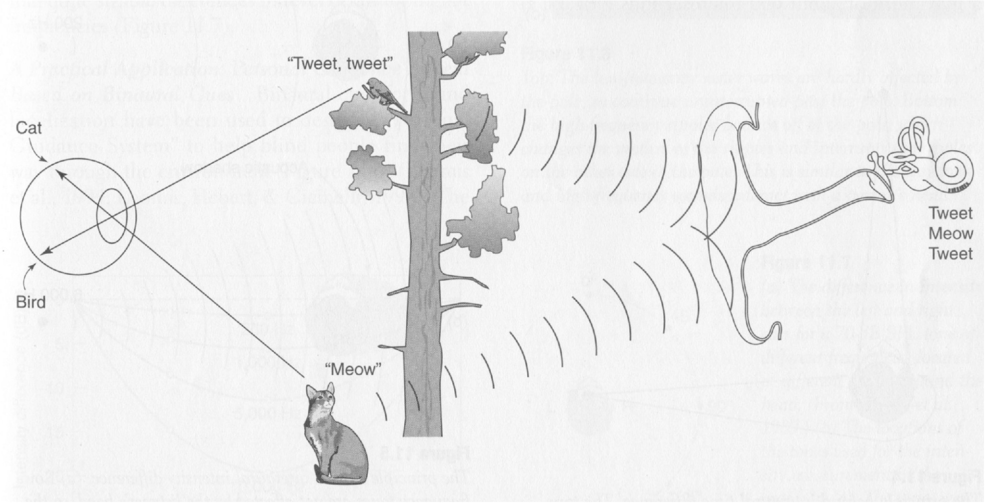


Figure 3.1: Images of two stimuli in vision and hearing (Adapted from Goldstein (2002)).

In this chapter these cues, their use in localization and their limitations are explained in Section 3.1. Some aspects of localization as well as the importance of other senses in auditory localization are also mentioned in Section 3.2. The measurement of head-related transfer functions (HRTF) and binaural room impulse responses (BRIR) is also briefly explained in Section 3.3. As the purpose of this work is to evaluate the quality of reproduced sound also in terms of timbral aspects, the timbre of a sound and some of its characteristics are presented in Section 3.4. A short summary of the chapter is also given in Section 3.5.

3.1 Location of sound event

The relative location of a sound event around the listener can be described with the help of three variables: azimuth (defining the angle in the horizontal plane), elevation (defining the angle in the median plane) and distance. The auditory space defined by these variables is illustrated in Figure 3.2, where the origin of the variables is located at the centre of the head at the level of the entrance to the auditory canal. One must however remember that the coordinates of a sound event in this space are head-related and not stationary. This means that the coordinates change to match the movement of the head when the person turns or moves his or her head to a certain direction. A formal definition of absolute and relative coordinate system has also been defined. More information about this coordinate system can be found in the publication by Paavola et al. (2005).

The ability to localize sound events has been the interest of researchers for years and different theories of the fundamental reasons behind this ability have been suggested over time. The first consistent theory was presented by Lord Rayleigh (1907). His research results indicated that low-frequency tones are analyzed based on the interaural time difference (ITD) and the high-frequency tone analysis is based on the interaural level difference (ILD). These two differences are called the binaural cues and the next section explains their origins and limitations.

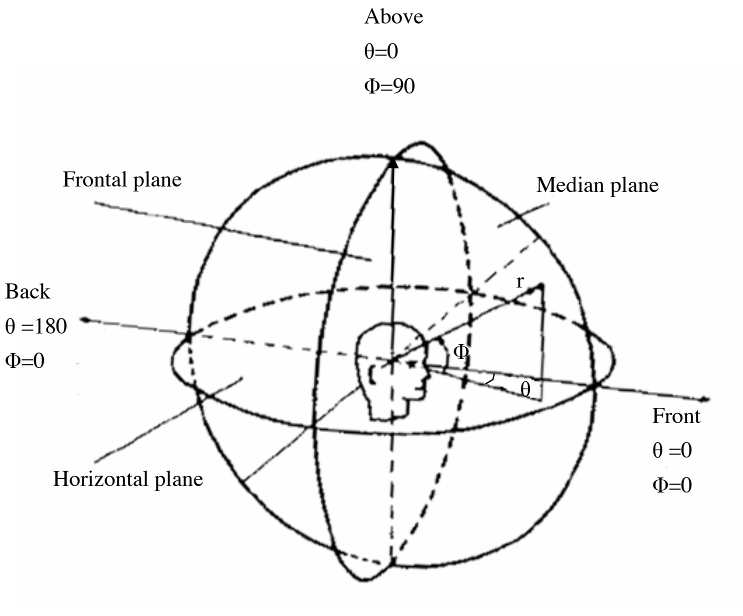


Figure 3.2: Auditory space defined by head-related coordinates (Edited from Blauert (1997)).

3.1.1 Binaural cues

When the sound source is located at side of the head (at certain azimuth angle), the head and the pinna create differences to the paths of the sound to the two ears. The first of these differences is in the length of the path, as the sound has to travel a longer path before it can enter the contralateral ear[2] compared to the one to the ipsilateral ear[3] . Hence the sound arrives later to the contralateral ear, resulting in ITD between the ears. Figure 3.3 illustrates this difference by showing the situation viewed from above. Here θ denotes the azimuth angle (Figure 3.2) and r is the radius of the head.

With continuous signals, this difference in time of arrival is harder to perceive. So the ITD analysis is based on the phase information. Due to the path length difference, the phase

[2] Contralateral ear refers to the ear that is on the opposite side of the head compared to the sound source.

[3] Ipsilateral ear refers to the ear that is on the same side of the head as the sound source.

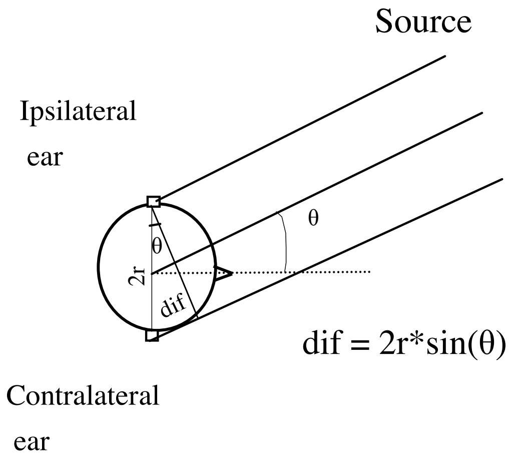


Figure 3.3: Difference in length of path to two ears that creates ITD.

of the signal is different in the ipsilateral and contralateral ears, and based on the delay in phase, the brain is able to estimate the ITD. As the frequency of the sound gets higher, the sound's wavelength or its multiple gets however closer to the difference in the path length. As a result of this, there may be a difference of multiple cycles between the phases of the signal in the two ears. This makes the phase analysis almost impossible. Therefore the ITD-based analysis works more consistently with low frequency sounds (below 1.5 kHz).

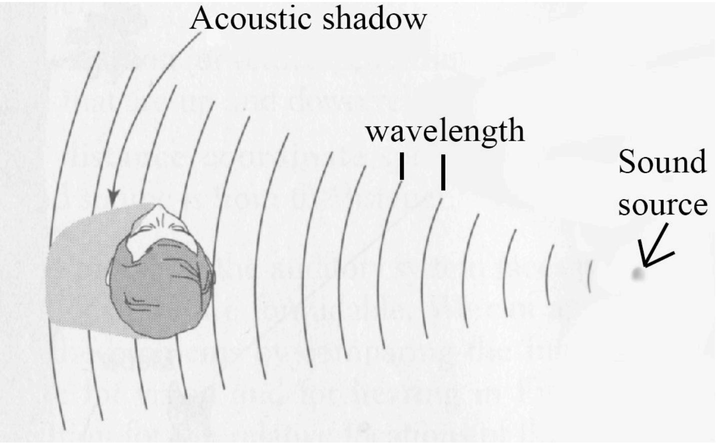


Figure 3.4: Acoustic shadow caused by the head (Edited from Goldstein (2002)).

The head also creates an acoustic shadow to the path of the sound to the contralateral ear. This shadow is the source of the interaural level difference (ILD), since this acoustical shadow attenuates the incident sound and thus lowers the remaining intensity of the sound, which is perceived in the contralateral ear. As the shape and position of this shadow and thus the amount of caused attenuation are dependent on the location of the sound, the brain can once again determine the localization of the sound source, based on this information. Figure 3.4 illustrates the forming of acoustic shadow by showing the situation, where the sound source is located directly at right of the listener.

This cue is however also frequency-dependent, since the wavelength of the sound has an effect on the size of the acoustic shadow and the wavelength is then dependent on the frequency of the sound. This means that with high frequency sounds, whose wavelength is small, many periods fit inside the acoustic shadow and thus more of the sound is attenuated. In the lower frequencies, the wavelength is however larger and so the sound 'bends' (as a result of diffraction) more smoothly around the head, and the amount of attenuation is therefore smaller (Moore 1997). This is why the ILD works best in high frequency areas (above 2 kHz). This confirms the duplex theory presented by Lord Rayleigh.

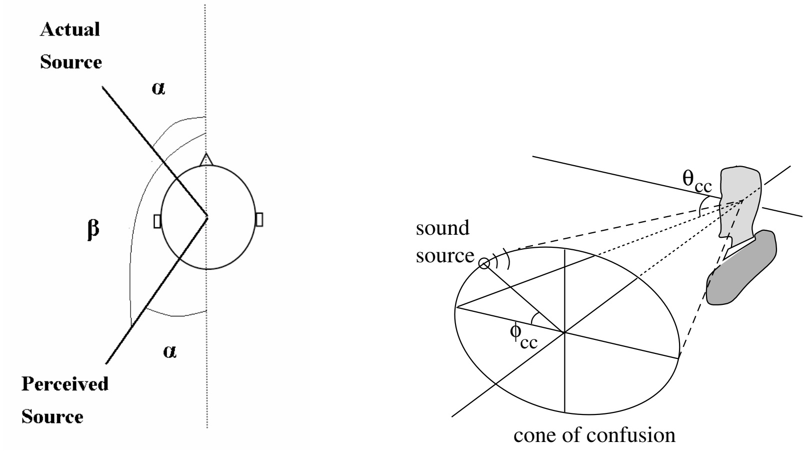
3.1.2 Spectral cues

The above-mentioned effects of the head on the path of the sound are mainly responsible for the binaural cue values. Only small extra information is obtained by the change of the transfer function on different angles due to the pinna. This is one of the reasons, why the pinna has been considered to be insignificant to hearing perception in the past (Section 2.1.1). The binaural cues do not however exist in the median plane (Figure 3.2), since the path of the sound to both ears is the same. This is why the elevation of the sound source must be perceived by using also other cues. This is where the change in the transfer function due to the pinna is one of the crucial cues. As mentioned before (Section 2.1.1), the pinna collects the sounds and transfers them to the auditory canal by reflecting and diffracting them via its different surfaces. As the reflections amplify and attenuate some of the frequencies and the path of the reflections is dependent on the elevation (and azimuth) of the sound location, the pinna has an effect to the sound entering the auditory system. This effect, together with reflections and attenuation of the sound caused by the head, are referred to as spectral cues (Goldstein 2002). Although the binaural cues (mostly ILD) also change a bit as the elevation changes, the spectral cues are mainly responsible for the perception of elevation of the sound source.

Since the binaural cues provide ambiguous information outside their strongest areas (for ITD below 1.5 kHz and for ILD above 2 kHz), the brain must analyze the cue values separately and compare the information these analyses provided in order to evaluate the location of the sound source. When the cues provide mismatching information, the task is more difficult. Studies with dichotic listening[4] on broadband signals have shown that ITD cues dominate small ILD differences in these situations if there is low-frequency content in the signal. On the other hand, large ILD values restrict the possible location area to be near the ear, even if the ITD values would provide other kind of information. More information of these studies can be found in Blauert (1997) and in Shinn-Cunningham et al. (2000).

3.1.3 Problems with binaural cues

The shape of the head can be estimated to be round, where the ears locate at its surface along the centre axis. This estimation and the small effect of the pinna on the binaural cue values cause the ITD and ILD values to be almost identical, weather the sound source is located in front or back of the listener at the relatively same azimuth angle (Figure 3.5(a)). The similarity of the binaural cue values creates problems in localization of a sound event, when the listener cannot see (Section 3.2.3) the sound source and move his or her head (Section 3.2.3).



(a)

(b)

Figure 3.5: a) Perception of sound source location at back due to front-back confusion and b) cone (torus) of confusion describing area where binaural cues are similar (Adapted from Pulkki (2001)).

In these situations, the person is likely to estimate the sound coming from back of the head, when the actual source is located at the front. The opposite relationship, between the perceived and actual location, is also possible. This tendency of mislocalization is referred

[4] Dichotic listening refers to situation in headphone listening, where signals to the left and right ear are not the same.

to as the front-back confusion, and one example of it is illustrated in Figure 3.5(a).

The second problem with binaural cues occurs, when the sound source is located at the right (or the left) of the listener at a certain distance. In these cases the binaural cue values are similar within a certain volume, which is denoted as the torus (or cone) of confusion (Figure 3.5(b)). Therefore the listener can only localize the sound to be somewhere within the torus (Shinn-Cunningham et al. (2000)). As Figure 3.5(b) shows, the uncertain area of localization grows as sound source gets further away from the listener. Hence the localization gets less accurate with longer distances.

3.2 More aspects of localization

In normal listening environments, objects present in the vicinity can absorb sound waves and reflect them back and towards the listener. Depending on the properties of the environment, the sound can reflect many times on different surfaces before reaching the listener ear (adding therefore to the complexity of the received signal) and still the sound will be audible when it arrives to the ear.

Thus the sound from the source can travel to the ears of the listener both directly and indirectly. The reflections of the sound arrive to the ears later than the direct sound and could therefore cause false localizations of sound sources. The summing localization and precedence effect help the localization process by time-domain analysis of the incident sounds. Due to these effects two similar signals that arrive within 30-40 ms time limit are perceived as one singular sound event whose perceived location is dependent on the delay between the arrivals. For instance in normal stereophonic listening situation, where loudspeakers A and B are positioned at 30° and -30° azimuth angles (Figure 3.6), the perceived sound source direction varies relative to the delay.

If the signals from two sources are the same and there is no delay between the signals, the localization cues cause the listener to perceive the presence of a single sound event locating at the centre (at zero azimuth). By adding a delay to the signal from loudspeaker B, the perceived direction of the sound source moves towards the loudspeaker A due to summing localization effect, and reaches that direction when the delay is larger than 1-2 ms. After this time limit the precedence effect starts to affect the perception. Even though more delay is added to the signal from loudspeaker B, the listener still perceives the presence of a single sound source at the position of the loudspeaker A. The perception remains the same until the delay exceeds the 30-40 ms limit, after which two separate sound sources are perceived at the directions of the loudspeakers A and B.

It is also possible to get the listener to perceive the sound source locating above the head, when the loudspeakers C and D are positioned directly at left and right of the listener at same distance (Figure 3.6) and there is no delay between the signals. More detailed information about summing localization, precedence effect and the different studies made in this field can be found in Blauert (1997).

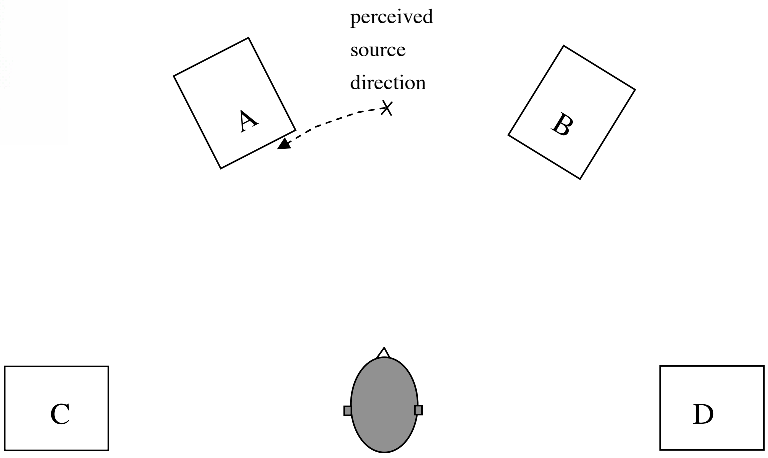


Figure 3.6: Perception of virtual sound sources in stereophonic listening over loudspeakers.

3.2.1 Accuracy of direction perception

Over the years there hav been several studies on the human ability to localize sound events. As the sound stimuli used (and the test subjects) in these tests vary between tests, the results obtained from these tests also differ (Karjalainen 1999). Naturally the way the listener points the direction of the sound source has also significant effect on the results. These differences being considered, the results show that the listeners can point the direction of the sound source within 10° accuracy both in horizontal and vertical directions (Makous & Middlebrooks 1990).

Another experiment covers the minimum change in the sound direction that the listener can perceive. The research results of this localization blur state that the listeners can detect as small as 1° changes in the direction when the source is directly at front of the listener and that the ability gets worse when the source is placed more to the side of the listener. More information about the different tests and the associated results can be found in Makous & Middlebrooks (1990) and in Blauert (1997).

3.2.2 Distance perception

The distance of sound source is however estimated with worse accuracy, which results from the ambiguity of the available cues. In the tests, loud sources are usually estimated to be close to the listener and quiet ones to be afar, because the loudness (Section 3.4.1) of the sound decreases as the distance increases. So the overall loudness is one of the cues for distance, but this cue can be misleading if a quiet source is positioned close to the listener and a loud source is positioned at a greater distance. Hence the estimation of distance based on loudness works accurately only if the listener has prior knowledge of the sound source and if the source is at least 1 meter away from the listener (Shinn-Cunningham 2000).

The second available cue relates to the frequency content of the sound. As the sound travels through air, the high frequencies are absorbed on the way more than the low frequencies because they carry less energy (Goldstein 2002). So, if the sound stimulus has more frequency content on low frequencies than on the higher ones, the sound source is perceived to be afar. This cue works however only with certain type of stimuli and requires also prior knowledge of used stimuli.

As the distance between the source and the listener increases, the amount of reflections of the sound also increases in non-anechoic conditions. Hence the relationship between the energies of the direct sound and its reflections can be used as a cue for distance (Goldstein 2002), especially with distant sources (Shinn-Cunningham 2000). This cue also requires some prior knowledge of the sound environment to work accurately (Karjalainen 1999).

Another available cue of distance with near sources is the change of ILD, since the closer the sound source is to the ipsilateral ear, the more signal is blocked from entering the contralateral ear by the acoustic shadow (Figure 3.4). So the ILD grows as the distance gets smaller. With more distant sources, the changes of ILD over distance are not anymore that drastic and therefore the ILD-based estimation is most accurate with near sources (Shinn-Cunningham 2000).

If the sound source is not stationary and moves for instance horizontally, the change of binaural cue values (ITD and ILD) can also be used to estimate the distance. At greater distances the changes are smaller so a slowly moving source is estimated to be afar and a faster moving source is estimated to be close. This can however be misleading since the prior knowledge of the listener about the stimulus has an effect on this estimation.

3.2.3 Role of other senses

The above-explained torus of confusion, front-back confusion and the problems with distance perception show that the human auditory system has difficulties to localize sounds accurately in complex listening situations. There is however supportive (and sometimes conflicting) information available from other senses to make this localization more precise. The most important source of information is naturally vision, since it has the greatest accuracy in three-dimensional localization.

If the listener is able to both hear and see the sound source simultaneously, the vision corrects the possibly wrong direction perception, caused by front-back confusion. Also our accuracy in perception of sound distance is greatly improved, when we are able to see the sound source (Riederer 2005). Vision can however provide also information, which is conflicting with the information received by hearing. As a result of this, the perception of an auditory event may be false. The most common examples of this cross-modal induction in spatial hearing are the previously mentioned McGurk effect (Chapter 2) and the Ventriloquism[5] effect.

As the binaural and spectral cues are dependent on the relative position of the sound source with respect to the head, the head movement affects these cue values. When we are unable to localize the sound accurately due to the front-back confusion or the torus of confusion, by observing the changes in the binaural cues relative to the head movement, we are able to improve the localization accuracy, since our auditory system is capable to perceive even the slightest changes in the binaural cue values. Consequently the localization of an auditory stimulus is usually the result of a co-operation between the auditory system, head movements and the vision.

3.3 HRTF and BRIR

The ultimate goal of sound reproduction is to reproduce the sound to the ears of the listener in a way that the listening experience matches the one in the recording situation. The achievement of this goal would require numerous microphones to be placed in the recording environment and the same amount of loudspeakers placed in the listening environment to match the placement of the microphones. As this is not possible (in most cases), there is a need for an alternative way to reach the goal.

Previous sections of this chapter have shown how the human auditory system is capable to localize sound events and what the limitations of this localization process are. Binaural technology claims that using knowledge of these properties it is possible to reproduce an authentic auditory experience, where the synthetic signals brought to the eardrums match those of the real-life listening experience (Riederer 2005). This is usually implemented using head-related transfer functions (HRTF) or binaural room impulse responses (BRIR). The latter parts of this section concentrate on presenting the measurement of these

[5] Spatially biased perception of the auditory stimulus to the same point as the simultaneous visual stimulus (Riederer 2005).

responses, as the description of their usage in applications would cover a thesis work by itself.

Blauert (1997) defines the free-field transfer function (FFTF), also known as HRTF, as the sound pressure measured at the listener's ear canal divided by the measured sound pressure at the position of the centre of the head, when the subject is absent. According to the definition, the head-related transfer functions or hear-related impulse responses (HRIR) are measured in anechoic conditions, with a listener placed at the centre and a loudspeaker positioned around the subject in different locations in the auditory space (Figure 3.2). Small microphones (earplugs), which record the reproduced sound, are inserted Into the subject's ear canals. By comparing these measurements to the one measured with same microphone at the position of the centre of the head in absence of the subject, the acoustic transfer functions from that location to the ipsilateral and contralateral ears of the subject are obtained. The head, pinna and torso effects on the sound are hence covered in these measurements, so the head-related transfer functions include all the spatial information needed for the auditory system, although some equalization must be applied to them before they can be used (Riederer 2005).

The binaural room impulse response measurements (BRIR) are an extension of the HRTF measurements. As the HRTF measurements are made in anechoic conditions, only the direct sounds enter the ears (and are therefore recorded in these measurements). The BRIR measurements on the other hand are made in different listening environments, where also the reflections of the sound can enter the ears of the listener. The measurement process is otherwise the same as in HRTF measurements, but due to the room effect, the binaural room impulse measurements are highly dependent on the listening environment and on the relative position of the subject in the environment.

In the BRIR and HRTF measurements, the microphone can be placed at any place inside the auditory canal of the subject (Hammershøi & Møller, 1996), as the definition does not state it specifically, as long as the positioning is taken into account in later stages. The most common placements are at the eardrum, at the entrance to the open ear canal and at the entrance to the blocked ear canal (Figure 3.7) and all of them have their own benefits and drawbacks in the applications (Riederer 2005). More information about the different microphone positions and their effects on the measurement results can be found in Hammersøi & Møller (1996) and in Riederer (2005).

The shape of the head, pinna and auditory canal however differ between human listeners, and despite the accurate positioning of the subject, the human listeners tend to move their head (unintentionally) during the measurements. Hence the transfer function measurements are hard to repeat exactly, and due to individual differences, the authentic listening experience in the applications would be possible only with the same subject. Therefore it is

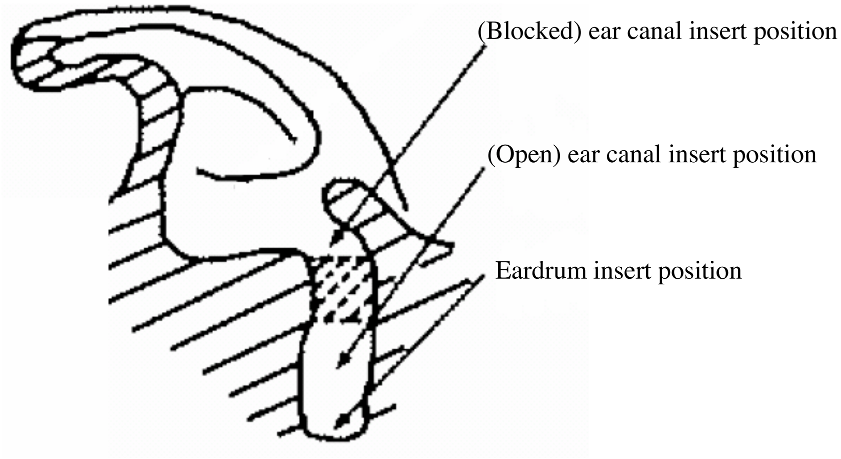

(Blocked) ear canal insert position
(Open) ear canal insert position
Eardrum insert position

Figure 3.7: Microphone positions in transfer function measurements (Edited from Riederer (2005)).

beneficial to use an artificial head (a dummy head) to measure the transfer functions, where the built-in microphones are always at the same place and the head does not move during the measurements. As the physical characteristics of an artificial head are an average from a large amount of people (ITU-T R. S. P.58 1996), the responses of an artificial head are statistically closer to a subject's own transfer functions than the ones measured from a random subject. Therefore the transfer functions from an artificial head have better usability in applications.

The use of an artificial head in HRTF and BRIR measurements is more repeatable than the ones made by human test subjects, but still it requires a lot of time for the preparations of the measurement setup and for running the test in different locations. One must also remember that the test equipment always adds some errors to the measurements. Therefore the use of a computational approach to obtain the impulse responses (or transfer functions) is an attractive alternative, as it can produce an arbitrary number of locations for the sound source without the time consuming effort required from the researcher. In the computational approach, the test subject, test environment and the sound source are all modelled into a computer simulation. Besides this, in binaural room impulse response simulations, one needs to simulate also the reflections, which is made by tracking the sound reflections from different surfaces. Needless to say, the computation becomes quite demanding, but still computational approaches are powerful tools in obtaining transfer function databases (Riederer 2005).

More information of the head-related transfer functions, measurement repeatability and idiosyncrasy of the transfer functions can be found in Riederer (2005). An example of computational approach to evaluate head-related transfer functions with a good description of the process is given in the work by Kirkeby et al. (2007).

3.4 Timbre

Sound timbre is a very important aspect in music since it allows diverse hearing experiences from similar tones. For instance in an orchestra, a flute and a bassoon are playing the same note with the same volume level, but still the listeners can easily hear the two instruments separately, since they differ in timbre (Goldstein 2002). Hence timbre has been defined as the attribute of auditory sensation, in terms of which listeners can judge two sounds with similar pitch and loudness as being dissimilar (Moore 1997). The broad definition of timbre makes it however a multi-dimensional concept, since sound can differ in multiple ways within the definition. Therefore it is not possible to measure timbre with just one value.

When inspecting the spectra of recorded sounds played with different instruments, one can find that the spectra differ in the way the energy of the sound is spread over the frequency range. Therefore timbre is considered to be dependent on the frequency content of the sound. A simple FFT (Fast Fourier Transform) spectrum is however not adequate for timbre inspections, since it does not take into account the effects of the human auditory system in its calculation (Karjalainen 1999). A more accurate evaluation of timbre can be achieved by inspecting the specific loudness of the sound. Before this concept can be described, a brief description of loudness must however be given first.

3.4.1 Loudness

The loudness of sound is a quantity closely relevant to the psychoacoustic quantity, loudness level, which describes how loud the sound is perceived by a listener. Consequently loudness is a fairly complex concept, but it still behaves logically as relation to the change of sound level (Moore 1997). Since the different parts of the auditory system amplify and attenuate different frequencies according to their transfer functions (Section 2.1), the loudness level and loudness of a sound depend on its frequencies.

Using hearing tests with different test subjects and stimuli, audio researchers have come up with equal loudness contours (ELC) (Figure 3.8), which describe the loudness level of a sound relative to its frequency and to the sound pressure level (SPL) (Equation (3.1) (Zwicker & Fastl 1999)). The pressure levels on the contours are perceived as equally loud. So the measurement of loudness with simple sounds (such as sinusoids) is quite straightforward. First, one measures the sound pressure level in dB and then converts that measure to loudness level in phones using the equal loudness contours. Finally the loudness in sones is obtained by using Equation (3.2) (Zwicker & Fastl 1999). Here N denotes the loudness in sones, p denotes the sound pressure and L_L denotes the loudness level.

$$L_p = 20 \times \log_{10}\left(\frac{p}{p_0}\right) \tag{3.1}$$

$$N' = 2^{\left(\frac{L_L - 40}{10}\right)} \tag{3.2}$$

where

$$p_0 = 20\mu\text{Pa}.$$

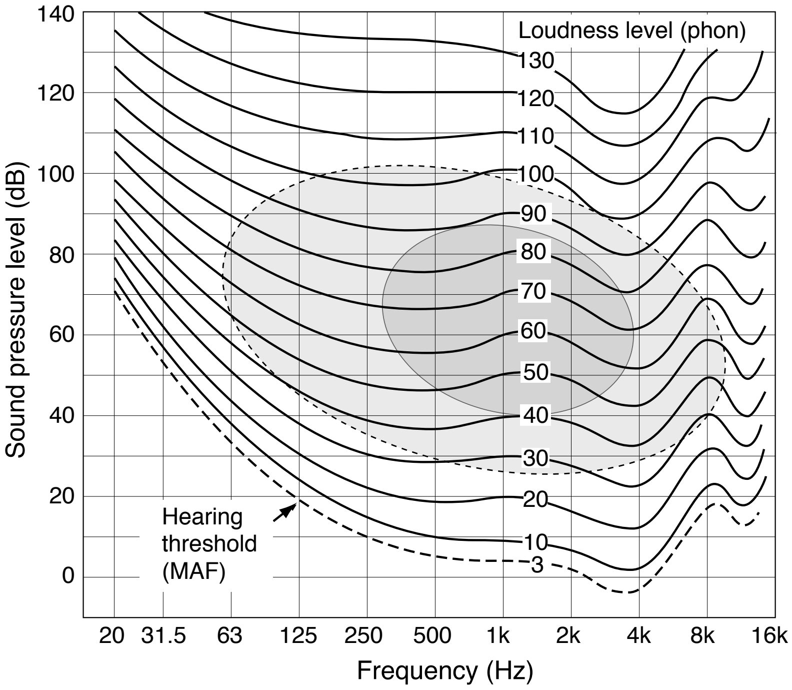


Figure 3.8: Hearing threshold and equal loudness contours as function of frequency and sound pressure level (Adapted from Karjalainen (1999)).

With broadband signals, the process is not however that simple, as the energy of basilar membrane motion at a certain place (frequency) spreads to adjacent places (frequencies) and the overall loudness is calculated from the spread values. Research on hearing has shown that the auditory system analyses broadband sounds with critical band resolution and that the overall loudness is calculated from these values (Karjalainen 1999). Therefore in perceptual auditory models the specific loudness of the sound is calculated.

3.4.2 Specific loudness

Specific loudness models the above-explained process of the auditory system by summing the energies on each critical band based on the frequency and the critical band separation (Table 4.1). Consequently, specific loudness describes the loudness per critical band. As described above, the excitation energy at given frequency spreads also to neighbouring critical bands and that spreading has to be taken into account also in specific loudness calculations. Therefore in the auditory models, the specific loudness values are calculated from the excitation patterns of the basilar membrane models (Section 2.3.2) using Equation (3.3) (Zwicker & Fastl 1999). Here z denotes the number of the critical band and the scalar c must be selected so that a 40 dB 1 kHz sinusoidal tone has the loudness of 1 sone.

$$N'(z) = c \times E\left(z\right)^{0.23} \tag{3.3}$$

As there exist two main approaches to model the selectivity of the basilar membrane (Section 2.3.2), the specific loudness values can also be calculated using the equivalent rectangular bandwidth (ERB) separation instead of Bark bands to discriminate loudnesses on different frequencies. More information about this approach can be found in Moore et al. (1998).

Regardless of the selected approach, the overall loudness value is then obtained, by simply summing the specific loudness values. In binaural auditory models the overall loudness and the specific loudness are calculated by summing the specific loudness values on the two ears at each critical band. More information about the Bark scale based approach can be found in Zwicker & Fastl (1999).

3.4.3 Timbral aspects

From the specific loudness values one can then evaluate the timbral quality aspects of the sound. The multidimensionality of timbre comes from the fact that quite many aspects relate to it. Therefore in this section only the ones, which are the most relevant to this work, are explained. In this case the interest of research is in studying the differences between different loudspeakers in music reproduction. The research conducted by Lorho (2007) shows that the most common timbral aspects that listeners can use to discriminate different loudspeakers in music reproduction are low-frequency emphasis (bass), high-frequency emphasis (treble) and sound balance differences.

Bass and treble are somewhat complementary concepts, as bass refers to the amount of low frequency content in the sound and treble on the other hand to the amount of high frequency content. So the higher the specific loudness value is at the first critical bands, the more bass the sound has. Vice versa, the higher the specific loudness values are at the highest critical bands, the more treble there is. As some loudspeakers are good in producing both bass and treble (and some in neither), a third aspect is needed to check the balance between these two. Besides evaluating the balance between bass and treble, this aspect also checks whether some frequency range is emphasized or lacking in the reproduced sound. Hence the specific loudness of a sound with good balance follows closely the specific loudness values of the same sound reproduced by a reference device.

Other aspects relative to these three and to timbre overall include (among others) sharpness, roughness, sensory pleasantness, tonality, dissonance and consonance. A good overview of these aspects can be found in Karjalainen (1999).

3.5 Summary

As a person grows older, the ossicles wear down and some of the hair cells and connections in the auditory nerves are destroyed. Therefore the hearing thresholds increase over the years. The ability to localize sounds however improves as the person learns to use the different cues more accurately over time. This results from the increased knowledge of different surroundings, stimuli and increased capability to discriminate binaural cues.

In this chapter the ability to localize sounds was explained, by first describing the binaural and spectral cues, which provide the main information about the sound direction. The limitations of these cues were also presented, as well as the different methods for sound distance evaluation. The important role of other senses in localization was also mentioned to show that the localization is not only based on auditory information.

The next section of the chapter gave a brief introduction of head-related transfer function and binaural room impulse response measurements to show how the source position dependent head, torso and pinna effects can be measured or simulated. The last section of the chapter introduced timbre and explained how its aspects can be evaluated from the specific loudness values. The timbral quality aspects relevant to reproduced music were also presented.

Chapter 4

Binaural Auditory Model

4.1 Introduction

In the previous chapters we have learned that in human auditory perception, the high level processes (such as perception of pitch, source location, etc.) are handled in different areas in the brain by comparing the neural information from the two ears. Despite the fact that the exact role of different areas is still to be resolved, this high processing level makes the modelling of the hearing perception complex, since one needs to take into account all the effects the torso, the head, and different parts of the auditory system have on the sound, in order to get an accurate evaluation of how a human subject might perceive a certain sound stimulus. Still one can, if needed, simplify the model a bit by focusing only on perception of one certain aspect. In this case, one can model only the parts of the auditory system, which are relevant to the current task and still get a fairly good estimate of the aspect.

For instance from the localization point of view, the only parts of the auditory system, whose transfer function is dependent on the direction of the sound source, are the pinna and to some extent the auditory canal, as described earlier in Chapter 2. Hence the effect of the cochlea and the middle ear can be regarded as insignificant to the task, and a simpler model can be used. As an example of this kind of approach Avendano & Jot (2004) used cross-correlation of the frequency content information of the left and right ear inputs to estimate the binaural cues. They were thus able to develop a real-time application for creating multichannel upmix for stereo, based on the binaural cue estimates.

When the auditory model is however used to evaluate two or more (different) aspects of the sound, all the parts of the ear and their effects have to be modelled and a simplified approach cannot be used. Thus the model becomes quite complex, but at the same time the model can now evaluate these aspects more accurately, and so with the help of the model, one can get a better idea about the overall perception of sound.

This chapter presents a new binaural auditory model where the effects of different parts of the auditory system have been modelled. Consequently this model and its MATLAB (The Mathworks 2007) implementation can be used to evaluate both spatial and timbral aspects of the reproduced sound. This work is inspired by the Binaural Cue Selection Model created by Faller & Merimaa (2004). Section 4.2 of the chapter presents the proposed binaural auditory model, including the estimation of the specific loudness and binaural cue values. An approach of Mapping sound source directions based on the binaural cue estimates and a lookup table is presented in Section 4.3. Also a short summary of the chapter is given in Section 4.4

4.2 Model description

4.2.1 Peripheral hearing model

The main goal of this auditory model is to model how a human listener might perceive the sound in terms of spatial and timbral aspects. This is why the inspected sound is reproduced via a loudspeaker setup in an anechoic chamber and the sound is recorded with a Head and Torso Simulator (HATS), which has a microphone inserted in the eardrum reference point (DRP) (ITU-T, (1996) in the left and right (artificial) ear of the HATS (Figure 4.2). The recording is made with a 48 kHz sampling rate. As each of these recordings can be quantified as an acoustic transfer function of the sound from free-field to the eardrum of a listener (B&K 2006), we have a binaural recording of the sound, where the combined influence of the torso, head, pinna and the auditory canal has thus been modelled.

This binaural recording is then used as an input to the model, whose peripheral structure is illustrated in Figure 4.1. Due to complexity of the model, running it on a normal computer requires quite a lot of calculation power (and time). In order to make the calculations less demanding, without being forced to settle for a less complete analysis of the aspects, the necessary calculations are made by inspecting only a small part of the recordings at a time. Therefore the input signals are divided into rectangular, 200 ms long time windows in the windowing phase (Figure 4.1). As the sample rate is 48 kHz, there are 9600 samples in one time window. To ensure that the results from different time windows have a more continuous form over time, the adjacent time windows are set to be 50% overlapping. Hence there are 4800 samples in the beginning of a time window, which are the same as the ones in the end of the previous time window.

The next logical phase is to model the transfer function of the middle ear, as the binaural recordings were made at the eardrum reference point. In this model this is implemented by filtering the binaural input signals with an FIR (finite impulse response) filter whose magnitude response resembles the middle ear transfer function presented by Moore et al.

(1998). The filtering is made with the help of a function from the freely available Matlab toolbox by Irino & Patterson (1997) and the magnitude and phase responses of the middle ear compensation filter are illustrated in Figure 4.3.

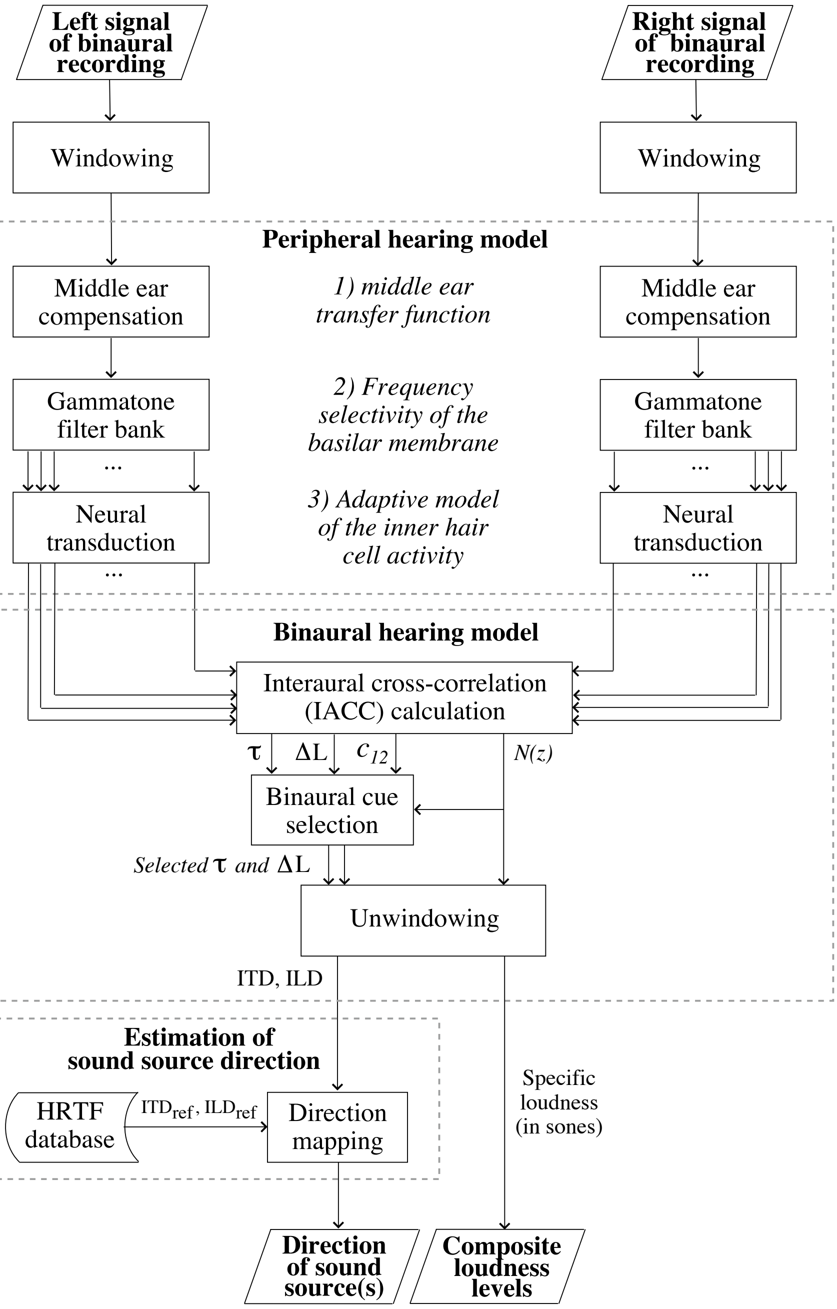

Left signal of binaural recording
Right signal of binaural recording
Windowing
Windowing
Peripheral hearing model
Middle ear compensation
1) middle ear transfer function
Middle ear compensation
Gammatone filter bank
2) Frequency selectivity of the basilar membrane
Gammatone filter bank
...
Neural transduction
3) Adaptive model of the inner hair cell activity
Neural transduction
...
Binaural hearing model
Interaural cross-correlation (IACC) calculation
τ
ΔL
c_{12}
N(z)
Binaural cue selection
Selected τ and ΔL
Unwindowing
ITD, ILD
Estimation of sound source direction
HRTF database
ITD_{ref}, ILD_{ref}
Direction mapping
Specific loudness (in sones)
Direction of sound source(s)
Composite loudness levels

Figure 4.1: Peripheral structure of the model.

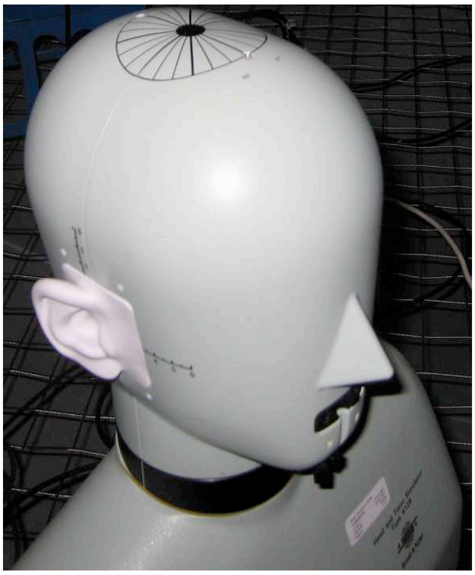


Figure 4.2: Head and torso simulator (B&K, 2006) that was used to record stimuli.

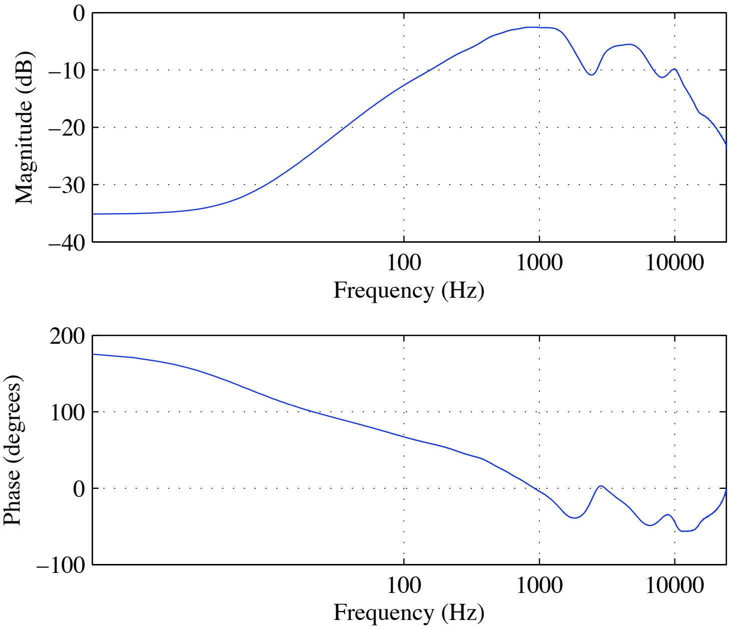


Figure 4.3: Magnitude (above) and phase (bottom) responses of middle ear compensation filter.

After the middle ear compensation process is finished, the filtered (binaural) inputs are passed to a gammatone filterbank (GTFB), which has been considered accurate enough to model the frequency analysis done in the basilar membrane (Merimaa (2006) & Section 2.3.2). In this model a 24-band filter bank is used to cover the whole used frequency range (from 0 to 15.5 kHz), where the characteristic frequencies[1] are the same as the centre frequencies of the 24 critical bands on the Bark scale. This one filter per critical band approach is used instead of the one filter on each 42 ERB type approach because of two main reasons.

Firstly, using 42 or more instead of 24 filters makes the calculations more demanding (and time consuming), since the computer has to handle a larger amount of data at each of the following phases of the model. The comparison of the results showed also that the increase of filter channels did not produce major improvements in the results. Hence the increasing of the amount of filters is not reasonable, regarding the purpose of this work. Secondly, with this approach the specific loudness (Section 3.4.2) in each critical band, as presented in Zwicker & Fastl (1999), can be estimated directly from the outputs of the neural transduction model.

The filter coefficient calculation and the actual filtering are implemented by using the freely available Matlab toolbox by Slaney (1998), where the used centre frequencies are listed in Table 4.1 and the magnitude responses of some of the filters are shown in Figure 2.11. Note that this toolbox is originally designed for ERB bands.

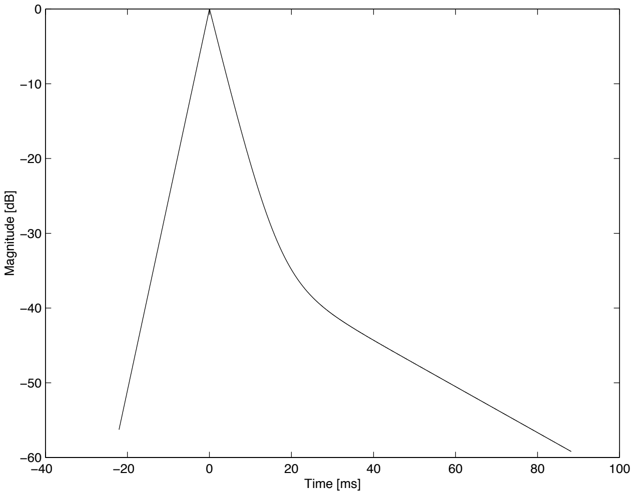


Figure 4.4: Magnitude response of temporal window introduced by Plack & Oxenham (1998) (Adapted from Härmä (1998)).

[1] Center frequencies of the GTFB filters, where the amplitude response of the filter is the highest.

Critical band No.	**Lower limit frequency [Hz]**	**Centre frequency [Hz]**	**Upper limit frequency [Hz]**
1.	0	50	100
2.	100	150	200
3.	200	250	300
4.	300	350	400
5.	400	450	510
6.	510	570	630
7.	630	700	770
8.	770	840	920
9.	920	1000	1080
10.	1080	1170	1270
11.	1270	1370	1480
12.	1480	1600	1720
13.	1720	1850	2000
14.	2000	2150	2320
15.	2320	2500	2700
16.	2700	2900	3150
17.	3150	3400	3700
18.	3700	4000	4400
19.	4400	4800	5300
20.	5300	5800	6400
21.	6400	7000	7700
22.	7700	8500	9500
23.	9500	10500	12000
24.	12000	13500	15500

Table 4.1: Critical band separation on Bark scale (Zwicker & Fastl 1999).

After this gammatone filter bank filtering is complete, the time window inputs have become 24× 9600 sample matrices to illustrate the basilar membrane movement at each characteristic frequency at different time instants. Before the binaural cue and the specific loudness values can be estimated, transformation of the basilar membrane movement into neural impulses has to be modelled. This is implemented by a neural transduction model, which models the activity of the inner hair cells (Section 2.1.3). In this work, a model based on the functional approach (Section 2.3.2) with a variable temporal window model is used to model the forward masking and the adaptation in the firing patterns of the inner hair cells (Section 2.3.2). From the different functional models listed earlier (Section 2.3.2) the model presented by Plack & Oxenham (1998) is selected, because its implementation is most compatible with the rest of the model, especially with the interaural cross-correlation (IACC) that is described in Section 4.2.2.

In the generation of the model the approach (defined by Bernstein et al. (1999)), which was used in the work by Faller & Merimaa (2004), was also tested. The comparison between the two approaches revealed that both of them work well with the binaural cue value estimation, but the model by Plack & Oxenham (1998) provides more accurate estimation of the specific loudness values and is therefore used in this model. The next paragraphs explain the process of the neural transduction model.

In the first step of the model of the inner hair cell activity, the GTFB filtered signal is half-wave rectified, so that the negative values are removed (Equation (4.1)). Here x denotes the signal from either left or right ear, i denotes the sample number and z denotes the critical band number. In the next two phases the smoothed envelope of these outputs is calculated to illustrate the firing pattern of the inner hair cells to the basilar membrane movement at that CF. In the first phase the basilar membrane movement is compressed in two separate regions, as recent results show that the output-input relationship of basilar membrane at a given CF can be divided to two regions (Plack & Oxenham 1998), depending on the signal level (dB in SPL) at that time. For this reason, the rectified outputs must be first transformed into dB by Equation (4.2), then compressed by Equation (4.3) and finally transformed back to the same form as before according to Equation (4.4).

$$\hat{x}\left[z,i\right] = \max\left(x\left[z,i\right],0\right) \tag{4.1}$$

$$L_p\left[z,i\right] = 20 \times \log_{10}\left(\frac{\hat{x}\left[z,i\right]}{p_0}\right) \tag{4.2}$$

$$\widehat{L_p}\left[z,i\right] = \begin{cases} 0.78 \times L_p\left[z,i\right], \ L_p\left[z,i\right] \geq \ 35 \text{ dB}, \\ 0.16 \times L_p\left[z,i\right], \ L_p\left[z,i\right] \leq \ 35 \text{ dB}. \end{cases} \tag{4.3}$$

$$x\left[z,i\right] = p_0 \times 10^{\frac{\widehat{L_p}\left[z,i\right]}{20}} \tag{4.4}$$

where

$$p_0 = 20 \ \mu\text{Pa}.$$

In the second phase each excitation signal is convoluted with a temporal window function (Figure 4.4) in order to get a smoothed envelope pattern of the neural output signal with the effects of forward masking taken into account. The output matrices of this process are from here on denoted as x_1 (the left output) and x_2 (the right output). The neural transduction is implemented using functions from the freely available Matlab toolbox by Härmä & Palomäki (1999).

4.2.2 Binaural hearing model

This model uses the interaural coherence to estimate the binaural cue values from the outputs of the neural transduction model in the peripheral hearing model (Figure 4.1). The physical basis for this approach was founded by Jeffress (1948) in his psychophysical studies of human hearing perception (Merimaa 2006) and the idea to estimate binaural cue values for each critical band separately was presented in the work by Merimaa (2006). The implementation of this is based on the approach presented in the works by Faller & Merimaa (2004) and by Merimaa (2006) and Equations (4.5), (4.6), (4.8), (4.9) and (4.10) are modified versions of the ones presented in those publications. The first part in this process is to calculate the normalized interaural cross-correlation (IACC), denoted as γ, between the two signals. Here *m* denotes the time lag, which is used to check the similarity of the two inputs at a given sample and critical band and i denotes the sample number.

$$\gamma\left[z,i,m\right] = \frac{a_{12}\left[z,i,m\right]}{\sqrt{a_{11}\left[z,i,m\right]a_{22}\left[z,i,m\right]}} \tag{4.5}$$

where

$$\begin{aligned}
a_{12}\left[z,i,m\right] &= \alpha x_1\left[z,i-\max\left(0,-m\right)\right]x_2\left[z,i-\max\left(0,m\right)\right] \\
&\quad +\left(1-\alpha\right)a_{12}\left[z,i-1,m\right], \\
a_{11}\left[z,i,m\right] &= \alpha x_1\left[z,i-\max\left(0,-m\right)\right]x_1\left[z,i-\max\left(0,m\right)\right] \\
&\quad +\left(1-\alpha\right)a_{11}\left[z,i-1,m\right], \\
a_{22}\left[z,i,m\right] &= \alpha x_2\left[z,i-\max\left(0,-m\right)\right]x_2\left[z,i-\max\left(0,m\right)\right] \\
&\quad +\left(1-\alpha\right)a_{22}\left[z,i-1,m\right].
\end{aligned}$$

The limits within which the time lag can change are set according to the maximum allowed interaural time difference (ITD) between the left and right ear signals. In this model this limit is set to be ±1 ms, which was also used in Faller & Merimaa (2004). This time lag is adequate for the direct sound to arrive to the both ears within the time lag, as the sound travels approximately 33 cm in 1 ms in air (at 20°C temperature) and the diameter of the head is approximately 17 cm.

The second unknown factor needed for IACC calculations is the forgetting factor α, which defines the used time resolution *T*. Here f_s denotes the used sample rate, which is 48000 Hz.

$$T = \frac{1}{\alpha f_s} \tag{4.6}$$

Selection of this time resolution is difficult since researchers have noticed that different time integration might be used in different cases. This model uses the same 10 ms time resolution, which was used in the model presented by Faller & Merimaa (2004), and which according to Merimaa (2006) is close to the smallest ones that are used in the studies of temporal resolution of binaural hearing. More details about these studies and about selecting the time resolution value can be found in Merimaa (2006).

The next phase of the process facilitates finding the most probable values for the interaural coherence (IC). In this phase the IACC values, for a given time lag, are multiplied with the values (at the same sample number and the same time lag) in the neighbouring frequency bands. The reason for this process is that the peaks at prominent frequency bands at given time lag are considered to be more relevant in localization compared to single peaks at one of the frequency bands (Pulkki & Karjalainen 2001).

$$\hat{\gamma}\left[z,i,m\right] = \gamma\left[z-1,i,m\right]\gamma\left[z,i,m\right]\gamma\left[z+1,i,m\right] \tag{4.7}$$

Following this phase, the binaural cue values can be estimated. The ITD, denoted as τ, can be estimated by looking for the time lag value *m*, with which the interaural cross-correlation (IACC) receives its highest value.

$$\tau\left[z,i\right] = \arg\max_{m}\left(\hat{\gamma}\left[z,i,m\right]\right) \tag{4.8}$$

As the estimate is obtained in samples, it must be transformed to milliseconds, by dividing it with the sample frequency (in kHz). It indicates how much sooner the sound arrived to the left ear compared to the right ear. At the same time also the needed estimates for the interaural coherence, denoted as c_{12}, are obtained from the maximum values of the IACC.

$$c_{12}\left[z,i\right] = \max_{m}\left(\hat{\gamma}\left[z,i,m\right]\right) \tag{4.9}$$

Finally, the interaural level differences (ILD), denoted as ΔL, can be estimated from the energies of the two signals revealing how much higher level (in dB) the left signal is with respect to the one on the right. The calculation is implemented as suggested by Merimaa (2006).

$$\Delta L\left[z,i\right] = 10 \times \log_{10}\left(\frac{L_1\left[z,\tau\left(i\right)\right]}{L_2\left[z,\tau\left(i\right)\right]}\right) \tag{4.10}$$

where

$$\begin{aligned} L_1\left[z,\tau\left(i\right)\right] &= \alpha\left(x_1\left[z,i-\max\left(\tau\left(i\right),0\right)\right]\right)^2 \\ &\quad +\left(1-\alpha\right)L_1\left[z,\tau\left(i-1\right)\right], \\ L_2\left[z,\tau\left(i\right)\right] &= \alpha\left(x_2\left[z,i-\max\left(-\tau\left(i\right),0\right)\right]\right)^2 \\ &\quad +\left(1-\alpha\right)L_2\left[z,\tau\left(i-1\right)\right]. \end{aligned}$$

Specific loudness estimation

As described earlier in this work (Section 3.4.2), the specific loudness is generally calculated at each critical band using Equation (3.3), where E denotes the energy at the given critical band and the scalar c must be selected so that a 1000 Hz sinusoidal sound at 40 dB level has loudness of 1 sone (Zwicker & Fastl 1999). Then the specific loudness on other critical bands is calculated by scaling them according to the equal loudness contours (ELC) and using the loudness of 1 kHz sound as a reference.

In this case the scalar is ignored and set to one, because in the earlier phases of the model, the binaural inputs have been attenuated and amplified according to the external ear, middle ear and the cochlea transfer functions. Therefore the relative loudness in different frequency bands are already scaled to an approximately equal level. Hence the scaling is required only to obtain accurate evaluation of the absolute loudness level and is therefore not implemented at this point, as in the evaluation of differences between different signals the relative loudness information is adequate. In the later stages this scaling should however be included in the model if the model is to be used to evaluate also the absolute loudness level of the sound. Hence the (binaural) specific loudness at each critical band is currently estimated in the model according to the idea presented by Pulkki et al. (1999), where the fourth root estimates the exponent used in Equation (3.3) and I denotes the number of samples in a time window (I = 9600).

$$N'\left(z\right)=\sqrt[4]{\frac{1}{I}\sum_i L_1\left[z,\tau\left(i\right)\right]}+\sqrt[4]{\frac{1}{I}\sum_i L_2\left[z,\tau\left(i\right)\right]} \tag{4.11}$$

This approach to calculate the specific loudness for the given critical band by summing the loudnesses of the left and right ear signals in sones is however only a rough approximation of the actual binaural (loudness) summation. Recent research on the binaural loudness and binaural summation has shown for instance that both the frequency of the sound and the interaural level difference (ILD) between the two ears affect on how the loudness is increased in binaural hearing compared to the monaural hearing. More information about these measurements and tests can be found in the publication by Sivonen & Ellermeier

(2006). Considering the goal of this work, this approximation is however accurate enough, as the model can be used to compare the specific loudness differences between different stimuli despite this approximation.

Binaural cue selection

In this model the binaural cues are not always considered to be accurate or relevant enough and two criteria are used to check whether the ILD and ITD values are accepted or not. The selection is developed from the idea presented by Faller & Merimaa (2004). In this model the binaural cue values are set as missing or NaN (not a number) values in the Matlab implementation when both criteria are not met. The missing values are used here because setting the ILD and ITD values to zero when the used criteria are not met would cause problems in the estimation of the sound source direction (Section 4.3), as the model would detect a sound source in the front (or at the back) of the listener even if the criteria are not met, which is not desirable. By setting the binaural cues to missing values ensures that the model detects a sound source located in the front (or at the back) only when the criteria are met.

The first criterion is used for the interaural coherence (IC) value to test the reliability of the cue estimates. Due to the normalization in the calculation of the value (Equation (4.5)), the interaural coherence values lie by definition between 0 and 1. The less noisy the recording environment and the better the quality of the recording are, the closer the IC value is to 1 (Merimaa 2006). Therefore the selection of the criterion value depends on the situation and the purpose of the study. In this model a value of 0.98 is used as the criterion value, because the recordings are made in anechoic conditions (IC values are close to 1). This way the model will discard uncertain estimates for binaural cues and still provide enough values for post processing. Note that high IC values were also used by Faller & Merimaa (2004) and by Merimaa (2006). More details about selecting this criterion value can be found in Merimaa (2006).

The second used criterion inspects the specific loudness values. This is needed, since the interaural cross-correlation only checks how similar the two input signals are at current point and provides therefore high IC values also, when there is little or no signal power on both inputs. To remove the false estimations caused by this situation, a criterion value is used to check that the specific loudness (Equation (4.11) at a given critical band meets the loudness criterion and the model therefore localizes sound events only when a signal is actually present. The ’correct’ loudness criterion value depends on the recording environment and on the type of sound source. In the model testing (Chapter 5) and in the case study with mobile loudspeakers (Chapter 6) a loudness of 0.5 sones was used as the loudness criterion. The same loudness criterion value can be used for all of the critical bands, since the effects of the external and the middle ear have been modelled on the inputs.

Unwindowing step

After this interaural coherence phase of the model is completed, the binaural cue values and specific loudness values for the given time window are stored (in the memory), and the processing of the next time window begins (Figure 4.1). When the last window of the inspected sound has passed this chain of processes, these values are fetched from the memory for the post-processing in the unwindowing phase of the model. Here, as the name of the phase hints, the windowing, made in the beginning of the model, is reversed and a continuous-time form is created for both ILD and ITD cues to present their changes over time in a better way.

Since the windows created in the windowing phase were half overlapping, this process is quite straightforward. The mean value of the binaural cue values in adjacent time windows, which are referring to the same sample, is selected to present the binaural cue value at that time instant. The process of these calculations in the overlapping parts of the signal is presented in Equation (4.12), where *I* denotes the length of the time window, *j* denotes the number of the time window, *k* is the time instant (in samples) in the original continuous input signal and *i* is the sample number in the time window.

$$\begin{aligned} itd\left[z,k\right] &= \frac{1}{2}\left(\tau_{j-1}\left[z,\frac{I}{2}+i\right]+\tau_j\left[z,i\right]\right) \\ ild\left[z,k\right] &= \frac{1}{2}\left(\Delta L_{j-1}\left[z,\frac{I}{2}+i\right]+\Delta L_j\left[z,i\right]\right) \end{aligned} \tag{4.12}$$

where

$$\begin{aligned} 0 \le i \le \frac{I}{2}, \\ k = j\frac{I}{2}+i. \end{aligned}$$

The values that have no overlap (values at the beginning of the first time window and at the end of the last time window) are used as such. From the specific loudness values at different time windows an overall composite loudness level (CLL) spectrum is obtained with percentile analysis. This is made by taking the highest 5% of the short-term loudness values for the given critical band to present the overall loudness on that critical band. More information about this percentile analysis can be found in Zwicker & Fastl (1999).

4.3 Estimation of spatial direction

One of the goals of the model is to evaluate the reproduced sound in terms of spatial aspects, but finding the differences between different sounds just by looking at the obtained ILD and ITD estimates is a difficult task. The obtained ILD and ITD estimates are only providing cues of the spatial location of the sound source and are often providing conflicting information. Therefore the direction of the sound source needs to be estimated in order to evaluate spatial aspects of the reproduced sound. In this model the self-calibration idea presented by e.g. Macpherson (1991) is used. According to this idea, a lookup table is generated for the binaural cues to tell which spatial location provides certain binaural cue values and the spatial direction analysis is done based on this information. The used approach for obtaining these reference values is explained next.

4.3.1 Lookup table for binaural cues

In the generation of the binaural cue lookup tables the location of a monophonic sound source is simulated in different directions around the HATS and this simulated signal is fed as an input to the model. The model then calculates the binaural cue values for the given simulation and these binaural cue values are then used as reference values for the spatial location that corresponds to the given simulation. The monophonic sound source is a simple stepped sine sound (Appendix A), which consists of a series of 200 ms sinusoidal signals, whose frequencies are the same as the characteristic frequencies (CF) (Table 4.1) of the gammatone filterbank (GTFB). There are also a 50 ms pause between the sinusoids, at the beginning and at the end of the signal. The reason for using this kind of simple signal is to ensure as high interaural coherence value as possible, as there is a simple signal at only one critical band at a time.

The sound source direction simulation is made by applying the HRTF database (Section 3.3) developed by Kirkeby et al. (2007). In their studies the authors simulated the same HATS (with and without the torso) in anechoic conditions to obtain the head-related transfer functions (HRTF) for different angle locations around the HATS in far-field conditions. In their results, they have estimated the HRTFs for both azimuth angle (in 2 degree accuracy) and also elevation (in 5 degree accuracy) (Figure 3.2), which represents a very high spatial resolution. As a result of the completeness of the data and the use of the same HATS in the same listening environment, the binaural cue lookup tables can be generated with the help of this dataset.

There is however a need for some additional processing of the transfer function values, since the data in the work by Kirkeby et al. (2007), is simulated with a microphone placed at the entrance to the blocked ear canal (Figure 3.7) (Section 3.3) and the model presented in this work begins at the eardrum reference point (DRP). Therefore a correction filter was needed to model the effect of the auditory canals transfer function on the sound. This transfer function was generated from a set of HRTF measurements made with a dummy-head (Lorho 1998) at the two positions (blocked ear canal and DRP) and convoluted with all the head related impulse responses (HRIR). Although this transfer function has some effect on the measurements, these effects can be considered as insignificant to reference value estimation because of two reasons.

Firstly, the same transfer function is implemented for both left and right ear inputs, so the relative differences between the two inputs stay unchanged. Secondly, the spatial information is already included in the original data and the transfer function has effects only on the frequency content of the signals. Since this data is used only to evaluate spatial information, this change in timbral aspects is insignificant.

In this work the localization is limited to inspect locations only in the horizontal plane (zero elevation) and the azimuth angles are inspected at 10 degree accuracy throughout the whole azimuth plane, which is the same accuracy that human listeners can point the direction of the sound source (Section 3.2.1, Makous & Middlebrooks (1990)). These limitations are used to reduce the needed calculation power. The ILD and ITD reference values are the mean values of the estimated binaural cue values over different time windows respectively (Section 4.2.2), where the non-existing values are ignored (Section 4.2.2).

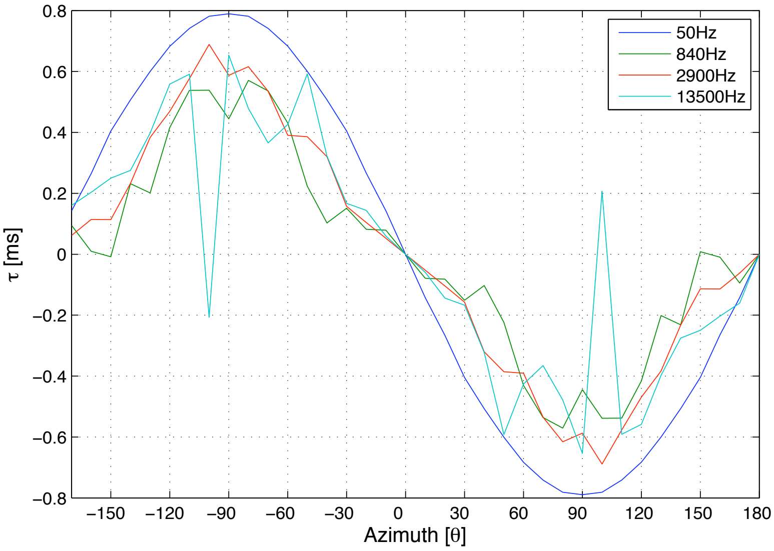


Figure 4.5: ITD as function of azimuth in different critical bands. Critical bands are presented by their characteristic frequency.

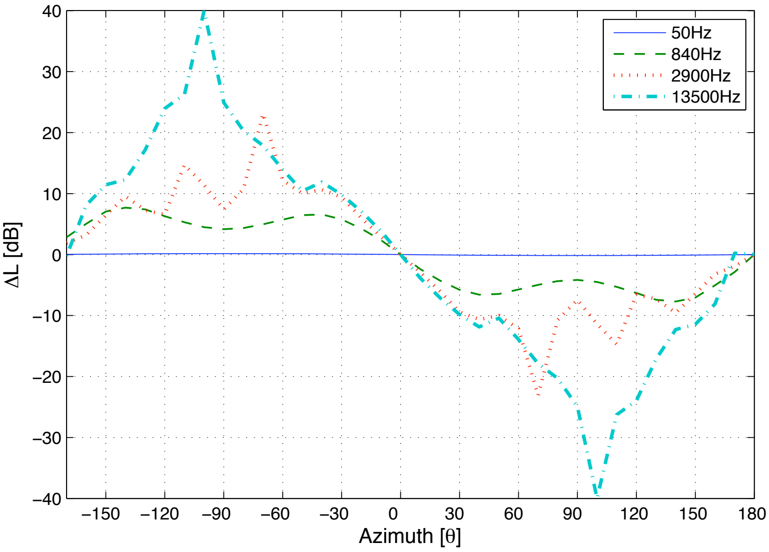


Figure 4.6: ILD as function of azimuth in different critical bands. Critical bands are presented by their characteristic frequency.

Figures 4.5 and 4.6 present the behaviour of the ITD and ILD reference values in different critical bands, whose characteristic frequencies are 50, 840, 2900 and 13500 Hz, respectively. The azimuth angles are positive on the right side of the HATS and negative on the left side of the HATS (Figure 3.2). Figures 4.5 and 4.6 illustrate that ITD provides bigger and more consistent differences in the low frequency areas and ILD on the other hand on the higher frequency areas. These results are consistent with the binaural cue, front-back confusion and the cone of confusion theories presented earlier (Section 3.2)

4.3.2 Mapping of sound source direction

Now that the necessary elements are ready the model can be used to localize sound events spatially from the reproduced sound. In this phase of the model, the obtained estimates of binaural cues (Section 4.2.2) are compared (one sample at one frequency band at a time) to the reference values in the binaural cue lookup tables (Section 4.3.1). As the recent results and theories about the binaural cues (Section 3.1.1) present, in the low frequencies the interaural time difference (ITD) is mainly responsible for the localization and in the high frequencies, the localization is based on the interaural level difference (ILD). At the same time these theories tell us that there is a frequency zone between these areas, where the localization is a result of both of these cues, and that in this area the weights of ITD and ILD cues in localization are still unknown.

This work presents a new attempt in localizing sounds spatially in different frequency bands. Here the localization is separated into two different categories based on the characteristic frequency (CF) of the frequency band in the GTFB, and each of these categories uses a different approach to find the most likely azimuth angle for the sound source at a given time instant. In the first category (where the CF of the frequency band is below 1100 Hz) the localization is based entirely on the ITD information and in the second category (where the CF is higher than 1100 Hz) the localization is based both on ITD and ILD. The ITD based mapping is made by finding the azimuth angle (denoted as θ) for which the squared difference between the estimated ITD (denoted as *itd*) and the ITD reference (denoted as ITD_{ref}) value in the lookup table is the smallest. In the higher frequencies the squared proportional errors (denoted as ΔITD and ΔILD) are calculated to model the uncertainty of the mapping to a given azimuth angle based on the binaural cue. As a result of these uncertainties, the azimuth angle having the smallest scalar product of these two uncertainty measures is selected as the most probable source direction.

$$\theta\,[z,k] = \arg\min_{\theta} \begin{cases} (ITD_{\text{ref}}\,[\theta,z] - itd\,[z,k])^2, \; cf\,[z] \leq \text{1.1 kHz} \\ \Delta ITD[z,k,\theta] \times \Delta ILD[z,k,\theta], \; cf\,[z] > \text{1.1 kHz} \end{cases} \tag{4.13}$$

where

$$\begin{aligned} \Delta ITD\,[z,k,\theta] &= \left(\frac{ITD_{\text{ref}}\,[\theta,z] - itd\,[z,k]}{itd\,[z,k]}\right)^2, \\ \Delta ILD\,[z,k,\theta] &= \left(\frac{ILD_{\text{ref}}\,[\theta,z] - ild\,[z,k]}{ild\,[z,k]}\right)^2, \end{aligned}$$

ILD_{ref} denotes the reference value for ILD and *ild* denotes the obtained ILD estimate. This frequency separation in the localization algorithm was selected based on evaluating the behaviour of the binaural cue values in the lookup table, where the ITD works consistently in the frequency range below 1.1 kHz and above this limit the binaural cues provide also ambiguous information. Although the ILD works consistently again in the higher frequencies, this frequency separation provided more precise results in practice and was therefore used.

In this localization algorithm the model is not forced to localize the sources into specified area in the azimuth plane. Since the binaural cue values for given azimuth (at either right or left side of the head) in front of the head are similar (almost identical) to the ones at the back of the head, the model is likely to estimate the source locating either in the frontal plane or at the back (Figure 3.5(a)) due to front-back confusion (Section 3.1.3). Hence the model is likely to suffer from the same difficulties in localization as a human listener in anechoic conditions. Depending on the use of the model, this feature can however be easily removed by using only half of the azimuth plane in the reference table.

4.4 Summary

In this Chapter a new binaural auditory model (BAM) of the human auditory system was described. Partly similar approaches have been presented previously in literature by Macpherson (1991), Karjalainen (1996), Pulkki & Karjalainen (2001), Faller & Merimaa (2004) and by Merimaa (2006), and the model presented in this work is built by combining together some aspects from these models. The composite loudness level (CLL) evaluation is formed similarly as in Pulkki & Karjalainen (2001) although this model uses the 24-band Bark-scale approach instead of the 42-band equivalent rectangular band (ERB) based approach in the gammatone filter bank (GTFB) to model the frequency selectivity of the basilar membrane.

The nonlinear approach for modeling the temporal aspects of the basilar membrane functionality was used first in Karjalainen (1996). In this model however the temporal window presented by Plack & Oxenham (1998) is used instead of the one presented in Karjalainen (1996), because the interaural cross-correlation (IACC) calculation worked better with it. The binaural cue estimate calculation is based on the approach presented by Faller & Merimaa (2004) and by Merimaa (2006), although in those publications the author(s) inspected the binaural cues only within one frequency band of the GTFB whereas in this model the values are inspected in all of the 24 frequency bands.

The middle ear compensation filter has been added as a new feature to allow more precise evaluation of the loudness level. This modeling of the middle ear transfer function also sets the localization capability of the model to the human performance level as well as the adding of the internal noise to the signal in Merimaa (2006) does. Therefore no internal noise is added to the signals in this model. The mapping of binaural cue estimates to corresponding azimuth locations by using a lookup table is also added in this model.

Chapter 5

Model Testing

In order to verify the functionality of the developed binaural auditory model (BAM), it was tested with various sound stimuli in anechoic conditions. This chapter presents these test procedures and shows the results from the different tests. The recording environment used in the tests is described in Section 5.1. The generation of the stimuli used in the tests is explained in Section 5.2 as well as the reasons for stimulus selection. The results of the different tests (with some discussion) are presented in Section 5.3. As the purpose of this model is to evaluate both spatial and timbral aspects from the reproduced sound, the test scenarios are divided into two categories. The first category describes the testing of spatial localization and the used sound stimuli in these tests and the latter category focuses on the evaluation of specific loudness estimation accuracy. Section 5.4 gives a short summary of this chapter.

5.1 Test environment

The tests were made in the large anechoic room[1] in Nokia Oyj premises at Tampere. The room volume is approximately 218 m^3, with a length of 5.96 m, a width of 5.96 m and a height of 6.14 m. The room has been designed to have a cut-off frequency of about 100 Hz, below which the intensity of the reflections grows and the condition for anechoic room is not met. Although the loudspeakers are able to produce sound also below this limit frequency, the intensity of the reflections are still considerably smaller than the intensity of the direct sounds. The test environment can be considered as anechoic in practice. The used test stimuli also have most of their energy and content in frequencies above this cut-off frequency.

[1] In anechoic chamber the measured sound pressure decreases within ±1 dB of the theoretical inverse square law when moving a point microphone away from a point source (Beranek 1998).

The recordings were made with the head and torso simulator (HATS) (B&K 2006) standing at the centre and loudspeakers positioned at 1 m distance at ±30° angles in front of the HATS. The distance to the loudspeakers was measured from the centre of the head to the front plate of the loudspeakers. Figure 5.1 illustrates the loudspeaker setup used in the tests as viewed from above. Here θ denotes the azimuth.

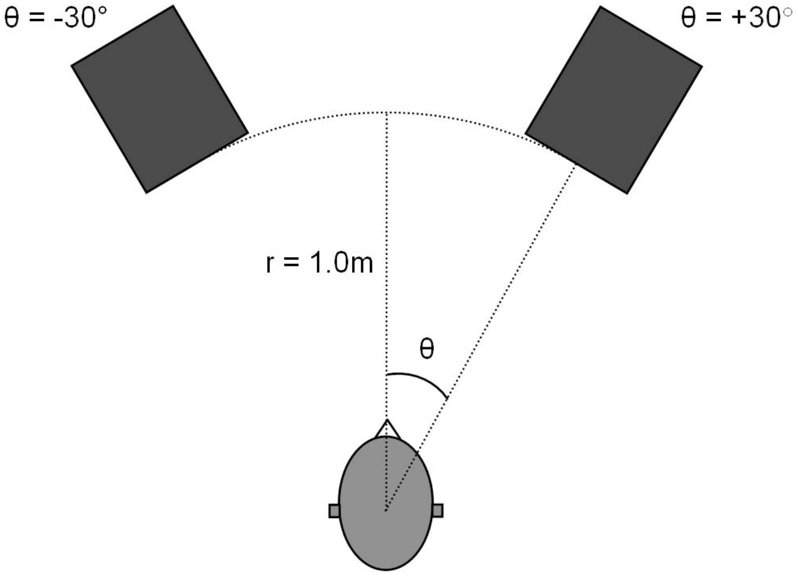


Figure 5.1: Relative positions of HATS and loudspeakers in test setup. Situation is illustrated from above.

The loudspeakers selected to reproduce the sounds in the tests were Genelec model 8020A loudspeakers (Genelec 2005). They were used in the test because of their controlled directivity, which is essential for evaluating spatial localization accuracy. These loudspeakers have also an almost flat frequency response (± 2.5 dB) in the frequency range between 66 Hz and 20 000 Hz. Therefore they are able to reproduce (almost) the whole frequency area without too much effect on the relative sound energies at different frequencies. Hence the timbral aspects of the original sound are reproduced transparently.

The loudspeakers were placed on top of plastic surfaces for this recording session so that the loudspeakers were at the same height as the ears of the HATS. The purpose of this placement was to ensure a good directivity of the sound to the ears and to have the elevation (Figure 3.2) of the sound sources to be zero, which is the same as the elevation used in the creation of binaural cue lookup tables from the simulated HRTF data (Section 4.3.1). This way the localization of sound sources (using the binaural cue value estimates (Section

4.3.2), would be as accurate as possible. As the floor material in the anechoic room is a suspended elastic grid, there was also a need to ensure that the loudspeakers and the HATS do not move during the measurements. For this purpose, the HATS and the loudspeaker 'stands' were fastened to the grid.

In order to minimize the noise in the recordings caused by the testing equipment, the different sounds were played on a computer outside the anechoic room and transmitted to the loudspeakers through cables. Also the binaural recordings from the microphones in the ears of the HATS were transmitted via audio cables to another computer outside the anechoic room. All signals in the cables were transmitted in analog format, so the only required analog-to-digital and digital-to-analog conversions were made in the two computers with high-quality digital sound cards in order to minimize the errors in the conversions. Figure 5.2 shows a simplified drawing of the test setup.

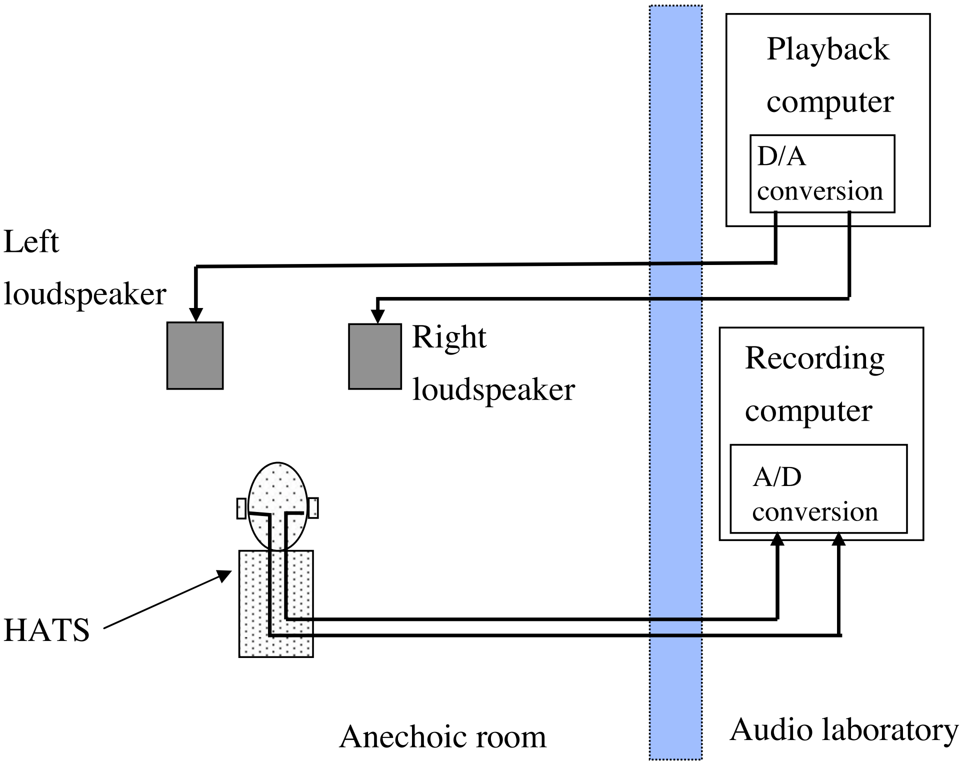


Figure 5.2: Schematic presentation of test environment.

Before the actual recording could begin, two calibration measurements were required. In the first measurement the sensitivity difference between the two microphones in the ears of the HATS was measured. This was made by playing a 1 kHz sine wave at 97.1 dB SPL through a calibration tool (B&K 2006) placed at one of the ears of the HATS at a time. In the second measurement the loudspeakers were adjusted to same level by measuring the level differences in the binaural recording of a pink noise signal, and by adjusting manually the output levels of the loudspeakers in a manner that the difference in the recorded signal powers matched the difference in the sensitivity between the microphones. After a series of trials the loudspeakers were adjusted in a way that the left loudspeaker was only 0.27 dB louder than the right one. This difference was considered insignificant as the ILD values (Figure 4.6) are larger than this difference. Consequently, all the recordings in the tests were afterwards scaled in a way that the left and right ear recordings were at equal level by using the HATS calibration measurement difference as a reference. This was realized so that the different sensitivities of the microphones would not create interaural level difference (ILD) to the recordings and thus mess up the localization process.

5.2 Test stimuli

The spatial localization accuracy of the BAM was tested with five different stimuli in the described test setup (Figure 5.1). The relative sound source directions and the types of the sound sources were known in these stimuli so that the localization results from the model could be easily evaluated. Since with more complex sounds, such as music, even a trained human listener has difficulties in localizing sound events accurately especially if the listener is required also to evaluate the time at which the sound event occurred in that direction. Therefore the evaluation whether a sound event actually occurred in the direction the localization results indicate is not possible to do reliably with complex sounds. Hence this quantitative approach where the expected localization results are known is more reliable for testing the model accuracy.

This knowledge of sound sources in the stimuli is not however adequate for evaluating the accuracy of the specific loudness and the overall loudness estimation of the model, and reference data from other measurements is needed for this purpose. In this work the reference data is obtained by measuring the loudnesses of the binaural recorded stimuli with the loudness meter developed by Tuomi & Zacharov (2000). The outputs of the model are then compared to the outputs of this meter in order to evaluate the accuracy of the specific and overall loudness estimation for the given stimuli.

This selected approach of using reference measurements for both spatial localization and specific loudness validation was also beneficial in the development phase of the model, since the improvement or impairment of the result accuracy due to changes in the model parameters was usually easily perceived. Hence the model could be nicely tuned to obtain better accuracy. The stimuli used in the tests are listed in Table 5.1 where the first five signals were used in the spatial localization testing and the last three in the testing of specific loudness estimation.

5.2.1 Generating stimuli

The stimuli were generated from 'raw' samples (i.e. anechoic recordings of speech and relatively dry music instruments) with the 'Cool Edit Pro 2' software package (Syntrillium Software Corporation 2002). As the source materials were stereophonic recordings, the first step in the generation process was to create monophonic sound sources from these. This was made by taking only the left channel of the given recording for each sound source.

In the next phase of the process the sound pressure levels (SPL) in the different sound signals were set to an equal level so that each sound source would be 'perceivable' in the test signals. This was implemented by looking at the maximum and at the average sound powers in the signals, and amplifying or attenuating the signal then according to these values. This phase also ensured that the signals did not clip in the playback phase. After this process, the stimuli could be created from these signals. Despite the fact that the amount of sound sources is different in the different stimuli (Table 5.1), the procedure was identical for all of them. In all cases, each sound source was panned independently to the desired direction according to the panning law (Equation (5.1)) presented by Bennet et al. (1985)

$$\frac{\tan\theta_T}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{5.1}$$

where θ_T is the perceived sound direction, θ_0 is the direction of the loudspeakers and g_1 and g_2 are the gains of the signals from the two loudspeakers. This panning law states that by adjusting the gains of signals to the loudspeakers 1 and 2, it is possible to get the perceived location of the sound source by a listener to anywhere between the loudspeaker locations (Figure 5.3). As in this work, the two loudspeakers were positioned at ±30° (Figure 5.1), a 5.48 dB (from approximate) difference in the gains equals to a perceived location of 10° and a 12.89 dB difference is needed for perceived location of 20°.

After the needed sound sources are obtained by this stereo panning procedure, the stimulus is formed simply by adding the different sources together into the same sound stimulus. So in test stimuli #4 and #5, there are at least two independent sound sources present simultaneously. More information about the (amplitude) panning and creation of virtual sound sources can be found in Pulkki & Karjalainen (2001), Pulkki (2001) and in Rumsey (2001).

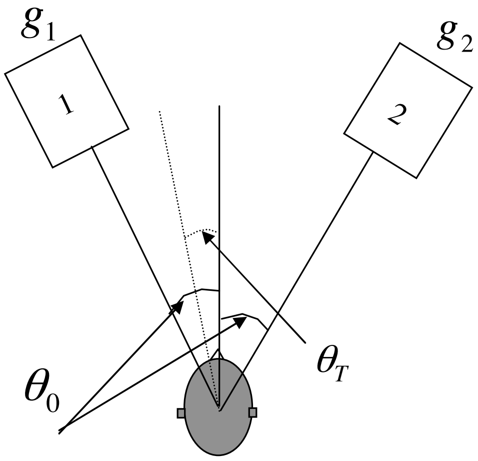


Figure 5.3: Dependence of perceived source location on gains of loudspeakers in (amplitude) panning law by Bennet et al. (1985).

Spatial localization		
No.	**Sources**	**Panning**
1.	Piano	Panned to -30°
2.	Violin	Panned to 30°
3.	Speech (english, female speaker)	Panned to 0°
4.	Drum and piano	Drum panned to -30° and piano panned to 30°
5.	Tuba, singing (male singer) and violin	Tuba panned to -30°, singing panned to -10° and violin panned to 30°
Specific loudness		
No.	**Sources**	**Panning**
6.	HATS calibration signal	Signals played at ear entrance points
7.	Pink noise	Independent channels panned to ±30°
8.	Music (10 s sample of Steely Dan's track *'Cousin Dupree'*)	Reproduced from loudspeakers at ±30°

Table 5.1: Description of test signals selected for the verification of the model functionality.

5.3 Results and discussion

5.3.1 Spatial localization testing

In Chapter 4 it was presented that the resulting spatial localization output from the model for a given stimulus is a large three dimensional data set (24 critical bands × 19 azimuth angles in the frontal plane (37 in the whole azimuth plane) × the length of the stimulus in samples), which makes it hard to visualize. Therefore in order to present the results in more understandable format the probabilities (denoted as $P(\theta)$)of sound source presence at given azimuth angle are calculated according to

$$P(\theta) = \sum_z \left(p[\theta, z] \times \sum_i p[\theta, i] \right) \tag{5.2}$$

where $p(\theta, i)$ and $p(\theta, z)$ are the probabilities of source locations over time instants and over frequency bands respectively. The probabilities are also normalized in a manner that the total sum of probabilities equals one.

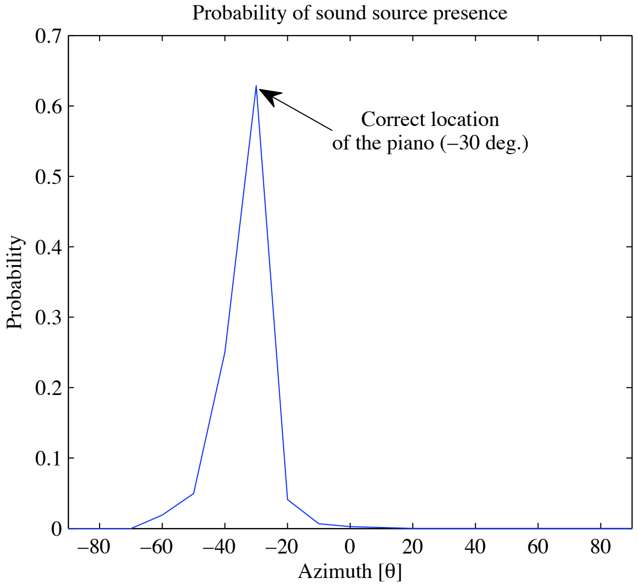


Figure 5.4: Probability of sound source presence at different azimuth angles for test signal #1.

In the first two stimuli there is only one source present at a time, and due to hard panning of the source there is sound coming from only one of the loudspeakers. Therefore the test scenario in these cases is similar as there would be either a piano playing at 30° on the left (stimulus #1) or a violin playing at 30° on the right of the HATS (stimulus 2). So with these stimuli the model output is expected to indicate the presence of a single source in the respective directions.

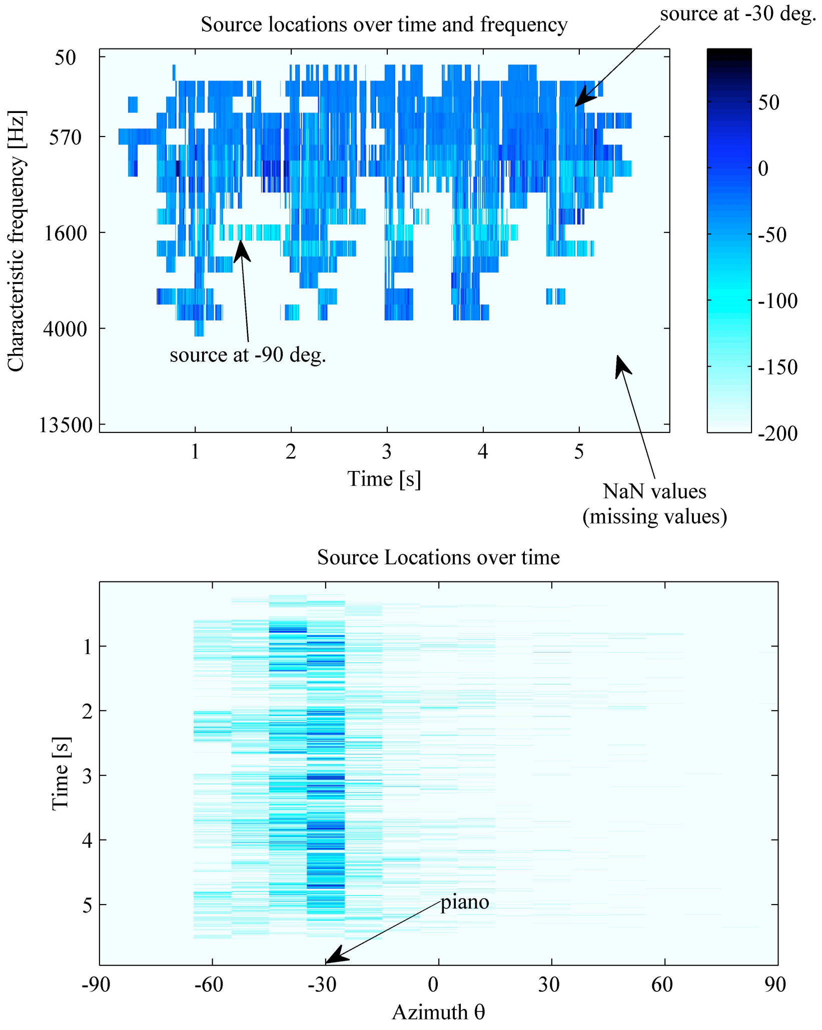


Figure 5.5: Topographic presentation of sound source locations (upper graph) and probabilities of source locations as function of time and azimuth (lower graph) for test signal #1.

Due to the way the sound source location is estimated (Section 4.3.2) the model is likely to localize sounds in ±10° accuracy, as the binaural cue values are closely similar in adjacent azimuth angles. Therefore with test signal #1 for example where there is a source only at -30° azimuth angle, the model is likely to localize the source to also in -20° and -40° azimuth positions. This is also illustrated in Figure 5.4, where the probability of sound source presence is presented as a function of azimuth angles for test signal #1. The reason why the 'tail' of the intensity graph is more biased towards the left is that the binaural cue values at -40° location are relatively closer to the ones at -30° location than the ones at -20°.

As there is only one sound source present at the time, the signals to the two ears are very much similar and hence the interaural coherence (IC) between the left and right ear signals is high throughout the whole frequency area. Therefore the interaural coherence criterion (Section 4.2.2) is met in all critical bands. So the loudness criterion (Section 4.2.2) is the one that limits the amount of accepted binaural cue values and therefore affects the frequency information in the results. This is illustrated in Figure 5.4, which shows the estimated source locations with stimulus #1. The white area in the graph illustrates the area where the criterion values are not met. In these areas the source locations are set to NaN (Not a number) values according to the idea presented earlier in Section 4.2.2. The graphs in Figure 5.4 illustrate that the model detects the source to be sometimes also in wrong directions. This results from the ambiguity of the binaural cues in frequencies between 1.5 kHz and 2 kHz (Section 3.1.1) (Figures 4.5 and 4.6).

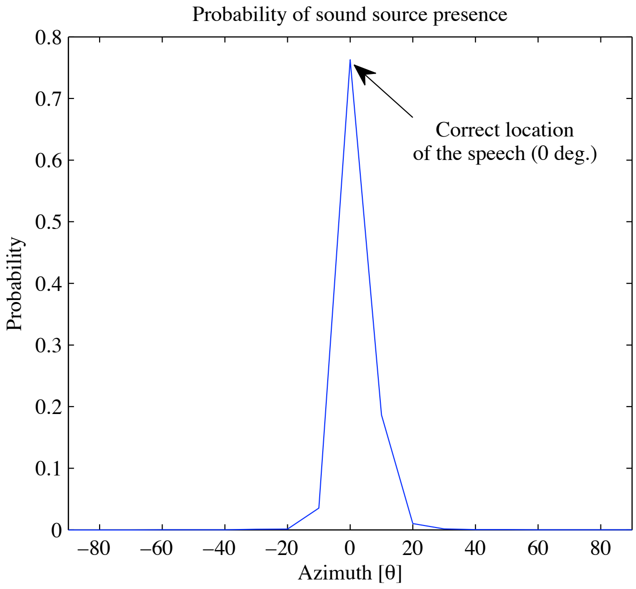


Figure 5.6: Probability of sound source presence at different azimuth angles for test signal #3.

In stimulus #2 there is also only one sound source present at a time. Although the sound source is different and the source is located on the opposite side of the head compared to the ones in stimulus #1, the test scenario is mostly similar. Therefore the results with stimulus #2 show similar localization accuracy on the right side of the head as stimulus #1 shows in the left side of the head. For this reason the results from tests with stimulus #2 are presented only in Appendix B.

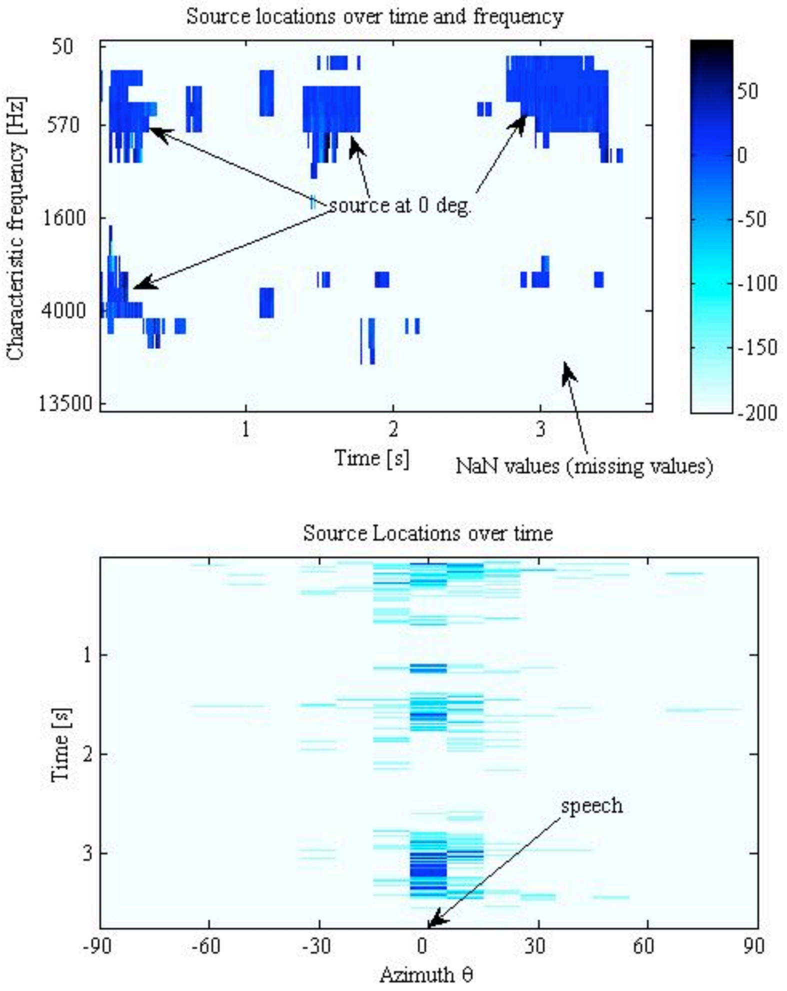


Figure 5.7: Topographic presentation of sound source locations (upper graph) and probabilities of source locations as function of time and azimuth (lower graph) for test signal #3.

In stimulus #3 there is however sound coming from both loudspeakers, although the same signal is reproduced from both loudspeakers. According to the stereo panning law (Bennet et al. (1985)) the model is in this case expected to detect the presence of a single sound source located at the centre (at zero azimuth). In this case the signals to the left and right ear are again, similarly to the scenario with stimuli #1 and #2, (almost) identical. Therefore the loudness criterion is again the limiting factor for the acceptance of binaural cue values.

In speech, the intensity of the sound varies by nature due to intonation, (word or sentence) stress and different phonemes. Hence there are again areas where the loudness criterion is not met although there is sound present in these areas. Figure 5.7 illustrates these properties, as the model indicates the presence of a source only at some time instants.

The localization itself is accurate with test signal #3 although there are again some false localizations. This is represented in Figures 5.6 and 5.7, which illustrates that the model actually detects the presence of a single sound source located at zero azimuth. In the tests with stimuli #1, #2 and #3 the test scenario was quite simple, as there was only one sound source present at a time, so the role of the interaural coherence criterion was quite small.

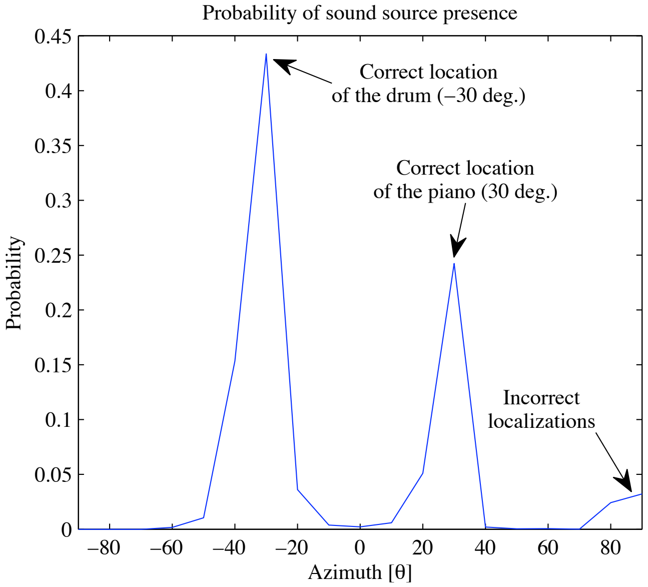


Figure 5.8: Probability of sound source presence at different azimuth angles for test signal #4.

In stimuli #4 and #5, there is however at least two independent and different sound sources present at the same time at different locations. Therefore the role of the interaural coherence criterion increases. The two sources in test stimulus #4 and the three sources in test stimulus #5 have each energy and content on all the frequency bands. As these contents are then mixed together before they enter the ears (Figure 3.1), they are also likely to cause false information in the binaural cues.

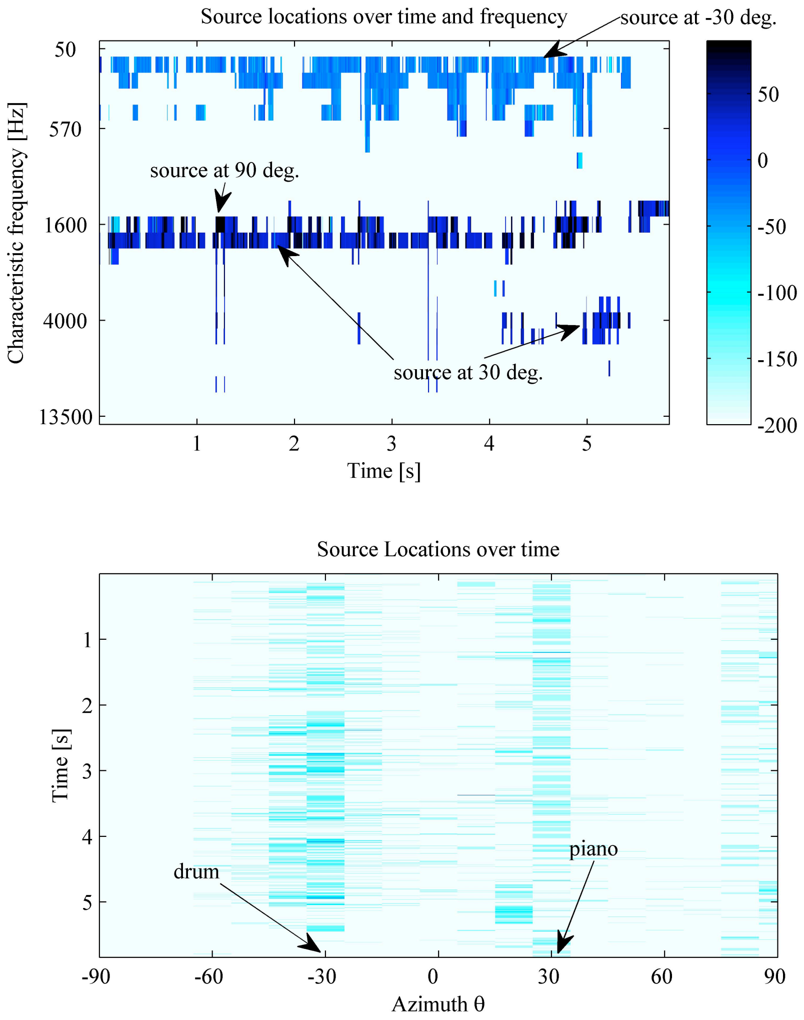


Figure 5.9: Topographic presentation of sound source locations (upper graph) and probabilities of source locations as function of time and azimuth (lower graph) for test signal #4.

This problem can be solved by using a higher interaural coherence criterion value for stimuli #4 and #5. Here the interaural coherence criterion value is set to 0.99. This on the other hand causes the model to detect the presence of a sound source only in the frequency bands and on the time instants where the sound source stands out from the others. This can be seen in Figure 5.9 where there are two distinctive frequency areas with different source locations. The drum, which was panned to -30° azimuth, has more low frequency content than the piano, and is therefore distinguishable in the low frequencies. The piano, which was panned to 30° azimuth, on the other hand has more high frequency content than the drum, and therefore stands out in the higher frequencies.

Unfortunately, the piano stands out mostly just in the frequency area where the binaural cues provide also ambiguous information (Section 3.1.1). Therefore the model localizes the piano to be sometimes also at 80° and 90° azimuth angles. This can be seen in Figures 5.8 and 5.9. Despite these false localizations the model is able to detect the presence of the two sound sources within ±10° accuracy. This is illustrated in Figure 5.9 where the piano is localized mostly to 30° azimuth and the drum is localized mostly to -30° azimuth (on the left of the HATS). The probabilities of the incorrect localizations is insignificant in comparison to the 'correct' ones.

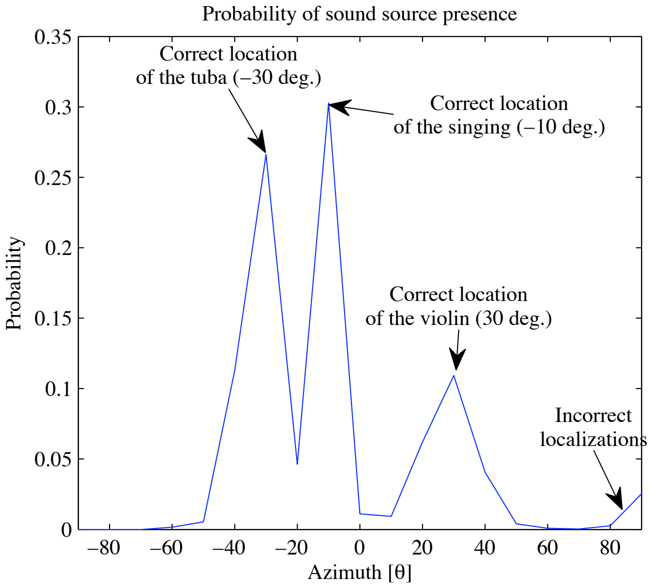


Figure 5.10: Probability of sound source presence at different azimuth angles for test signal #5.

In test stimulus #5 the tuba and the violin are again nicely distinguishable by their main frequency content. The singing however contains a lot of energy on both low- (due to the fundamental frequency of the singing) and high-frequency (due to the second and third formants in vowels) areas. Therefore the model localizes the source at the position of the singing in both in low and in high frequencies the main localizations being in the higher frequency areas. This is illustrated in Figure 5.11.

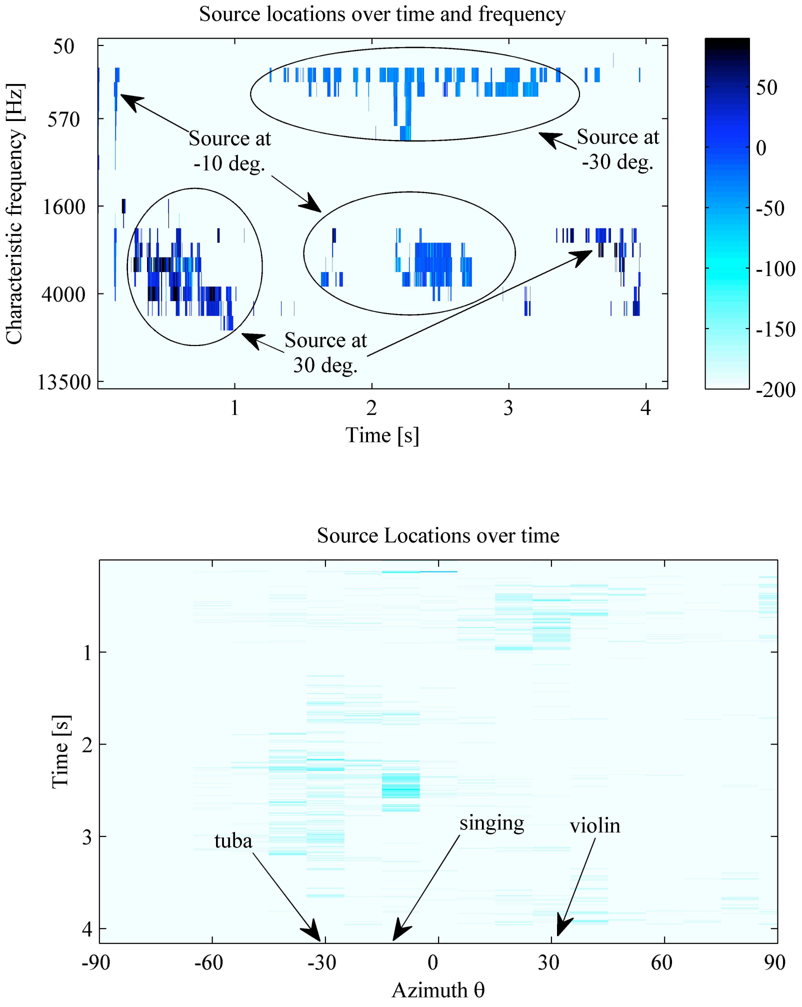


Figure 5.11: Topographic presentation of sound source locations (upper graph) and probabilities of source locations as function of time and azimuth (lower graph) for test signal #5.

As the singing is localized at both high and low frequencies and the tuba and violin on the other hand only at one frequency area, the probability of sound source presence is highest in the direction where the singing was panned. Since the singing stands out more in the higher frequencies, the model localizes the source in the position of the tuba also quite often. The 'main frequency area' of the violin on the other hand overlaps with singing in the high frequency areas. Therefore the model localizes the source in the position of the violin less often than the others. Another reason for the smaller probability of source presence (Figure 5.11) in the position of the violin lies in the false localizations due to the ambiguity of the binaural cues in this frequency area. These properties are illustrated in Figures 5.10 and 5.11. Figure 5.10 reveals that the model is again still able to detect the presence of the three sources in correct locations in the azimuth plane.

Discussion on spatial localization results

The results from the localization accuracy tests show that the model is at this point capable to localize the sound events (or sources) within $\pm 10^{\circ}$ accuracy. This accuracy can be however increased by increasing the amount of inspected azimuth angles in the creation of the binaural cue reference value tables (Section 4.3). There is however a risk to lose some of the models capability to detect the amount of present sources in this process, as due to the increased similarity of the binaural cue values in the neighbouring azimuth angles may result in less clear probability graphs. Therefore in the process of increasing the azimuth angle density, the localization algorithm (Section 4.4) also needs some re-evaluation and some limit for the probability (of sound source presence) needs to be generated in order to be able to evaluate the amount of sound sources from the probability graphs.

When the model is used to evaluate the spatial locations in more complex sounds such as music, one should recall that the model is capable of only detecting the presence of an instrument at given location when the instrument stands out from the others in some critical band. Therefore in a sound where a single instrument (such as singing) generally "dominates" other instruments while the others stand out only on short time instants, the overall probability graph can not be used to evaluate the sound in spatial aspects. In this kind of scenario there is a need to inspect the intensities in shorter time windows.

5.3.2 Specific loudness evaluation

A reliable validation of the accuracy of the specific loudness calculation would ideally require organising a listening test with a large number of test subjects on the stimuli used in this tests. Only after this the model could be calibrated accurately. This is a very time-consuming process. Therefore the loudness calculation cannot be validated at this point. However, this is not necessary either when considering the goal of this work, which is to compare different devices (and signals) in terms of loudness and timbre, and for this purpose the relative differences in specific loudness between the different devices are more than adequate.

Hence we content ourselves in this phase with the comparison of the specific loudness values of the model to the outputs of the loudness meter by Tuomi & Zacharov (2000). This loudness meter gives the specific loudness values (also) according to the generally accepted and standardised (ISO-532 1975) Zwicker's loudness model. The approach used in this work and the Zwicker's model use the same critical band based separation to represent the frequency selectivity of the basilar membrane so the results from the two models are comparable from this point of view. One should however recall that the approaches to evaluate the binaural loudness in the two models are different. In the developed binaural auditory model (BAM) the binaural loudness values at the given frequency band are calculated by taking into account the ITD information (Equation (4.11)) whereas the loudness meter calculates the specific loudness values for left and right signals separately. The loudness meter uses also a calibration signal to evaluate the absolute loudness level, which is not included in the BAM. Due the different calculation approach, it is not expected to obtain identical results from the two models. It is however expected that the specific loudness values on the different critical bands in these two approaches show similar patterns.

Like in the case of the spatial localization testing the stimuli in this loudness evaluation were also scaled to an equal level according to the difference in the HATS microphones sensitivity before they were passed to the model (Section 5.1). This was not however implemented to the "versions" of the stimuli that were tested with the loudness meter as the meter sets the channels to equal level by itself in the calibration phase.

The stimuli #6 and #7 (Table 5.1) were selected to this evaluation because they are 'steady-state' signals from the loudness point of view. This means that the sound pressure level (SPL) resulting from these audio signals does not vary much over time. The Zwicker's model and therefore the loudness meter is known to work well with steady-state sounds. Hence the evaluation of the specific loudness estimation accuracy is easiest with this type of signals. Stimulus #8 (Table 5.1) on the other hand is not a steady-state signal, so the evaluation of the accuracy of the model is more difficult. It was nonetheless selected here, because it is used later in Chapter 6 to evaluate the differences between different devices in terms of loudness and timbre.

As the calibration signal is played with the calibration tool at one ear at a time, stimulus #6 was formed by creating a signal where the left and right ear calibration measurements are occurring simultaneously. The phases of the signals in the left and right channels were not however synchronised, since in this evaluation the spatial information is not important. This stimulus was also used in the loudness meter to calibrate the meter by informing it that the signals in the channels of this sound have a loudness level of 97.1 dB. Figure 5.12 shows the plots of the specific loudness calculation results in the meter (Figure 5.12(a)) and in the model (Figure 5.12(b)). These figures show that the patterns are similar and that the highest loudness value occurs in the same critical band in both graphs. One should however notice that the values in the Figure 5.12(a) show the loudness on the left and right ears separately and the binaural loudness is in this meter formed by adding them together. Therefore the (binaural) specific loudness values in the meter are twice as high as the ones provided by the model. This difference results from the fact that the loudness meter is calibrated and the model of this study is not.

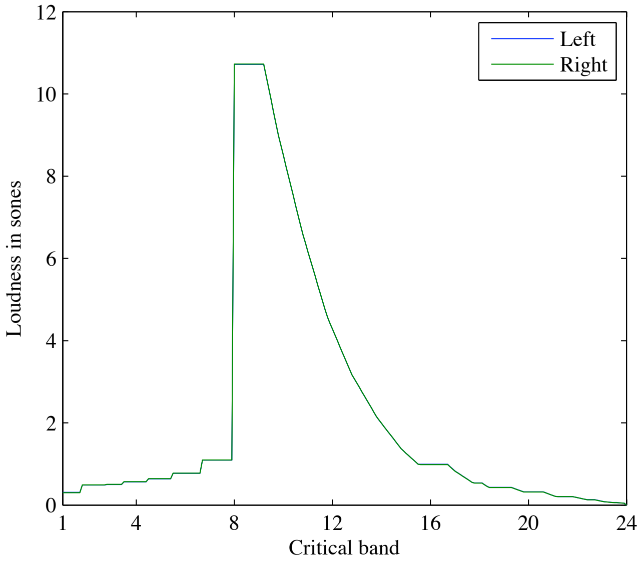
Stimulus #7 was reproduced with the loudspeakers so the recording includes also the spatial information. The left and right channel signals in the stimulus were independent and uncorrelated so the effect of this spatial information on the loudness calculation in the model is not significant. Figure 5.13 shows the plots of the specific loudness evaluations with this stimulus in the meter (Figure 5.13(a)) and in the model (Figure 5.13(b)).

Figures 5.13(a) and 5.13(b) show that the graphs have again a more or less similar pattern although the values are not as close as with the previous stimulus. Besides the unimportant difference in the overall magnitude of the specific loudness values between the two models, there is also some difference in the relationship of loudness values in different critical bands between the two graphs. For instance the loudness values around the eighth critical band (whose characteristic frequency is 870 Hz) are relatively larger in the results from the BAM, and the values around the 18th critical band (CF is 4 kHz) are on the other hand relatively smaller in the results from the BAM. These differences are illustrated in the results from stimulus #7 (Figure 5.13) and #8 (Figure 5.14) and they result from the different approach to calculate the values and the fact that the loudness meter was calibrated with the previous stimulus. Hence the meter can evaluate the loudness more accurately by comparing the sound pressure level of the stimulus to one used calibration signal.

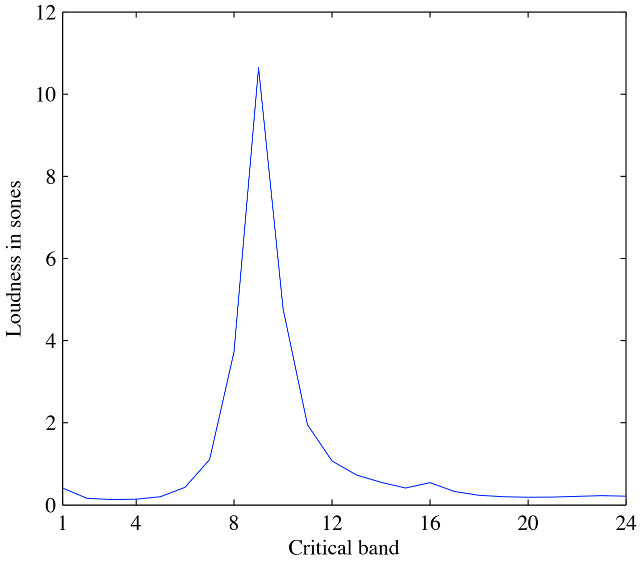
Stimulus #8 is the most challenging one, as the sound pressure level and the frequency information in it vary over time and therefore the specific loudness calculation should take into account the masking and spreading effects in both the frequency and temporal point of view in order to get an accurate evaluation of the loudness with this kind of sound. So the results from the model and the meter are only indicative, but still adequate to evaluate the relative differences between different devices. Figure 5.14 shows the results from the specific loudness evaluations in the loudness meter (Figure 5.14(a)) and in the BAM (Figure 5.14(b)). Again there is a difference in the magnitude of the specific loudness values between the two models and the relative difference between the values in different critical bands are not the same in the two graphs. The fundamental pattern is however similar in both cases.

Discussion on specific loudness evaluation

The patterns of the specific loudness evaluation graphs show that the model provides results similar to the Zwicker's loudness model for different type of stimuli. There are however

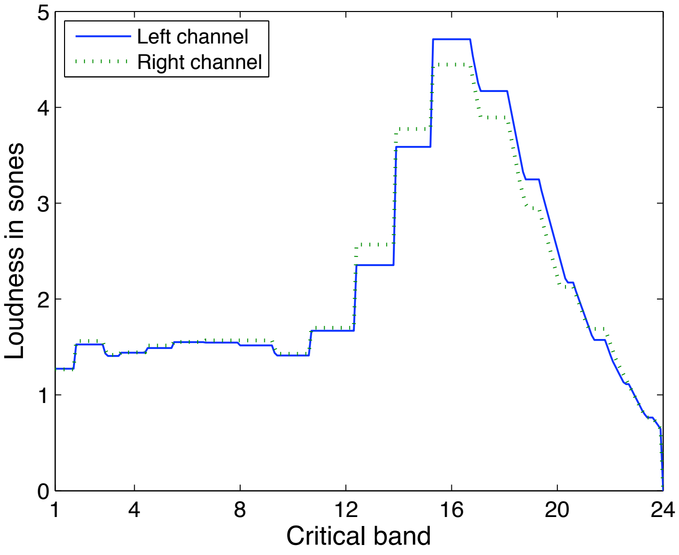


(a)

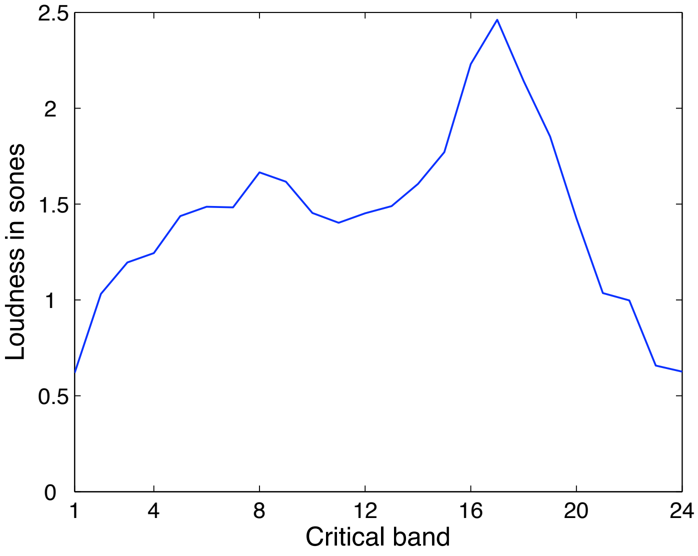


(b)

Figure 5.12: a) Specific loudness spectrum from loudness meter (above) and b) composite loudness level spectrum from BAM (below) for stimulus #6.

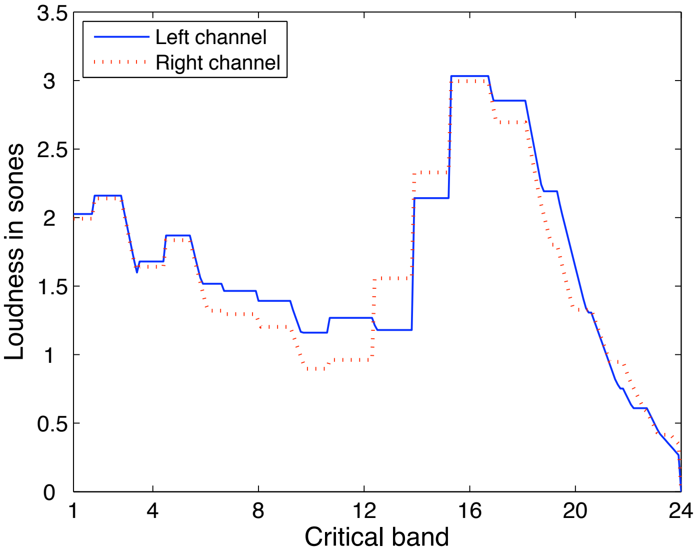


(a)

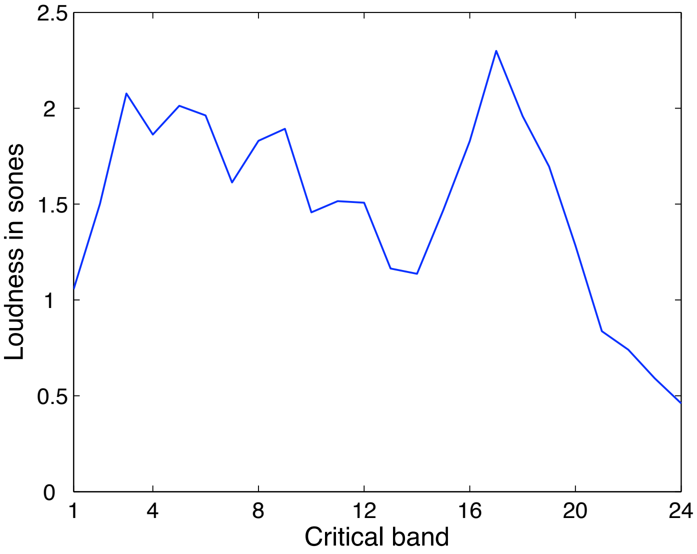


(b)

Figure 5.13: a) Specific loudness spectrum from loudness meter (above) and b) composite loudness level spectrum from BAM (below) for stimulus #7.



(a)



(b)

Figure 5.14: a) Specific loudness spectrum from loudness meter (above) and b) composite loudness level spectrum from BAM (below) for stimulus #8.

differences in terms of overall level and in the relative magnitude of values in different critical bands between the two models. This is due to the difference in the approach used to calculate the specific loudness values and to the fact that the loudness meter is calibrated in the testing process whereas the model presented in this work is not.

Therefore the model must also be calibrated before it can be used to evaluate the specific loudness accurately. One should however recall that Zwicker's loudness model is also a model, which has its own limitations due to approximations in the calculations. Consequently, Zwicker's model does not present the absolute truth in these tests. Nonetheless it is good that the developed model provides similar results as a standardised loudness model, because now the model can be used to evaluate the differences in terms of both timbral and loudness aspects with some confidence on its functionality.

5.4 Summary

In this chapter an experimental method was used in order to verify the functionality of the developed binaural auditory model (BAM). The idea for this kind of validation of the model was used also by Macpherson (1991). The used test setup was described and the (necessary) phases in the test were presented in this chapter. The used test signals were described and the selection of test signals was also reasoned. The localization accuracy of the BAM was tested with test signals (Table 5.1), with which the expected localization results were known beforehand. The loudness estimation of the BAM was evaluated by comparing the results from the BAM to a generally accepted and standardised loudness model.

This verification of the model functionality is however still preliminary as the tests were only made in anechoic conditions and the model functionality was not tested in reverberant environments at this stage. Therefore the testing of the model functionality in listening rooms needs to be done before the functionality of the model can be verified. The results from this preliminary verification are however promising, as the model can detect the sound source locations accurately also in the presence of multiple different sound sources. The $\pm 10°$ accuracy of the localization is due to the resolution of the binaural cue lookup table. There is still some work to be done before the model can be used to localize sound sources from more complex sounds, such as real music samples, with confidence on the accuracy of the results, as the test signals in this verification were relatively simple.

The results from the loudness testing illustrated that the verification of the loudness calculation is however a more difficult task and would require organizing listening tests with a large amount of test subjects in order to be accurate. At the same time these results illustrated also that the loudness results from the developed BAM and a standardised loudness model provide similar results. Therefore the BAM can at this point be used to evaluate differences in terms of timbral aspects (e.g. sharpness, low- and high-frequency emphasis) by comparing the relative specific loudness values between different stimuli. However BAM needs to be calibrated before the overall loudness of the stimuli can be evaluated.

Chapter 6

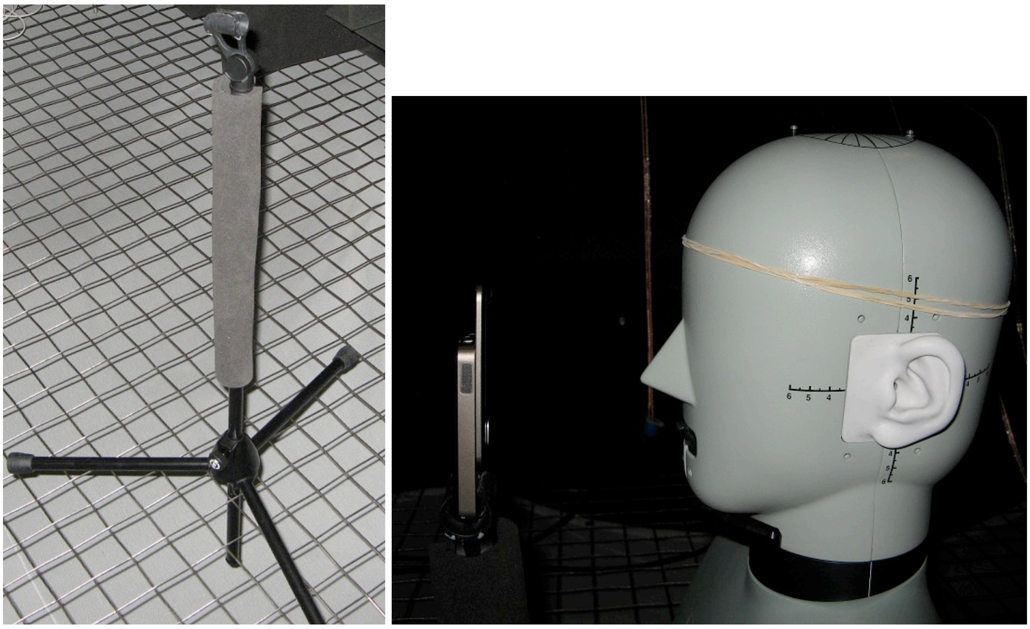
Case study on mobile loudspeakers

In this chapter a scenario is presented where the developed binaural auditory model is used in a case study on mobile loudspeakers. The users of the mobile loudspeakers demand high quality sound reproduction, but due to the small size of the devices (and their parts) these devices cannot achieve similar sound reproduction quality as normal loudspeakers. Therefore the manufacturers aim at different aspects of sound reproduction with their devices. Consequently, there are substantial differences between different devices, as has been presented e.g. in Lorho (2007). In the present study a few existing devices were selected to reproduce a couple of test stimuli. The purpose of this study was to test what kinds of differences the developed model can find between the devices at this stage and what are the limitations of the BAM. The selection of mobile devices is presented and reasoned in Section 6.1. The description of test environment used in this case study is also presented in that section.

The actual study is divided into two categories. The first one focuses on evaluating the differences in the "spatial image" of the reproduced test stimulus on a variety of devices. This study and its results are presented in Section 6.2. The second category focuses on evaluating the differences both in terms of overall loudness and timbral characteristics between the selected devices in music reproduction. This study and its results are presented in Section 6.3. A brief summary of the chapter is also given in Section 6.4.

6.1 Setup

Like the recordings in the model testing, the recordings in this case study were also made in the large anechoic chamber (Section 5.1) in Nokia Oyj premises at Tampere. Also the same Head and torso simulator (B&K 2006) was used to record the stimuli and the HATS position was also kept the same as before. In an attempt to demonstrate a 'normal' listening situation on these mobile devices, the devices were positioned in front of the HATS on a stand in a manner that the loudspeaker(s) of the devices were facing the HATS.



(a)

(b)

Figure 6.1: a) Microphone stand that was used as phone holder (left) b) Test setup where device is placed in front of the HATS at zero azimuth, zero elevation and at 24 cm distance from centre of head (right).

As the devices and the loudspeaker positions in them are different, the stand needed to be adjustable, so that each device could be placed in the desired position. Therefore a microphone stand with adjustable height was used to align the different devices into the desired height. The microphone holder in it was also turned in a manner that the device was facing the correct direction. The devices were positioned so that the loudspeaker(s) of the device was at the same level as the ears of the HATS (at zero elevation). This was done in order to obtain good reproducibility of the tests in this case study and to be able to compare the results from the devices to the ones from the loudspeaker recordings in the model testing phase (Chapter 5).

The distance from the loudspeakers to the centre of the HATS head was 24 cm for each device. This relatively small distance between the HATS and the devices was selected to ensure that the stereo information would not be lost, as it would if the device had been placed further away from the HATS. Figure 6.1(a) shows the used phone holder having isolating material around the pole to reduce reflections caused by the equipment. Figure

6.1(b) shows the recording setup where a mobile device is placed on the phone holder in the desired position in front of the HATS.

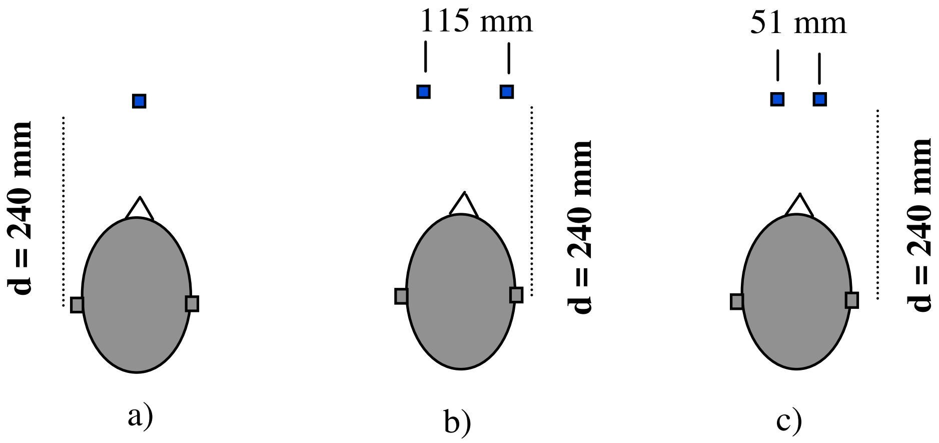


Figure 6.2: Recording setup illustrated from above and loudspeaker position(s) in a) monophonic device b) stereophonic device 1 and c) stereophonic device 2.

6.1.1 Mobile loudspeakers

A variety of mobile devices were selected for this case study to represent the different devices that are currently available on the market. At this point of the study the selection was however limited in a manner that each selected device was distinctively different from the others for instance by the number of loudspeakers, loudspeaker positioning or other point of view. This was to ensure that there would be large enough differences between the devices for the model to detect. The selection of devices included one monophonic device and two stereophonic devices. In stereophonic device 1 the two loudspeakers are placed as widely as it is probably possible in this kind of devices. In the other stereophonic device the distance between the loudspeakers is smaller but this device can use a stereo enhancement algorithm[1] to widen the perceived spatial image. Figure 6.2 illustrates the positions of the loudspeaker(s) in the different devices and the recording setup with each of them (scale of

[1] In mobile devices the loudspeakers cannot be placed in a manner that the traditional stereophonic reproduction of the sound would be possible. Therefore stereo enhancement algorithms are used in some mobile devices to create an (artificial) image of sound sources locating more widely than the actual loudspeaker base width would allow. This type of algorithm is usually based on the cross-talk cancellation formulated by (Atal & Schroeder 1966) and a subjective comparison of such applications can be found in (Olive 2001) and (Lorho 2006).

the graph is not accurate). The recorded samples were also scaled afterwards in order to correct the sensitivity difference between the two microphones in the ears of the HATS. This procedure was made similarly as described earlier in Section 5.1.

6.2 Evaluation of differences in spatial image

For this part of the study test signal #5 (Table 5.1) from the model testing phase was selected to test also the differences between the devices. This stimulus was selected, because the model was able to function correctly with this stimulus in normal stereophonic reproduction (Section 5.3.1) and because this stimulus also has sound sources that are panned somewhere between the two loudspeakers. Therefore this stimulus is more interesting and challenging than a more simple stimulus where the sound source(s) come from just one loudspeaker.

The loudspeaker base width differences between the different devices are quite small. Therefore the $\pm 10°$ accuracy in localization (Section 5.3.1), which results from the resolution of the binaural cue reference tables (Section 4.3.1), is likely to be inadequate to pick out the differences in the spatial image between the different devices. Hence there was a need to increase the resolution of the localization. For this purpose, the binaural cue reference tables were recreated with a resolution of 6° in the horizontal plane (Figure 3.2).

In this study all the three different devices were selected for the test. The stereophonic device 2 was tested with and without the stereo enhancement algorithm to see whether the model is able to detect the differences between the stereo enhanced reproduction and the 'normal' reproduction of the stimulus. This way there were four different cases in this test.

Due to the differences between the devices in terms of loudness and specific loudness, the same interaural coherence (IC) and loudness criterion could not be used for all the devices. The device's ability to produce sound power in a certain frequency band affects greatly on whether a sound source is localized or not because of the loudness criterion. Therefore some 'fine-tuning' was needed in order to localize the sound sources in the stimulus. These differences in the ability to produce sound power on different critical bands also affected the amount of times a sound source is localized in a certain position. Hence the intensities for each source location reflect also the timbral characteristics of the sound.

In the monophonic device all sound sources are reproduced with the same loudspeaker. Hence the amplitude panning that was applied to the sources in the stimulus becomes insignificant. Therefore it is expected that the model localizes all sound sources to the direction of zero degrees in the azimuth plane (horizontal plane). This is also shown in Figure 6.3, which illustrates the probability of sound source presence as a function of azimuth angle. The sound source is localized at -6° azimuth. This 6° offset is likely due to the combined effect of localization resolution and a small error in the phone positioning accuracy.

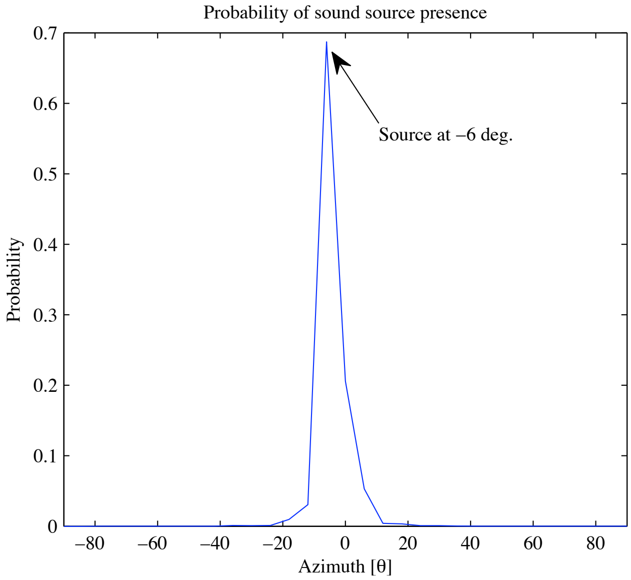


Figure 6.3: Probability of sound source presence at different azimuth angles for monophonic device.

In stereophonic device 1 the two loudspeakers are placed at approximately 15 cm distance from each other. This stereo base width combined to the panning law indicate that the tuba is expected to be localized at -12° azimuth, the singing at approximately -6° azimuth and the violin at approximately 12° azimuth. Figure 6.4 illustrates the localization results with stereophonic device 1, where the most prominent values in the intensity graph are in the correct locations. Since the resolution of the localization is 6°, which is the same difference the sources are located in the stimulus, the tuba and singing are not as distinguishable as they were in reproduction via traditional stereo setup (Section 5.3.1).

The reader is advised to notice that the probabilities with stereophonic device 1 (Figure 6.4) have been calculated with a slightly different approach than with the other devices. In this case the probability of sound source presence at a given azimuth angle has been calculated only from the probabilities of sound source presence (as a function of azimuth angle) at different critical bands according to Equation (6.1), whereas with the other devices the intensity is formed according to Equation (5.2) where also the temporal information of source probabilities is taken into account. This different approach is selected, because the sound reproduction with stereophonic device 1 lacks energy in the high-frequencies (Section 6.3). Therefore, based on the loudness criterion the violin is localized less often as the others and the intensity of source locations at that point is significantly smaller when the intensities are calculated in the same manner as with the other devices. This can be seen in the probability graph that is calculated according to Equation (5.2). This graph is listed in Appendix C.

$$P(\theta) = \sum_{z} p[\theta, z] \tag{6.1}$$

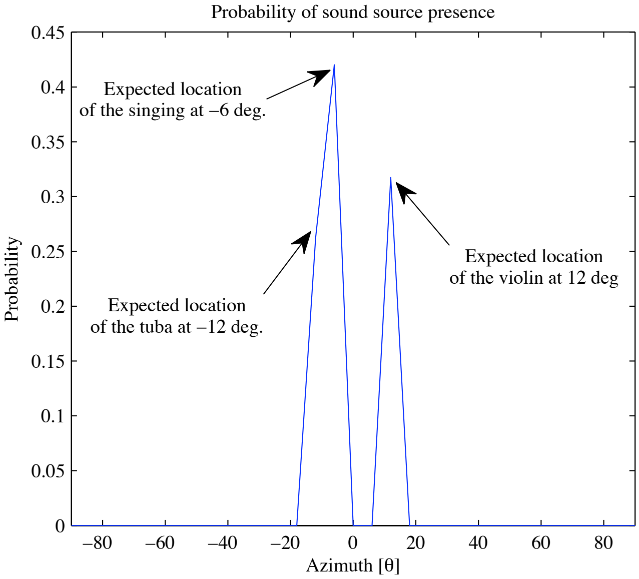


Figure 6.4: Probability of sound source presence at different azimuth angles for stereophonic device 1.

In stereophonic device 2 the loudspeaker base width is approximately 5.1 cm. Therefore the expected locations for the sound sources with this device are -6° azimuth for the tuba, -2° azimuth for the singing and 6° azimuth for the violin. The used resolution in the azimuth plane is however 6°. Hence the singing is likely to be localized either to -6° azimuth (with the tuba) or to 0° azimuth. Figure 6.5 illustrates the localization intensities with the stereophonic device where the localized sources and the expected locations of the sources are marked.

Figure 6.5 illustrates that the model localizes two sources from the stimulus, one at -6° azimuth (tuba and singing) and one at 12° azimuth (violin). There is therefore a 6° difference between the expected and the localized direction of the violin, which reflects the localization resolution and possibly a small error in the phone positioning. The source at -6° azimuth contains both the singing and the tuba. Although there are two sources there and in the results from the normal loudspeaker reproduction (Figure 5.11) these sources were the most often localized sources, the intensity of localizations at -6° azimuth is considerably smaller with this device than at the 12° azimuth. This difference in the intensities is due to the fact that the high-frequencies (where the violin is most prominent source) are emphasized in this device (Section 6.3). Therefore the violin is localized more often than the other two sources.

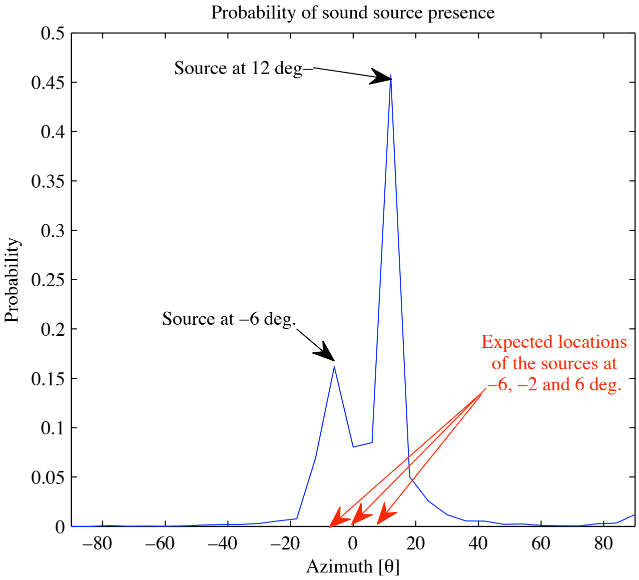


Figure 6.5: Probability of sound source presence at different azimuth angles for stereophonic device 2 without stereo enhancement algorithm.

Headphone listening of the binaural recording of the stimulus that is reproduced with stereo device 2, with the stereo enhancement algorithm set on, reveals that the stereo enhancement algorithm pans the three sources to a wider spatial area around the HATS. However at this stage the probability analysis from the model localizations is unable to reflect this functionality, since Figure 6.6 illustrates that based on the localization probabilities there would be only a monophonic source locating at 18° azimuth.

The analysis of the topographic presentation of source location intensities in different critical bands in Figure 6.7 on the other hand illustrates that the model actually localizes sources at other azimuth locations and that the sources are located in more wide area in the azimuth plane. This indicates that the stereo enhancement algorithm works as it is supposed to. The reason why these localizations do not show well in the overall intensity graph is that the high-frequency area is emphasized in the reproduction with this device (Section 6.3) and in the other frequency bands there is no prominent source location that would show also in Figure 6.6. There is also a problem in the interpretation of the localization accuracy in this case. Since at this point we do not have the knowledge where the sources are panned with the stereo enhancement algorithm, we cannot interpret the localization results with confidence at this point.

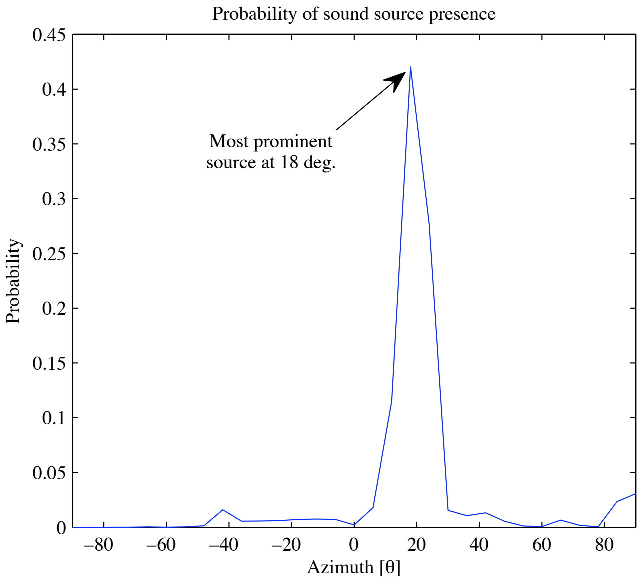


Figure 6.6: Probability of sound source presence at different azimuth angles for stereophonic device 2 with stereo enhancement algorithm.

Discussion on spatial image comparison

The test process and the offsets in the localization results (Figure 6.3 and 6.5) illustrated that there is a need for more careful phone positioning approach before the differences between different devices in the spatial localization of sound sources can be evaluated accurately. Although the azimuth resolution of the model has the most effect on the results, the small devices must also be positioned accurately in order to compare the differences. As even a slight change of phone position at this distance has an effect on the results, it is essential that the phone stays stationary and can be placed accurately to the desired position in the measurements. The phone holder that was used in these measurements was not rigid. Very accurate positioning of the device was not possible with it and it allowed the device to move a bit during the measurements.

The devices have different capabilities to reproduce sounds in different frequencies.

Therefore the loudness criterion needs to be tuned to meet the capabilities of the device if the model is to be used to evaluate the number of sound sources in the stimulus. These capability differences create also some difficulties for the analysis of spatial width from the results as the peaks in the intensity graphs are not prominent in all devices. It is therefore hard to determine from the intensity graph, which peaks can be counted as sound sources and which cannot.

There is also a need to create an alternative binaural cue reference table for the measurements with small devices. The current reference table is generated with a database of (simulated) far-field HRTFs and is therefore not suitable for near-field binaural recordings, as the HRTFs are different in the two cases. Hence the model accuracy in spatial localization could be improved by using a lookup table derived from near-field HRTFs.

The difficulties in the localization analysis with stereophonic device 2 with stereo enhancement showed that there is also a need to get more detailed information about the stereo enhancement algorithm. This information is needed before the spatial aspects from the stereo enhanced sound reproduction can be evaluated with confidence. This kind of information is only available through listening tests. At this point the expected locations of sound sources are unclear and therefore it is problematic to evaluate whether the sound sources actually are in the locations that the results from the model indicate.

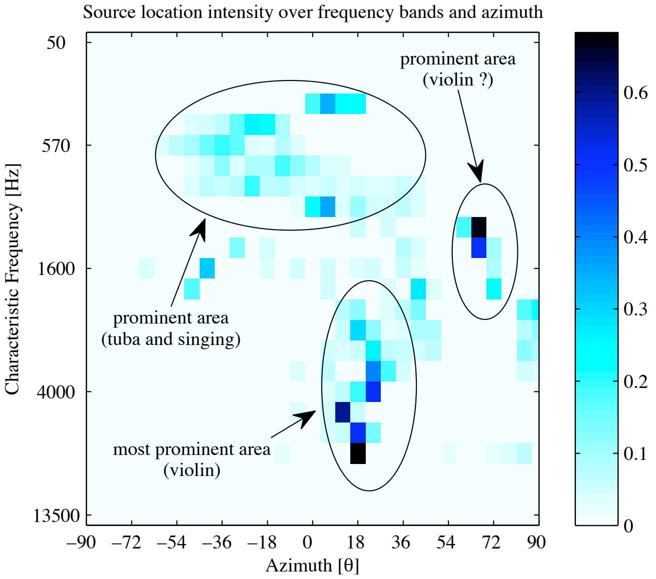


Figure 6.7: Source location intensities over critical bands and azimuth angles with stereophonic device 2 with stereo enhancement algorithm.

6.3 Evaluation of differences in terms of loudness and timbre

In this part of the study the selected devices were the monophonic device, stereophonic device 1 and stereophonic device 2 without the stereo enhancement algorithm. The effect of the stereo enhancement algorithm on the sound timbre was therefore excluded from this study as it is not supposed to change the sound in terms of loudness and timbre. The selected stimulus in this study was the same 10 second sample of Steely Dan's track '*Cousin Dupree*' that was used as stimulus #8 in the previous chapter to evaluate the functionality of the model. The reason why a music sample was used as the stimulus in this study is that it demonstrates a normal use-case with the devices in music reproduction.

The differences between the devices are evaluated both in terms of loudness and timbral aspects in this part of the study. The loudness differences are evaluated based on the specific loudness values and the overall loudness that is formed by summing the specific loudness values together. Figure 6.8 illustrates the specific loudness values from the different devices and the calculated overall loudnesses are also presented in Figure 6.8. The specific loudness values in Figure 6.8 illustrate that the values from the monophonic device are in all frequency bands greater than or almost as great as the values from the two stereophonic devices. Stereophonic device 1 has greater specific loudness values than stereophonic device 2 in low frequencies whereas in high frequencies the relationship is opposite. Therefore the monophonic device is the loudest one and the stereophonic devices are almost equally loud. This loudness relation is also illustrated in Figure 6.8.

The reader is however advised to notice that in this study the output levels of the devices were not set to maximum in all the devices. Stereophonic device 1 was set to maximum output level whereas monophonic device was set to 60 % and stereophonic device 2 to 70 % of the maximum output level of the given device. These adjustments were made by setting the outputs of the devices to a comfortable listening level and to clean sound by listening to the outputs. The reason for this procedure was to facilitate the evaluation of differences in terms of timbral characteristics between the devices, as the timbral characteristics of the sound change in some devices due to the changes in output level (Lorho 2007). Hence the differences in loudness in this study do not present the absolute level differences between the devices. This part of the study however illustrates how the model can be used to evaluate even small differences in loudness between the different stimuli.

In the evaluation of differences in terms of timbral characteristics the reproduction of the stimulus via Genelec 8020A (Genelec 2005) loudspeakers was used as a reference. The specific loudness values from the different devices were compared to the ones from the stereophonic loudspeaker reproduction that was presented earlier in Chapter 5. The specific loudness values were also scaled to the same level in order to discard the differences in overall loudness between different devices. The scaling was made at the 9th critical band, which has a characteristic frequency of 1 kHz. Here the difference in specific loudness value between the given device and the loudspeaker reference is calculated and added to the specific loudness values from the given device. These adjustments were made to facilitate the comparison of the devices in terms of timbral characteristics, since after this adjustment it is easier to see, which frequency areas are emphasized in a given device, and which are attenuated in comparison to the reference signal.

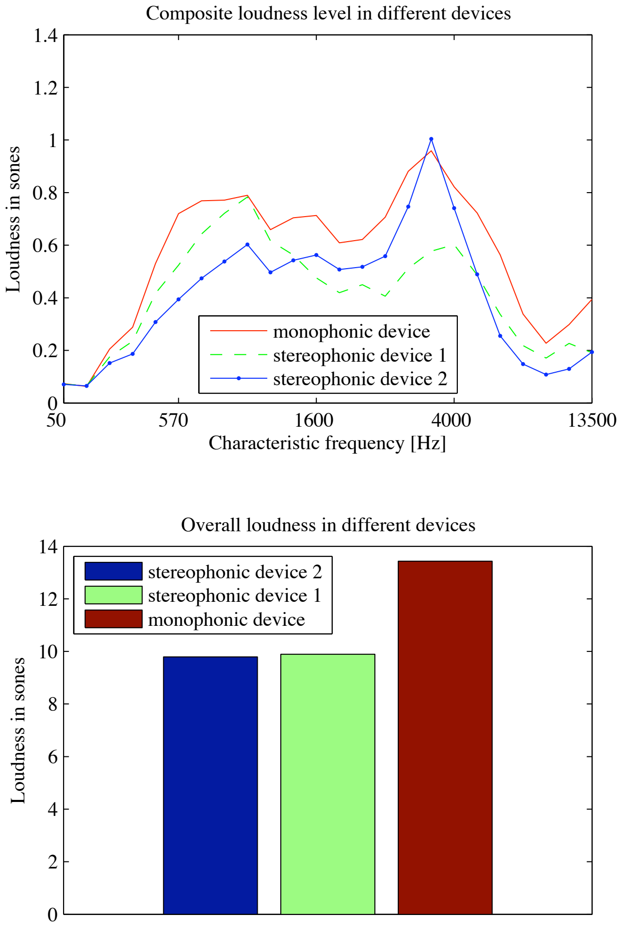


Figure 6.8: Composite loudness level values in different devices (above) and overall loudnesses in different devices (below).

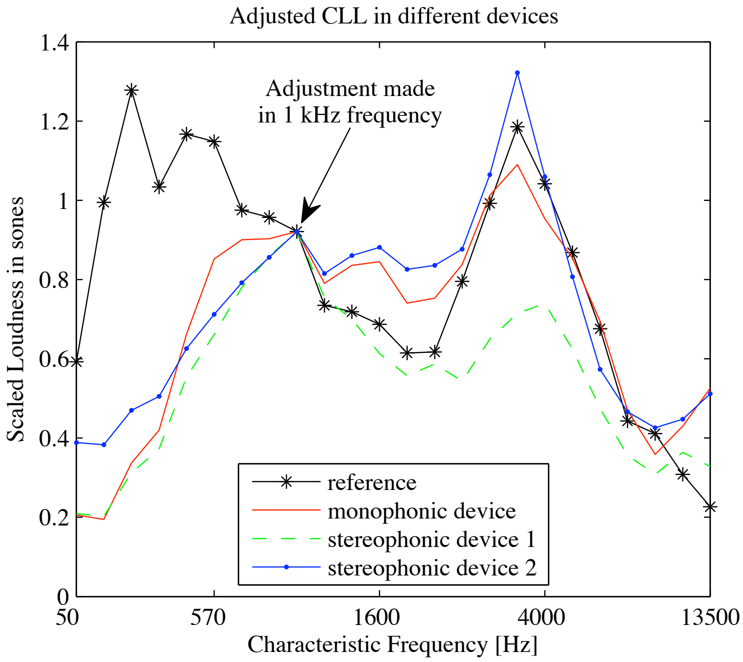


Figure 6.9: Adjusted composite loudness level values in different devices.

Figure 6.9 illustrates the composite loudness level values from the reference (loudspeaker) reproduction and the adjusted composite loudness level values from the different devices. The figure illustrates how the different mobile devices are unable to reproduce as much sound power in the low-frequencies (below 570 Hz) as the normal loudspeakers. This is understandable as the loudspeaker modules in the mobile devices are considerably smaller than the ones in a ’normal’ loudspeaker, which makes it impossible to reproduce as much bass in the sound with a mobile device. Figure 6.9 also illustrates that the frequency range between 1 and 2.5 kHz is emphasized in the reproduction with the monophonic device and with the stereophonic device 2 in comparison to the reference.

Based on the graphs in Figure 6.9 the differences in terms of timbral characteristics between the devices can be evaluated. The figure illustrates that the monophonic device produces the most balanced sound as its specific loudness values follow most closely to the ones in the reference. The monophonic device produces also most bass in the sound as the specific loudness values are highest in the frequency range between 450 and 1000 Hz. Stereophonic device 2 on the other hand produces the most treble in the sound as its specific loudness values are even higher than the ones in the reference in the frequency range between 1 and 4 kHz. The reproduction with stereophonic device 1 lacks the treble as the specific loudness values above 2000 Hz are considerably smaller than in the reference.

Therefore stereophonic device 1 produces also a "boomy" sound where the bass has been emphasized although it does not actually emphasize the low frequencies.

6.4 Summary and conclusions

In this chapter the developed binaural auditory model was used in a case study with mobile loudspeaker systems in order to evaluate what kind of differences one can detect between the different devices by inspecting the results of the BAM for the reproduced sounds. In this study the mobile devices were selected to present the variety of different devices that are currently available on the market. In this case study, the differences were evaluated in terms of spatial, loudness and timbral aspects.

The spatial difference evaluation results illustrated that the BAM is capable to detect differences in source locations also between these small devices. This evaluation demonstrated the importance of recreation of the binaural cue lookup table with higher resolution (in azimuth) and near-field HRTF database, as the loudspeaker base width in the devices is small and the devices must be placed closer to the HATS in order for the differences to be detectable. At the same time this study also showed the importance of accurate positioning of the devices, as even a slight misplacement of the device is likely to cause errors in the localization results. The evaluation of spatial differences also presented how the timbral aspects (high- and low-frequency emphasis, balance) of the reproduced sound have an effect on what sound sources are detectable in the sound. Therefore these timbral characteristics need to be taken into consideration when the criterion values for the binaural cues are set. For instance a lower loudness criterion could be used in the low frequencies with the mobile loudspeaker systems in order to detect sound sources also in that frequency range, as the mobile loudspeaker systems cannot reproduce as much sound power in the low frequencies as traditional loudspeakers.

The comparison of the specific loudness results of the reproduced sounds illustrated how the BAM can be used to evaluate differences in terms of loudness and timbral aspects. The evaluation of timbral differences was made by comparing the level aligned specific loudness values from the different devices to the reference values from the traditional stereophonic reproduction with traditional loudspeakers. By this comparison the differences in terms of timbral aspects (low- and high-frequency emphasis, and balance) between the selected devices were evaluated. This study on timbral differences however demonstrated that the selected frequency band where the level alignment is made has an effect on these results, as by adjusting the levels at a different frequency band it was possible to obtain different differences between the devices in terms of low- and high-frequency emphasis. Therefore in the evaluation of timbral differences, the effect of the level alignment position must be taken into account.

At this point the loudness evaluation of the BAM is not calibrated. Therefore in the loudness evaluation the absolute level differences between the devices cannot be evaluated accurately and only the relative differences can currently be measured. One should also notice that although the results from the evaluation of spatial, loudness and timbral aspects illustrated that there are differences between the devices, it is not possible to evaluate at this point, which differences are significant and which are not. In order to be able to do this, a listening test with human subjects must be organized so that one could know how big the difference (in terms of the given aspect) must be before a human listener can perceive it. There is hence still some work to be done before the model can be used to measure the differences in terms of spatial and timbral aspects of the reproduced sound.

Chapter 7

Conclusions and Future Work

The goal of this thesis work was to develop an auditory model that could be used to evaluate the reproduced sound in terms of quality characteristics relating to spatial and timbral aspects. After a brief introduction to the motivation of this thesis in Chapter 1 and to the relative aspects of the human auditory system in Chapters 2 and 3, a new binaural auditory model was presented. This model was built on some of the existing auditory models and its peripheral structure and functionality was described in Chapter 4.

The functionality of the developed model was verified with an experimental approach in Chapter 5. The results from these tests show good uniformity between the localization accuracy results and sound source locations in corresponding situations. It was also shown in these tests that the composite loudness level spectrum of the model and the specific loudness values from a standardized loudness model (ISO-532 1975) show similarities for different test stimuli, at least in qualitative perspective.

In this work, the developed BAM was also used in a case study with mobile loudspeakers. The purpose of this study was to illustrate what kinds of differences the BAM can find between the reproduced sound from different devices. The differences were evaluated both in terms of spatial and timbral aspects, and the different devices were selected to represent products that are currently available on the market. This case study demonstrated how the differences in terms of spatial aspects can be evaluated from the source localization results and how the differences in terms of loudness and timbral aspects can be evaluated from the CLL spectra.

The results from the functionality verification in Chapter 5 and from the case study in Chapter 6 illustrated that the localization capability of the model is dependent on the resolution of the binaural cue lookup table. Hence if a higher accuracy is needed at later stages, the lookup table resolution could be increased easily with the help of a simulated HRTF (head-related transfer function) database. It should however be noted that the verification of the model functionality is still preliminary, as the model was tested only in anechoic conditions. At this point the model works well with the tested signals, but using the model with more complex signals such as real music or recordings made in different environments is more problematic. The presented model can detect the presence of a sound source when it stands out from the others in (at least) one of the frequency bands and therefore the estimation of spatial width or number of sound sources from the probability graphs is more challenging with a real music sample where one sound source (singing) is usually most prominent and masks the others. The loudness evaluation should also take into account the temporal and frequency masking effects in order to be accurate with more complex sounds. The current results are however promising and therefore it is justified to assume that the functionality of the model could later be verified also in reverberant conditions and with more complex test signals.

The model is simplified at this point to detect only the arrival of the direct sound to the two ears, as the used time-lag (denoted as m) and time resolution (denoted as T) are too small for the reflections of the sound to arrive to the ears within them. This simplification is due to the lack of a precedence effect model (Pulkki & Karjalainen 2001). One suggestion for future approach to solve this would include using higher values for T and m and using smaller interaural coherence (IC) criterion values after the detection of the direct sound. This would possibly allow the model to detect also the early reflections of the sound and would therefore contribute to the evaluation of spatial image of the stimulus.

In the model verification and in the case study the sound source localization is evaluated in the frontal-plane with good accuracy. The binaural cue values are however closely similar at front and back at corresponding azimuth locations. Therefore inspecting the sound locations in the whole horizontal plane would cause false localizations to either front or back of the HATS (head and torso simulator). The sound direction estimation is also limited at this point to the horizontal-plane and the elevation of the sound source direction is not estimated, since according to Pulkki & Karjalainen (2001) the binaural cues are not adequate for evaluating elevation of the sound source direction. In Pulkki & Karjalainen (2001) the authors suggested that interaural cross-correlation (IACC) could be used to evaluate the elevation. Hence this elevation evaluation would be interesting to add to the BAM, as it would allow to localize sound sources in three-dimensions.

At this point the timbral aspects of the sound are 'hidden' in the composite loudness level (CLL) spectra and the evaluation of them from the reproduced sound is a natural step for the future work. The loudness estimation of the model needs however to be calibrated to an absolute level before this step can be taken as currently the model can be used to evaluate these aspects only with the help of a reference stimulus. Another suggestion for future work is adding the evaluation of other quality aspects from the binaural recorded stimulus. This addition could include for instance the evaluation of sound distortion, spatial width and evaluation of the impression of spatial image with the help of early reflections. Some suggestions for these evaluations have been presented in literature e.g. by Karjalainen (1985), Macpherson (1991) and by Thiede et al. (2000).

One should also notice that although the results from the case study illustrate differences between the devices in terms of spatial, loudness and timbral aspects, it is not possible to evaluate at this point which differences are significant and which are not. Listening tests must therefore be organized with human test subjects in order to know which differences are perceivable and which are not. The developed model could then be used to evaluate the reproduced sound in terms of defined metrics. Hence the presented model could be used as a building block for evaluating the perceptual aspects of sound.

Bibliography

Atal, B. & Schroeder, M. (1966), ‘Apparent sound source translator’, US Patent no. 3,236,949.

Avendano, C. & Jot, J.-M. (2004), ‘A frequency-domain approach to multichannel upmix’, *J. Audio Eng. Soc.* **52**(7), 740–749.

Bennet, J., Barker, K. & Edeko, F. (1985), ‘A new approach to the assesment of stereophonic system performance’, *J. Audio Eng. Soc.* **35**(5), 314–321.

Beranek, L. (1998), *Acoustical measurements*, revised edn, Acoustical Society of America.

Bernstein, L., van de Par, S. & Trahiotis, C. (1999), ‘The normalized interaural correlation accounting for nosπ thresholds with gaussian and "low noise" masking noise’, *J. Audio Eng. Soc. Am.* **106**(2), 870–876.

B&K (2006), ‘BP 0521-18, Product Data - Head and Torso Simulator - Handset positioner for HATS’.

Blauert, J. (1997), *Spatial Hearing*, revised edn, MIT Press.

Dau, T., Püschel, D. & Kohlrausch, A. (1996), ‘A quantitative model of the “effective” signal processing in the auditory system. i. model structure’, *J. Audio Eng. Soc. Am.* **99**(6), 3615–2622.

Faller, C. & Merimaa, J. (2004), ‘Source localization in complex listening situations: Selection of binaural cues based on interaural coherence’, *J. Audio Eng. Soc. Am.* **116**(5), 2075–3089.

Genelec (2005), ‘8020A Data sheet, Genelec Document BBA0034001’.

Glasberg, B. & Moore, B. (1990), ‘Derivation of auditory filter shapes from notched-noise data’, *Hearing Research* **47**, 103–138.

Goldstein, E. (2002), *Sensation and Perception*, sixth edn, Wadsworth-Thomson Learning.

Hammersøi, D. & Møller, H. (1996), ‘Sound transmission to and within the human ear canal’, *J. Audio Eng. Soc. Am.* **100**(1), 408–427.

Härmä, A. (1998), ‘Temporal masking effects: single incidents’, Internal document of Helsinki University of Technology.

Härmä, A. & Palomäki, K. (1999), Hutear - a free matlab toolbox for modeling of human hearing, *in* ‘Matlab DSP conference’, Espoo, Finland.

Houtgast, T. (1971), ‘Psychophysical evidence for lateral inhibition in hearing’, *J. Audio Eng. Soc. Am.* **51**, 1885–1894.

Irino, T. & Patterson, R. (1997), ‘A time-domain, level-dependent auditory filter: The gammachirp’, *J. Audio Eng. Soc. Am.* **101**, 412–419.

ISO-532 (1975), ‘Acoustics - method for calculating loudness level’.

ITU-T R. S. P.58 (1996), ‘Head and torso simulator for telephonometry’.

Jeffress, L. (1948), ‘A place theory of sound localization’, *J. Comp. Physiol. Psychol.* **41**, 35–39.

Karjalainen, M. (1985), A new auditory model for the evaluation of quality measurements and spatial hearing studies, *in* ‘IEEE on Acoust., Speech and Sig. Proc.’, Tampa, USA, pp. 608–611.

Karjalainen, M. (1996), A binaural auditory model for sound quality measurements and spatial hearing studies, *in* ‘IEEE on Acoust., Speech and Sig. Proc.’, Vol. 2, pp. 985–988.

Karjalainen, M. (1999), *Kommunikaatioakustiikka*, Helsinki University of Technology.

Kirkeby, O., Seppälä, E., Kärkkäinen, A., Kärkkäinen, L. & Huttunen, T. (2007), ‘Some effects of the torso on head-related transfer functions’, *In proceedings of the 122nd Int. Convention of the Audio Eng. Soc. (Vienna, Austria).*

Lord Rayleigh (1907), ‘On our perception of sound direction’, *Phl. Mag.* **13**, 214–232.

Lorho, G. (1998), Virtual source imaging system using headphones, Master’s thesis, University of Southampton.

Lorho, G. (2006), The effect of loudspeaker frequency bandwidth limitation and stereo base width on perceived quality, *in* ‘120th Int. Convention of the Audio Eng. Soc.’, Paris, France.

Lorho, G. (2007), Perceptual evaluation of mobile multimedia loudspeakers, *in* '122nd Int. Convention of the Audio Eng. Soc.,', Vienna, Austria.

Macpherson, E. (1991), 'A computer model of binaural localization for stereo imaging measurement', *J. Audio Eng. Soc.* **39**(9), 604–622.

Makous, J. & Middlebrooks, J. (1990), 'Two-dimensional sound localization by human listeners', *J. Audio Eng. Soc. Am.* **87**(5), 2188–2200.

Merimaa, J. (2006), Analysis, synthesis, and perception of spatial sound - Binaural localization modeling and multichannel loudspeaker reproduction, PhD thesis, Helsinki University of Technology.

Moore, B. (1997), *An Introduction to the Psychology of Hearing*, fourth edn, Academic Press.

Moore, B. & Tan, C.-T. (2003), 'Perceived naturalness of spectrally distorted speech and music', *J. Audio Eng. Soc. Am.* **114**(1), 408–419.

Moore, B., Glasberg, B. & Baer, T. (1998), 'A model for the prediction of thresholds, loudness and partial loudness', *J. Audio Eng. Soc.* **45**(4), 224–237.

Olive, S. (2001), Evaluation of five commercial stereo enhancement 3d audio software plug-ins, *in* '110th Int. Convention of the Audio Eng. Soc.', Amsterdam, Netherlands.

Paavola, M., Karlsson, E. & Page, J. (2005), 3d audio for mobile devices via java, *in* '118th Int. Convention of the Audio Eng. Soc.', Barcelona, Spain.

Patterson, R., Ninno-Smith, I., Wever, D. & Milroy, R. (1982), 'The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold', *J. Audio Eng. Soc. Am.* **72**, 1788–1803.

Plack, C. & Oxenham, A. (1998), 'Basilar membrane nonlinearity and the growth of forward masking', *J. Audio Eng. Soc. Am.* **103**(3), 1598–1608.

Pulkki, V. (2001), 'Localization of amplitude-panned virtual sources ii: Two- and three-dimensional panning*', *J. Audio Eng. Soc.* **49**(9), 754.

Pulkki, V. & Karjalainen, M. (2001), 'Localization of amplitude-panned virtual sources i: Stereophonic panning', *J. Audio Eng. Soc.* **49**(9), 739–743.

Pulkki, V., Karjalainen, M. & Huopaniemi, V. (1999), 'Analysing virtual sound source attributes using a binaural auditory model', *J. Audio Eng. Soc.* **47**(4), 203–207.

Riederer, K. A. (2005), HRTF Analysis: Objective and Subjective Evaluation of Measured Head-Related Transfer Functions, PhD thesis, Helsinki University of Technology.

Rix, A., Beerends, J., Kim, D.-S., Kroon, P. & Ghitza, O. (2006), 'Objective assessment of speech and audio quality - technology and applications', *IEEE Transactions on audio, speech and signal processing* **14**(6), 1890–1901.

Rossing, T., Moore, F. & Wheeler, P. (2002), *The Science of Sound*, third edn, Pearson Education.

Rumsey, F. (2001), *Spatial Audio*, Focal Press.

Schroder, M., Atal, B. & Hall, J. (1979), 'Optimizing digital speech coders by exploiting masking properties of the human ear', *J. Audio Eng. Soc. Am.* **66**, 1647–1652.

Shinn-Cunningham, B. (2000), Distance cues for virtual auditory space, *in* 'The First IEEE Pacific-Rim Conference on Multimedia', Sydney, Australia, pp. 227–230.

Shinn-Cunningham, B., Santarelli, S. & Kopco, N. (2000), 'Tori of confusion: Binaural localization cues for sources within reach of a listener', *J. Audio Eng. Soc. Am.* **107**(3), 1627–1636.

Sivonen, V. & Ellermeier, W. (2006), 'Directional loudness in an anechoic sound field, head-related transfer functions and binaural summation', *J. Audio Eng. Soc. Am.* **119**(5), 2965–2979.

Slaney, M. (1988), Lyon's cochlear model, Technical Report 13, Apple Computer.

Slaney, M. (1998), Auditory toolbox: Version 2, Technical Report 1998-010, Apple Computer.

Staffeldt, H. (1974), 'Correlation between subjective and objective data for quality of loudspeakers', *J. Audio Eng. Soc.* **22**(6), 402–415.

Syntrillium Software Corporation (2002), 'Cool edit pro version 2.00'.

The Mathworks (2007), 'Matlab, version 7.4.0.287 (r2007a)'.

Thiede, T., Treurniet, W., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. & Colombes, C. (2000), 'PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality', *J. Audio Eng. Soc.*

Toole, F. (1986), 'Loudspeaker measurements and their relationship to listener preferences: part 1', *J. Audio Eng. Soc.* **34**(4), 227–235.

Tuomi, O. & Zacharov, N. (2000), A real-time loudness meter, *in* '139th ASA meeting', Atlanta, USA.

Unoki, M., Irino, T., Glasberg, B., Moore, B. & Patterson, R. (2006), 'Comparison of the roex and gammachirp filters as representations of the auditory filter', *J. Audio Eng. Soc. Am.* **120**(3), 1474–1492.

Zielinski, S., Rumsey, F. & Bech, S. (2002), Subjective audio quality trade-offs in consumer multi-channel audio visual delivery systems. part ii: Effects of low frequency limitation, *in* '22nd Int. Conference of the Audio Eng. Soc.', Espoo, Finland.

Zwicker, E. & Fastl, H. (1999), *Psychoacoustics, Facts and models*, second updated edn, Springer.

Appendix A

Matlab code for "step-sine" sound

```
% parameters
fs = 48000;
cf = [50;150;250;350;450;570;700;840;1000;1170;1370;1600;1850;
    2150;2500;2900;3400;4000;4800;5800;7000;8500;10500;13500];
% length of one sinusoid is 200 ms
t = 0.2/(fs*0.2):(0.2/(fs*0.2)):0.2;

% signal creation for each critical band
sig = zeros(1,0.1*fs);
for i=1:length(cf)
        sig = [sig sin(2*pi*cf(i).*t) zeros(1,0.1*fs)];
end

% adjustment to avoid clipping
sig = sig./(1.1*max(max(sig)));
```

Appendix B

Model test figures

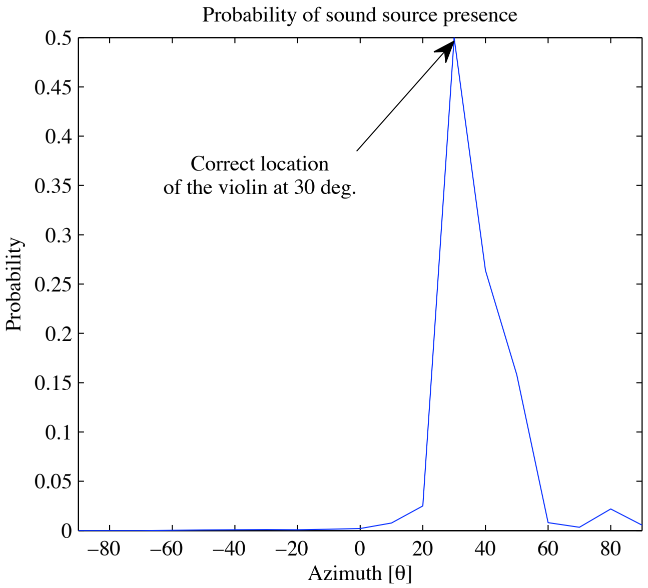


Figure B.1: Probability of sound source presence with stimulus #2.

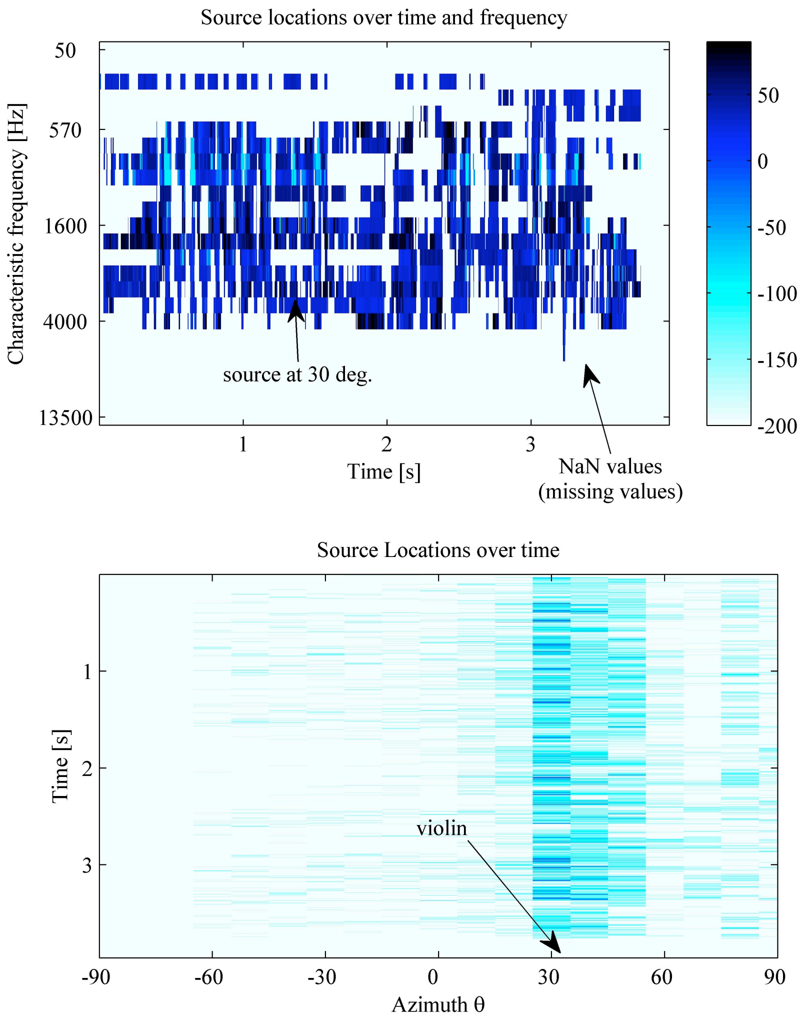


Figure B.2: Topographic presentation of sound source locations (upper graph) and probabilities of source locations as function of time and azimuth (lower graph) for test signal #2.

Appendix C

Case study figure

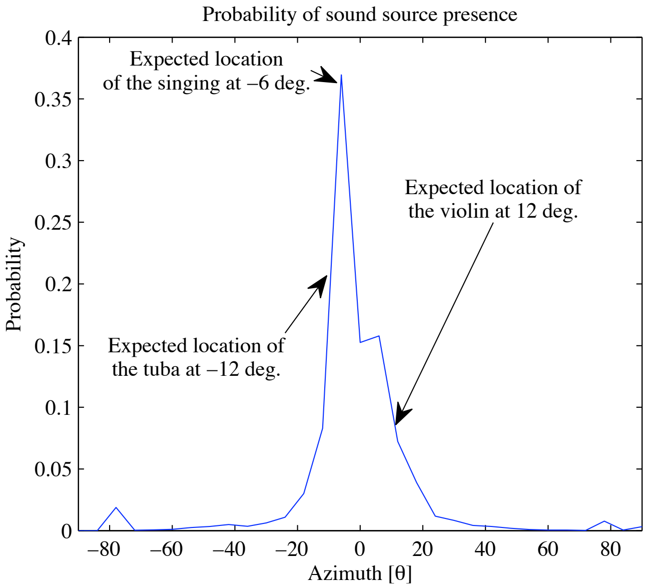


Figure C.1: Probability of sound source presence with stereophonic device 1. Probabilities calculated with Equation (5.2).