HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of electronics, communications and automation
Department of signal processing and acoustics

Timo Hiekkanen

Virtualized Loudspeaker Testing

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

Espoo, Feb 24, 2008

Supervisor:	Professor Matti Karjalainen

HELSINKI UNIVERSITY
OF TECHNOLOGY — ABSTRACT OF THE
MASTER'S THESIS

Author:	Timo Hiekkanen		
A Title:	Virtualized Loudspeaker Testing		
Date:	Feb 24, 2008	**Number of pages:**	71
Faculty:	Faculty of electronics, communications and automation		
Professorship:	S-89		
Supervisor:	Prof. Matti Karjalainen		

Loudspeakers cannot be accurately compared to each other in room conditions since the placement of loudspeakers has a significant effect on the results. Also, the comparison of multiple loudspeakers is limited by the poor auditory memory of humans since the repositioning of the loudspeakers cannot be done instantly.

In this thesis, the fundamentals of spatial hearing and sound reproduction are reviewed and a binaural method for loudspeaker comparisons is developed. The proposed method utilizes individual binaural responses together with an artificial head and headphones to enable fast, seamless and place-independent switching between different loudspeakers. The method overcomes the worst drawbacks of traditional loudspeaker listening tests but new problems related to the measurement accuracy and reproduction arise.

Formal listening tests were conducted to examine the differences between the binaural method and the real loudspeakers in a standard listening room. It was found out that the binaural method is close to imperceptible compared to reality when a speech signal is listened to. However, the difference to the real loudspeakers increases when audio material with more energy at high frequencies is used. The listening tests revealed that individually equalized responses of an artificial head are applicable to binaural synthesis almost as well as individual true-head responses.

Keywords: binaural techniques, virtual reality, loudspeakers, listening tests, spatial sound

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Timo Hiekkanen		
Työn Nimi:	Paikkariippumaton menetelmä kaiuttimien vertailuun		
Päivämäärä:	24.2.2008	**Sivuja:**	71
Tiedekunta:	Elektroniikan, tietoliikenteen ja automaation tiedekunta		
Professuuri:	S-89		
Työn valvoja:	Prof. Matti Karjalainen		

Kaiuttimien vertaileminen huoneolosuhteissa on vaikeaa, koska pienetkin erot kaiuttimien sijoittelussa vaikuttavat lopputulokseen. Ihmisen kuulomuistin lyhyys rajoittaa vertailtavien kaiuttimien määrää, jos kaiuttimet sijoitetaan vuorotellen samaan pisteeseen.

Työn kirjallisuusosiossa esitellään tilakuulon ja äänentoiston perusteita. Myöhemmässä osassa esitetään binauraalitekniikkaan perustuva menetelmä kaiutinten vertailuun. Menetelmä perustuu mitattujen tosi- ja keinopäävasteiden käyttöön yhdessä kuulokkeiden kanssa, ja sen avulla voidaan verrata useita kaiutinpareja viiveettömästi. Vaikka työssä esitetty menetelmä ohittaa perinteisiin kuuntelukokeisiin liittyvät ongelmat, sen käyttökelpoisuutta rajaavat mittausepävarmuus ja toiston tarkkuus.

Menetelmän toimivuutta tutkittiin kuuntelukokeessa. Kokeella selvitettiin, kuinka paljon binauraalinen toisto eroaa todellisesta kuuntelutilanteesta. Puhesignaalilla menetelmä oli lähes erottamaton todellisuudesta. Musiikki- ja kohinasignaaleilla koehenkilöt arvioivat eron olevan havaittava tai hieman häiritsevä. Koe osoitti, että yksilöllisesti korjatut keinopäävasteet soveltuvat binauraalisynteesiin lähes yhtä hyvin kuin tosipäävasteet.

Avainsanat: binauraalitekniikat, keinotodellisuus, kaiuttimet, kuuntelukokeet, tilaääni

Acknowledgements

This Master's thesis was carried out in the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology in co-operation with Genelec Oy.

First and foremost, I wish to thank Professor Matti Karjalainen for supervising this work. His excellent knowledge on acoustics and comments on grammatical issues have been helpful. I would also like to thank Aki Mäkivirta and Genelec Oy for their support.

I want to thank Mikko-Ville Laitinen, Juha Vilkamo and the whole spatial sound group for inspiring discussions related to acoustics, signal processing, computer games and life.

Otaniemi, February 24, 2008

Timo Hiekkanen

Contents

Abbreviations

ANOVA	ANalysis Of VAriances
ASD	Auditory Spectrum Distance
FEC	Free-air Equivalent Coupling
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
HATS	Head And Torso Simulator
HRIR	Head-Related Impulse Response
HRRTF	Head and Room-Related Transfer Function
HRTF	Head-Related Transfer Function
IEC	International Electrotechnical Comission
IHL	Inside-the-Head Localization
IID	Interaural Intensity Difference
ILD	Interaural Level Difference
IMD	InterModulation Distortion
IRS	Inverse Repeated Sequence
ITD	Interaural Time Difference
ITU	International Telecommunications Union
JND	Just Noticeable Difference
LTI	Linear and Time Invariant
MLS	Maximum Length Sequence
PDR	Pressure Division Ratio
PTF	headPhone Transfer Function
RT_{60}	Reverberation Time
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
THD	Total Harmonic Distortion

Chapter 1

Introduction

Traditionally, subjective evaluation of loudspeakers is done in room acoustics, usually in more or less standardized listening rooms. Listening tests are conducted to find out, does the loudspeaker meet the design goal or is some loudspeaker inferior to another. In the first stage, properties of the loudspeakers are compared by the designer and final evaluation is made by the consumer when making a buying decision. However, there are several aspects that prevent reliable direct comparisons between loudspeakers.

The human auditory memory is too short. Humans can not reliably remember complex sound images for longer than few seconds. Our long-term auditory memory does not give comparable sound images and our mood-of-the-day can affect severely the preference ratings if we try to compare current loudspeaker to some older piece of equipment that is not at hand. Loudspeakers cannot be replaced fast enough to make direct comparisons at the same location and on the other hand, if the loudspeakers are not located on exactly the same spot, results can be biased according to the loudspeaker positions.

Unfortunately, what we see is often what we hear. Visual cues can seriously affect the results of listening tests. This phenomenon is called ventriloquism [1]. For instance, if we can see the loudspeakers we are listening to, fancy-looking big loudspeakers will get better results than everyday modest loudspeakers even if the sound quality would suggest the opposite.

To achieve comparable results, all loudspeakers should be evaluated in the same room, at the same time and location. The listener should also be in exactly the same position all the time. Loudspeakers should be invisible or look the same. Obviously, these requirements can not be fulfilled in real life by any means. That is why an alternative approach is studied in this thesis.

According to the binaural theory, an auditory experience could be repeated if the same sound pressures are reproduced at the ear drums [2]. Now, the question arises if loudspeakers could be compared and evaluated through binaural recording and processing. If the full auditory experience consists of only the signals at the two ear drums of a human listener, shouldn't it be possible to repeat these signals and use this superior repeatability to make ideal listening tests, where multiple loudspeakers in the same room and position could be switched instantaneously and unnoticed.

Another question is, how well the binaural replica of the real world represents the reality. Could this binaural reproduction of reality be used simultaneously with the real loudspeakers to compare new loudspeakers to older ones which don't exist any more? And if we go even further, do we need expensive, big and impractical loudspeakers if we could listen the best loudspeakers in the world through our own headphones?

At least in some earlier research projects, binaural techniques have been used to ease the test method and make the listening conditions equal to every subject [3][4][5]. In [6], Blauert points out the benefits of binaural technology in measurement and evaluation of audio signals. In the present thesis, binaural methods are adapted to normalize the listening conditions of different loudspeakers. The listening room is simulated using individual head and room related transfer functions with artificial head and torso simulator and headphone correction performed. The method shares some properties with the method proposed by Mickiewizc in [7] to improve headphone listening in home conditions.

An earlier study performed by Ganjian and Preis suggests that a loudspeaker-room response can not be accurately represented with a linear filter [8]. In their study, the problem was investigated with physical measurements. In this thesis the problem is approached from the perceptual side.

The ultimate goal of this work is to introduce a new method for loudspeaker listening tests and to examine the capabilities of headphone reproduction. The main question is, how accurately the loudspeaker-room responses can be reproduced through headphones and does this reproduction match with the reality in terms of localization and timbre. Smaller partial questions could be:

- Can the head and torso simulator (HATS) responses be used instead or with the true-head responses?
- How accurate the measurements have to be? What is the tolerance between measurements to achieve perceptually similar reproduction?
- What is the repeatability of the headphone responses and how the headphones should be equalized for binaural reproduction?
- Can the inside-the-head localization (IHL) be defeated?

This thesis is organized as follows:

In chapter 2 some aspects of human hearing and auditory perception are introduced. Especially properties connected with sound reproduction are emphasized.

In chapter 3 attention is paid to loudspeaker and headphone reproduction. Binaural recording and reproduction techniques are examined.

In chapter 4 measurements needed are discussed in general level and the repeatability of the measurements is investigated.

In chapter 5 the proposed test method is revealed and documented. Measurement procedure as well as the signal processing aspects are discussed.

In chapter 6 the new method is verified by listening tests. Results are shown and analyzed.

In final chapter 7 conclusions are drawn and the results are discussed.

Chapter 2

Hearing and Auditory Perception

Human hearing is still the ultimate measurement and evaluation device of sound. It has prevailed the attacks of technology for decades. When thinking about complex recordings where several instruments are mixed with excess of room acoustics and noise, computers are ruled out right away. Only humans can separate the instruments, transcript them correctly and somehow exclude the reverberant environment. In many cases, when no difference is seen from frequency responses or from other measurements, the effect under study is still apparently audible.

Although in many cases measurements provide valuable information about loudspeakers, rooms or instruments, the final judgement comes always form us. There's no help to claim that some piece of equipment is perfect for its purposes if listeners don't agree with the measurements. That is why equipment meant for sound reproduction have always to be evaluated by humans.

The properties of human hearing differ significantly from linear scale used by standard measurement devices. Perception of pitch, loudness, and time domain effects are far away from linear; on the contrary they are very complex and non-linear.

In the next sections, some properties of human hearing and sound localization are revised and linked to the subject of the thesis. First, the coordinate system used is shown and equivalences between time and frequency domains are presented. Secondly, loudness perception is studied and human sensitivity to audio distortions is reviewed. Finally, attention is paid to directional hearing and spatial perception. The physical structure of human hearing organs is not studied, since the scope of this thesis lies more on the perceptual side. Good descriptions of hearing organs can be found from [9][10][1] or from many other books covering some aspects of human hearing.

2.1 Coordinates, Time and Frequency Domain Relations

In this thesis, the head-related spherical coordinate system is used. Use of the head-related coordinate system is advantageous since coordinates are fixed relative to the position of ears [9].

Figure 2.1 shows the orientation. φ denotes the azimuthal angle, zero degrees being in the front of the head. φ increases clockwise. δ denotes for elevation angle and it increases upwards, zero degrees being also in front of the head, at the ear level. Horizontal plane is the plane where δ is zero. In the median plane, φ is zero.

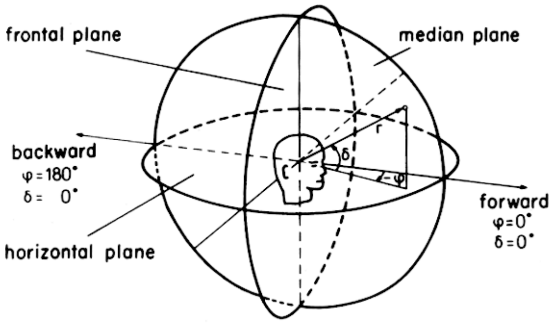


Figure 2.1: Head-related spherical coordinate system. Adopted from [9].

A time domain acoustic signal refers directly to a signal that can be measured. Here it represents the pressure fluctuations in the air. A frequency domain signal represents the frequency content of the signal in a specific time window. In general, the frequency domain signal, equivalent to the time domain signal, is achieved by the *Fourier transform* shown in Eq. (2.1).

$$X(f) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft}dt \tag{2.1}$$

$X(f)$ is the complex spectrum where the variable f represents the frequency in Hz, and $x(t)$ is the original time domain signal. The time domain signal can be achieved from the frequency domain signal by *inverse Fourier transform* shown in Eq. (2.2).

$$x(t) = \mathcal{F}^{-1}\{X(f)\} = \int_{-\infty}^{\infty} X(f)e^{i2\pi ft}df \tag{2.2}$$

In the digital domain, where signals are finite-length and discrete, the *discrete Fourier transform* is used. Here we omit further mathematical descriptions and accept that signals can be examined in the time and frequency domains. Time domain signals are denoted with lower case letters while frequency domain equivalents are denoted with capital letters. More about discrete Fourier transform and its inverse can be found from [11].

According to the theory of linear and time-invariant (LTI) systems [12], an LTI system is fully described by its impulse response, which is the system output when unit impulse is the input. The system response to any input can be calculated by the *convolution* operation denoted by $*$

and described in Eq. (2.3).

$$f(t) * g(t) = \int_a^b f(\tau)g(t-\tau)d\tau \tag{2.3}$$

An important property of the frequency domain signals is that convolution, which is computationally heavy operation, reduces to multiplication:

$$\mathcal{F}\{f(t) * g(t)\} = F(f)G(f) \tag{2.4}$$

2.2 Frequency Domain Resolution

The ear works as a frequency analyzer, coding the pressure fluctuations of air to place-specific vibration on the basillar membrane in the inner ear [13]. However, to reduce the amount of information, the hearing system has interesting physiological and neurological properties.

The frequency selectivity of human hearing has been traditionally tested with the following experiment. Narrow band noise and variable bandwidth noise with the same center frequency are played in a row. Sound pressure levels of the sounds are kept constant. The task is to adjust the signals to the same loudness perception. First, when the bandwidth of the noise is increased, the perceived volume levels (i.e. loudness) stay the same. If the bandwidth of the second sound is still increased, the perceived volume level starts to grow. The bandwidth where this happens is called *the critical bandwidth*. The width of the critical bandwidth varies with the center frequency. At low frequencies the bandwidth is nearly constant (about 90 Hz) and at higher frequencies it is roughly proportional to the center frequency (about one third octave) [13].

The Bark scale and Bark bands are achieved from the critical bands. One Bark band equals roughly to a critical bandwidth at the same frequency. The audible frequency range is divided to 24 Bark bands. The Bark scale, as in Figure 2.2, was first proposed by Zwicker in 1961 [14].

When analyzing complex tones, human hearing analyzes one critical band as one block. This affects greatly to the loudness perception and wide-band signal analysis. Although pitch detection by humans is more accurate than the critical bandwidth, the accuracy of pitch detection follows the widths of the critical bands being about $1/25$ of the critical band-width [15].

Another scale approximating the frequency resolution of humans is the ERB scale [16]. It reminds the Bark scale but it is closer to the logarithmic scale at low frequencies. The relation of the ERB scale to the resonance positions at the basillar membrane is somewhat better than with the Bark scale. Both scales are used in auditory models.

A loud tone can render other simultaneous tones completely inaudible. Traditionally it is studied with pure sine tone and bandpass or wide-band noise. The sine tone acts as probe sound to be detected and the noise signal is the masker. Uniform masking at all frequencies is achieved with wide-band noise which is constant to approximately 500 Hz after which it decreases about 3 dB per octave. The masking effect of noise is quite strong: relative noise level of 0 dB masks

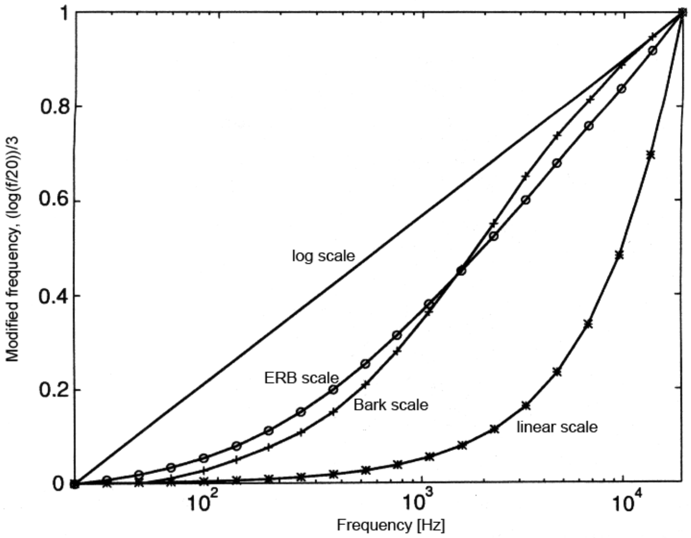


Figure 2.2: Relations between linear, logarithmic, Bark and ERB scales. Adopted and modified from [15].

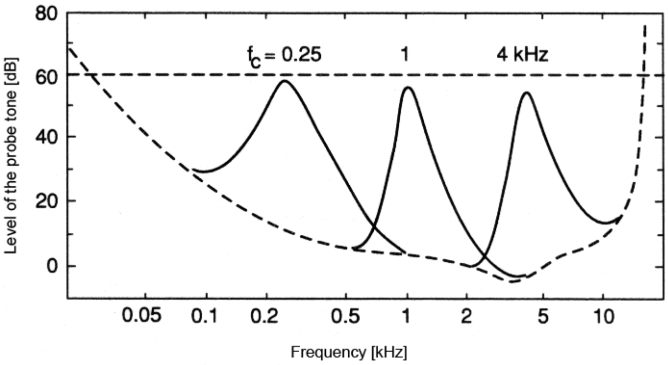


Figure 2.3: Masking effect in frequency domain caused by narrow-band noise. The center frequency of the masker is 250 Hz, 1 kHz or 4 kHz and the sound pressure level is 60 dB. Adopted and modified from [15].

all frequencies below relative level of −20 dB. Figure 2.3 illustrates the masking caused by narrow-band noise.

With narrow-band noise the masking effect is strongest at the point of center frequency and it decreases when moving to lower or higher frequencies. Louder sounds mask high frequency sounds better than sounds lower in level.

When thinking about sound reproduction or audio signal processing, the frequency domain masking gives new degrees of freedom to the designer. It is not required to get fully rid of distortions or artifacts, it is enough to reduce the level of the disturbances below the masking threshold. With loudspeakers, many kinds of nonlinearities may go undetected because of the masking.

2.3 Loudness

It is well-known that the perceived sound level doesn't always correspond to the levels indicated by sound level meter very well. The perceived sound level, *loudness*, depends on a variety of properties like actual sound level, frequency content and temporal structure of the sound. At the same time, loudness influences greatly other perceived aspects of sound. Moore defines loudness as *that attribute of auditory sensation in terms of which sound can be ordered on a scale extending from quiet to loud* [17]. In this section, the build-up of loudness and its measures are revised shortly and its effects to sound quality are discussed.

For a single steady tone, loudness is determined by equal loudness curves as in Figure 2.4 These curves are often referred to as the Fletcher-Munson curves since the curves were first obtained by H. Fletcher and W. A. Munson [18]. Based on the standardized curves, shown in Figure 2.4, the unit for loudness level has been set. Loudness level in *phons* equals sound pressure level (SPL) at 1 kHz frequency and towards low and high frequencies it decreases

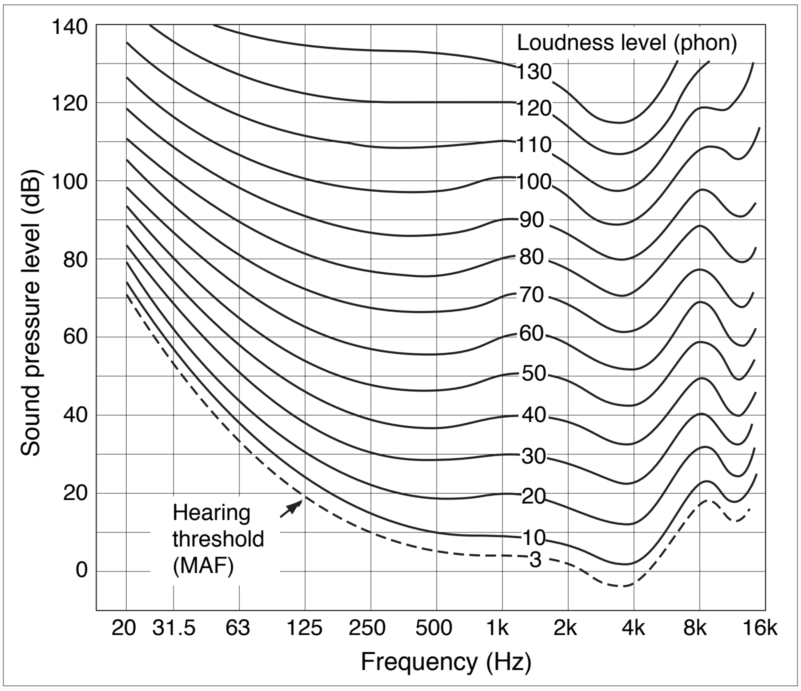


Figure 2.4: Equal loudness curves defined by standard ISO 226.

compared to SPL. Equation 2.5 describes the relation between the loudness level and loudness. Doubling the loudness, expressed in *sones*, increases the loudness level by 10 dB [15].

$$N = 2^{(L_L - 40)/10} \tag{2.5}$$

However, the loudness of complex tone sounds is a much more complex issue. Generally, wide-band sounds are perceived louder than narrow-band sounds even if the signal energies are equal. This is closely related to the concept of the critical bands presented in the previous section. Usually loudness models are somehow summing the energy on the critical or $1/3$ octave bands to get the overall loudness. One of the first methods to predict the loudness of complex signals was proposed by Zwicker in [19] and it has been standardized by ISO.

Loudness is not defined only by frequency content. Signal duration has a significant effect on the loudness sensation. With signal durations over 200 ms the perceived loudness is constant. With shorter durations the loudness decreases and with durations less than 100 ms the perceived loudness decreases approximately 10 phons per decade [15]. Also, a signal with 200 ms breaks between equal length noise bursts sounds as loud as a continuous signal.

Another problem in loudness evaluation is that the hearing system adapts itself depending on the incoming signal. *Temporary threshold shift* (TTS) happens when ear is exposed to a tone of known frequency and sound level for a period of time. If the hearing threshold is measured after this fatiguing, higher than usually results are received [17]. Moore describes in [17] an experiment where even 20 dB adaptation can be seen.

It has been studied that loudness affects greatly to the perceived sound quality. To some extent, louder sounds are considered "better sounding" and clearer than quiet ones. Gabrielsson et al. have studied the effect of frequency content and sound level of the signal to the perceived sound quality [20]. According to Gabrielsson et al., sound level, which usually highly correlates with the loudness, gives significantly better clarity, fullness, spaciousness, nearness and fidelity. On the other hand, higher levels decrease the softness of the sound.

The complexity of the loudness sensation and the significant effect of loudness to sound quality lay challenges when loudspeakers or other reproduction devices are evaluated. There is no doubt that the loudness of all signals should be normalized to get comparable results. This is not a trivial or easy task to do since the loudness sensation is always more or less a subjective matter. However, Toole and Olive state in [21] that B-weighted sound pressure level measurements may give as good results as much more complex methods when comparing loudspeakers. Even A-weighted measurements might be good enough.

2.4 Perception of Linear and Nonlinear Distortion

In ideal sound reproduction equipment, sound is reproduced exactly as it is fed to the input. This is not the case with loudspeakers. Some amount of linear (changes in magnitude, phase and group delay) and nonlinear distortion is generated always. That is why it is not possible to design distortion-free loudspeaker. A better approach is to minimize the unwanted and audible distortion while not spending time with effects that we can not perceive anyway. Some nonlinear distortion may be even wanted: sophisticated amount of second order harmonics may increase the perceived sound quality.

Nonlinearities of the hearing system make the perception of nonlinear distortion complex. New frequencies can be created in the ear itself. Green gives an example of so-called cubic difference tone in [22]. If two tones are produced to the ear, let's say $F_1 = 1000$ Hz and $F_2 = 1200$ Hz, a third tone, $2F_1 - F_2$, is evidently audible. How we can know if the distortion we hear is produced by the ear, not in the equipment we are evaluating?

The method for virtual loudspeaker evaluation proposed in Chapter 5 does not take into account nonlinear distortion since it can not be simulated with LTI filters used. In the following subsections the audibility of different distortions is discussed.

2.4.1 Linear Distortion

Linear distortion occurs in mechanical or electrical devices if some of the signal energy is absorbed or reinforced, different frequency components travel in different speeds, or some of the signal energy is stored and released later [23]. Differences can be observed from measured frequency response, but the cause of the distortion can not.

Human hearing is quite sensitive to changes in magnitude spectrum. Changes as small as ± 0.5 dB to ± 1 dB can be heard in good listening conditions within the audio bandwidth [23]. To be more accurate, the just noticeable difference (JND) is a ± 1 dB change with one Bark bandwidth. Direct comparison is required to achieve this accuracy [15]. This gives the ultimate goal for sound reproduction devices: the magnitude spectrum should fit between these tolerances. Lower JND values can be achieved by training. However, since humans poor auditory memory, as large modifications as 5 - 10 dB can be unnoticeable if direct comparison is unavailable.

Group delay, τ_{g}, is the derivative of the phase response with respect to frequency as in Eq. (2.6).

$$\tau_{\mathrm{g}} = -\frac{d\phi(\omega)}{d\omega}, \tag{2.6}$$

where $\phi(\omega)$ is the phase response of the system. Group delay is the delay that the envelope of a specific frequency experiences in the system. If the phase function $\phi(\omega)$ is constant or linear, envelopes at all frequencies have the same delay and signal comes out delayed but not distorted. Otherwise the frequency envelopes have unique delays and the signal is distorted in time domain. Group delay is useful measure when speaking about phase distortions since human ear detects changes in the time envelope of the signal and the group delay describes the relative delays of the envelopes at different frequencies [15]. More about mathematics of the issue can be found from [11][24].

In general, the human hearing system is not very sensitive to phase distortions. Jensen and Møller came to the decision that human ear is practically phase deaf [25]. In many cases complex modifications can be done to the phase response without audible difference. For instance, for vowels the phase can be inverted. On the other hand, different thresholds for group delay distortion detection have been received by Blauert and Laws, Deer et al. and Suzuki et al. [26][27][28]. Blauert and Laws found that minimum audible threshold for group delay distortion is 1 ms at 2 kHz varying from 3.2 ms to 1 ms between 500 Hz and 8 kHz. Deer et al. found the minimum audible amount to be 2 ms at 2 kHz in dichotic listening with headphones. Hearing is much more sensitive to the phase when listening with headphones [28].

Sensitivity to group delay distortions decreases with the frequency. Because at low frequencies hearing responds more to the actual waveform than to the envelope, phase delay instead of group delay is proposed as the measure of phase distortion [23]. Human sensitivity to phases in audio signals is still a controversial subject and research continues.

A 1 ms group delay difference, which is found to be just noticeable, means 0.001 s · 340 m/s = 0.34 m difference in flight time. This would suggest that the distance between the high frequency driver and the woofer of the loudspeaker is not a critical design parameter when designing small or medium sized loudspeakers.

2.4.2 Nonlinear Distortion

Nonlinear distortion occurs when the input and the output of a system are not linearly related, i.e. doubling the input amplitude does not double the output amplitude. In Figure 2.5 some nonlinear system responses can be seen. In the frequency domain, nonlinear distortion appears as new frequency components meaning that the system generates frequencies that don't exist in the input signal at all. These new frequency components can be identified as harmonic, sub-harmonic and intermodulation components [29]. Harmonic and sub-harmonic components are integer multiplies or divisions of the fundamental frequency. Intermodulation components appear when two frequencies modulate each other: sum and difference components are created.

There are two traditional measures for nonlinear distortion: total harmonic distortion (THD) and intermodulation distortion (IMD). THD is the ratio of the RMS output signal due to distortion to the total RMS output signal and it can be measured in a number of ways [21]. IM distortion is divided to amplitude modulation and frequency modulation. Amplitude modulation (AM) can be seen (or heard) when a low frequency tone modulates a second tone higher in frequency: the amplitude of the higher frequency changes periodically. In frequency modulation (FM) one tone modulates the frequency of the other tone. In both cases sum and difference products appear around the modulating frequency.

Perception of distortion in terms of THD and IMD depends the situation. Both Colloms and Geddes state that while levels of below 0.1% for harmonic and intermodulation distortion may be audible in amplifier stage, levels of 1% may be completely inaudible in loudspeakers [30][31]. According to Moir, at 400 Hz the audibility levels of 2nd and 3rd harmonic distortions are both about 1% when testing with pure tones. At 100 Hz, levels below 5% are inaudible for 3rd order distortion and for 2nd order distortion levels greater than 20% may be below the threshold of per-

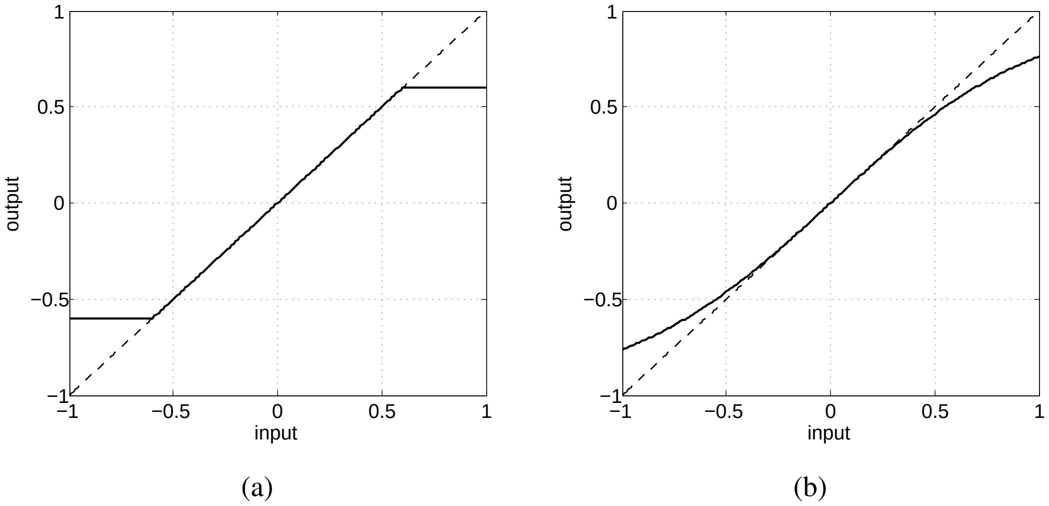


Figure 2.5: Nonlinear system responses. (a) is hard clipping, (b) is soft clipping, $y = tanh(x)$.

ception [32]. Fryer examined the human sensitivity to the first order intermodulation distortion in [33] and came to decision that the level of perception threshold lies between 4 and 5%.

In [34] Karjalainen defines auditory spectrum distance (ASD) and uses it to measure the perceptivity of nonlinear distortion. ASD is the maximum value of spectral deviation over time and Bark scales. Karjalainen found that nonlinear distortion is just perceivable when ASD is about 1.5 - 2.5 dB and undistorted reference is available.

Geddes states in [31] that THD and IMD are not useful measures of nonlinear distortion in means of perception. Justification to this is that human hearing masks differently different orders of harmonic and intermodulation distortions. At high sound pressure levels masking is stronger and high distortion levels can be unperceptible. At low sound pressure levels smaller amounts of distortion may be perceptible. Distortion products below the masking tone frequency are also more audible than products above. Another reasoning is that equipment act differently at different levels. While for amplifiers crossover distortion increases at low levels, loudspeaker distortion increases only when volume is turned up. Then the masking effect is stronger and high distortion levels may still go unnoticed.

Klippel separates eight sources of nonlinearities in loudspeakers, mainly caused by the geometry and physical limitations of the loudspeakers [29]. These are only mentioned here to remind that these properties are not modeled by the evaluation method proposed in Chapter 5.

Stiffness of suspension. Woofer suspension is nonlinear at high displacements. It causes harmonic distortion and amplitude intermodulation.

Force factor. Force factor describes the coupling between the mechanical and electrical sides of the transducer. Force factor depends on the displacement of the voice coil causing nonlinearities especially at high sound levels.

Voice-coil inductance. The input impedance depends on the position of the coil and is frequency dependent.

Nonlinear material properties. Vibrations in cone and other parts become nonlinear if the strain and stress are high.

Variation of geometry. Vibrations become nonlinear if the displacement is not small compared to the geometrical dimensions.

Port nonlinearity. Ports in vented systems have a flow resistance that is not constant, but depends highly on the velocity of the air inside the port.

Doppler effect. High frequencies radiating from woofer are frequency-modulated by lower frequencies.

Wave steepening. At high amplitudes a sound wave propagates at the maxima faster than at the minima, causing a gradual steepening of the wavefront.

2.5 Localization

According to Moore, the term *localization* refers to judgments of the direction and distance of a sound source [17]. Term localization is used when the sound source is located to be somewhere around us. If the sound source seems to be inside the head, term *lateralization* is used instead. Lateralization is discussed further in sections 3.5 and 3.6, when reproduction over headphones is studied.

Blauert uses the following separation when speaking about the actual location of the sound source and the perceived location. *Sound event* is created by sound source and its position can be measured directly. *Auditory event* is the perceived event caused by a sound event.[9]

The localization accuracy is quite good, but not as good as it is with vision. A point source produces an auditory event that is spread in space rather than being a single accurate point. The spreading is called *localization blur* [9]. The minimum value of localization blur depends on signal and how the experiment is conducted. According to Blauert, the absolute lower limit of localization blur in the horizontal plane is around $1°$. With head immobilized, Haustein and Schirmer report localization accuracy of ±3.6° at $\varphi = 0°$ and ±10° at $\varphi = \pm 90°$ [35]. Behind the subject the localization blur is approximately twice its value for forward direction.

Changes in the elevation angle are not detected very accurately and the result, again, depends on the material used. Minimum localization blur varies from ±9° in forward direction to ±22° at $\delta = 126°$. Also, auditory events are biased towards forward direction compared to actual sound events.

Mechanisms of human sound localization are not yet fully know. Although the physical side, which can be measured, is quite well known, the neural side remains undiscovered. Measurable localization cues don't explain all phenomena like inside the head localization.

Sound localization accuracy is one of the criteria when evaluating loudspeakers, and localization is closely related to the perception of space. Since the method proposed in Chapter 5 tries to imitate these properties with headphones, in the next subsections, the basic cues for sound source localization are reviewed and distance sensation is investigated. Panning techniques used in stereophonic reproduction based on cues reviewed in Subsection 2.5.1 are presented in Chapter 3, Section 3.3.

2.5.1 Interaural Localization Cues

The fundamental cues of sound localization are time and level differences between the ears, namely *interaural time differences* (ITDs) and *interaural level differences* (ILDs). The latter ones are frequently referred as *interaural intensity differences* (IIDs). With the aid of monaural cues, these cues determine the sound location in free field conditions [9][10][17]. The next paragraphs follow ILD and ITD descriptions given in [15] and [17] and authors own experiments, although much earlier references exist.

ITD and ILD occur because of the shape of the head and distance between the ears. The maximum time difference appears to be around 700 μs when sound arrives from far left or right. This could be expected as the distance between ears along the surface of a sphere, which diameter is approximately 17 cm, corresponds to similar flight time for the sound waves.

ILD is highly dependent upon frequency. The head shadows effectively frequencies above 2 kHz as can be seen from Figure 2.6 where magnitude difference between left and right ear is plotted. The measurement is made in anechoic chamber, the sound source being at $\varphi = -90°$ angle. At low frequencies ILD is close to zero while at high frequencies the difference can exceed 20 dB.

Since level differences are too small at low frequencies, ITD dominates localization there.

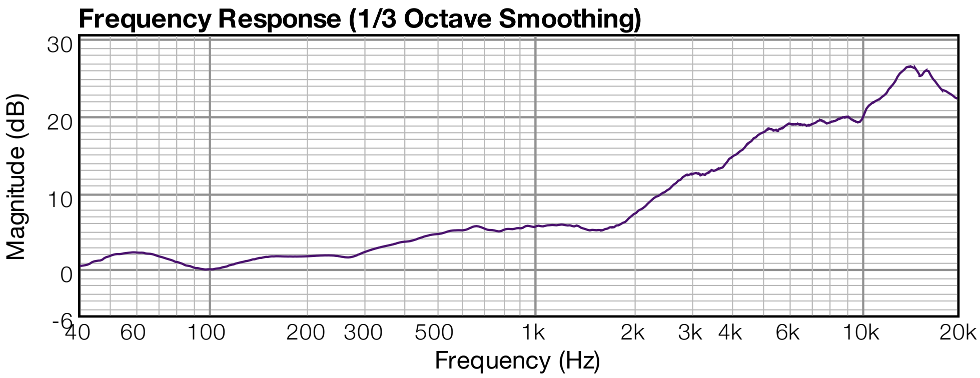


Figure 2.6: Magnitude difference of left and right ear measured in anechoic conditions. Sound source directly on the left side.

When moving to higher frequencies, the wavelength diminishes shorter than the distance between ears and phase information becomes unreliable. At high frequencies, approximately above 1.5 kHz ILD dominates sound localization. This phenomenon is called *the duplex theory* and it was first investigated by Lord Rayleigh in [36]. The duplex theory holds well with pure tones, but with complex sounds things get much more complicated. At frequencies above 1.5 kHz the time difference between time domain signal envelopes is used instead of phase information to determine the ITD.

If the head is approximated to be symmetrical, there are areas where ILD and ITD are equal between two points. For instance, sound sources at angles $\varphi = 0^\circ$ and $\varphi = 180^\circ$ produce the same ILD and ITD values. Planes where ITD and ILD values are inseparable are often referred as *cones of confusion*. Apparently, we are able to localize sounds between front and back and up and down which suggests that other localization cues than ILD and ITD must exist.

2.5.2 Other Localization Cues

One important aspect improving the directional hearing is the asymmetry of the head and torso. Head geometry and especially pinna cavities are highly individual and asymmetrical. The pinna works as a direction-dependent linear filter providing necessary differences to separate sounds arriving from different directions [9].

The filtering caused by torso, head and pinna to one sound source in anechoic conditions is referred as *head-related transfer function*, HRTF, in the frequency domain or *head-related impulse response*, HRIR in the time domain. HRTFs are strongly dependent on direction of arrival, highly individual and ear-dependent. Asymmetry provides localization cues needed when ITD and ILD are not enough. In room conditions, measurements are referred as *head and room-related transfer functions*, HRRTFs or *head and room-related impulse responses*, HRRIRs, correspondingly. If relative time and level differences are preserved in processing, it can be stated that HRTFs and HRRTFs include all auditory information needed to correct sound localization and reproduction.

Head movements provide important information that completes the ILD and ITD cues. If the head is kept still, front-back confusion may occur, especially in anechoic conditions. However, if head movements are allowed this kind of confusions almost never happen. It is understandable, since if the source is front-left or rear-left, different changes in ITD and ILD values are perceived depending on the direction of movement.

Visual cues are known to affect the localization. A classical example is that although the loudspeaker of old televisions is usually attached to one side, the voice of a news reader is localized to the center of the screen.

Room reflections affect the spatial information we receive. Barron states in [37] that in some cases early reflections diminish the localization errors. Begault et al. have found that reverberant conditions decrease the amount of localization confusions but increase the localization blur [38][6].

2.5.3 Distance

In anechoic conditions, distance perception is based on the sound level of the sound source. If the source is moved away from the listener, auditory event agrees very well with the actual sound event inside two meters. If the distance is still increased, distance of the auditory event asymptotically tends to about 10 meters, which is called *the acoustical horizon* [15]. Increasing the distance of the sound event does not increase the distance of the auditory event beyond the acoustical horizon.

Since the distance perception heavily depends on sound level, the familiarity to the sound material affects greatly the sensation of remoteness. According to Blauert, the distance of an auditory event corresponds very well with the distance of sound event if familiar sound material is used with typical sound level. If unfamiliar material or unnatural sound level is used, distance perception is altered. Surprisingly, the distance of the sound source seems not to have effect at all with unfamiliar sounds. Distance perception depends nearly only on the sound level instead of the location of the sound event.[9]

Reverberation eases the distance localization tasks significantly. We all can evaluate distances of hundreds of meters when sufficient reflections are present. Think about barking or distant traffic sounds. The auditory event is apparently very far although accurate meters can not be given. Begault noted in [38] that artificial reverberation had an effect on perceived distance. Added reverberation moves the auditory event away from the listener.

A special case of distance sensation is the zero distance, meaning inside the head localization. This happens when there is no reflections or reverberation in the binaural signal and/or other localization cues are unnatural or lacking completely.

2.5.4 Precedence Effect

The precedence effect (or the law of the first wave front or the Haas effect) helps us to localize sounds in room environment and prevents us to get confused of rapid reflections. The effect has been known for decades and although there are relatively old articles describing it [39], this description follows information given in [9][15][1].

While listening two sound sources, like in stereophonic listening, where speakers are in sixty degree angle from the viewpoint of a listener, delaying the left speaker localizes the sound to the right and vice versa. Delays above one millisecond force the sound completely to the right speaker if the levels of left and right are equal. When delay is increased, localization stays somewhat stationary until about thirty milliseconds is reached. With over thirty milliseconds delays, we start to hear two different signals localized left and right.

When speaking about the precedence effect, the experiment described earlier can be generalized. *The first wave front we perceive decides the localization of the sound.* Our hearing system rejects localization cues coming after direct sound in a 30 – 40 milliseconds time window independently of the direction of the sounds. This effect can be defeated by adjusting the level of the

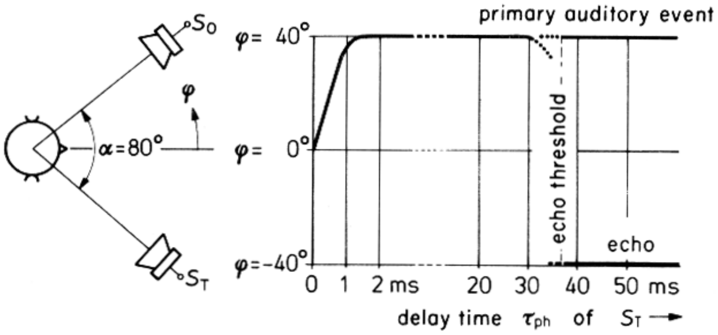


Figure 2.7: Sound localization and precedence effect. Adopted from [9].

later sound much higher. Localization cues of softer sounds are also more easily rejected. Figure 2.7 illustrates the effect.

When listening to loudspeakers in a room, the precedence effect plays a significant role. Firstly, it allows us to localize the sounds to the loudspeakers or between them. Without the precedence effect we would localize several sound sources around the room because of the reflections coming from the walls, floor and ceiling. Secondly, in stereophonic or multichannel listening, this effect makes the accurate loudspeaker and listener positioning important, since if one loudspeaker is closer to the listener than other, the sound localizes to the nearer loudspeaker.

Chapter 3

Sound Reproduction

Sound reproduction is the phase where an electrical signal is transformed to acoustical vibrations. This final stage of a recording and reproduction chain is critical: earlier efforts are undone if the transducer used is of low quality.

A variety of loudspeakers can be used to produce a propagating sound field to a listening space. Common properties are some kind of a membrane which vibrates according to the electrical signal fed to the device and an enclosure which prevents backward radiation to cancel the forward radiation. Reproduction over loudspeakers is heavily affected by room acoustics that often have greater effect than the free-field properties of the loudspeaker. That is why listening room acoustics and loudspeaker-room interaction are discussed in the following sections before considering the actual loudspeaker reproduction. Big variations in room response are the strongest motivation to the development of place-independend evaluating systems like the one described in Chapter 5.

Headphones offer another point of view: sound field is generated in close range to the ear. Room acoustics has minimal effect on the reproduction unless the environmental noise level is high. While kind of an place-independent reproduction is achieved, new problems arise. Accurate and repeatable reproduction is difficult to achieve because of bass reproduction problems and individual head and ear shapes. Different headphone designs and other issues like lateralization are discussed in Section 3.5.

Binaural reproduction is here used in the meaning of "reproduction of binaural recordings or binaural synthesis". Section 3.6 reviews some fundamentals of binaural techniques and sums up some problems related to the issue.

3.1 Listening Room Acoustics

In traditional auditorium and concert hall acoustics, the acoustical properties are handled statistically. Received sound divides to direct sound, early reflections and reverberation. Direct sound is the audio signal that propagates directly from a sound source to a receiver. Early reflections are the first part of the sound that arrives indirectly, meaning that the signal reflects from surfaces before it is received. Sources for some of the reflections can be found by measuring path lengths and travel time differences. When the amount of reflections grows so high that individual reflections can not be separated, the term reverberation is used. Usually the reverberation level decays exponentially towards a noise floor. In many applications, much effort is put to optimize the reverberation time, RT_{60}, which is the time required for a sound to decay by 60 dB.

Although controlling the early reflections is important also in small rooms like listening rooms, the statistical approach renders useless when room size diminishes, at least below certain frequency called the *Schroeder frequency*. Kuttruff formulates the Schroeder frequency in [40] as follows:

$$f_s = \frac{5400}{\sqrt{V\delta}}\ \mathrm{Hz}, \tag{3.1}$$

where f_s is the Schroeder frequency, V is the total volume of the room in cubic meters and δ is the average damping constant. In large rooms this frequency is so low that it can be ignored. In listening rooms, which usually are much smaller, the Schroeder frequency can be as high as 200 Hz. Below this, room characteristics are dominated by individual *modes*.

At certain frequencies, room becomes very responsive. These frequencies, at which standing waves are created in the room, are called the modes. In a rectangular space, the modal frequencies can be derived from the wave equation. The derivation is not showed here, since there are excellent references [40][31], but the result is:

$$f_m = \frac{c}{2}\sqrt{\left(\frac{n_l}{l}\right)^2 + \left(\frac{n_w}{w}\right)^2 + \left(\frac{n_h}{h}\right)^2}. \tag{3.2}$$

In Eq. (3.2) f_m is the modal frequency, n_l, n_w and n_h are integers and l, w and h are corresponding room dimensions. It can be seen that when n_l, n_w and n_h are of low order, low frequencies are received. When integers grow, possible combinations increase rapidly. It means that at higher frequencies there are a lot more modes than at low frequencies, i.e. the modal density is much higher.

Because of sparse spacing of the modes at low frequencies, there are some frequencies that are extensively boosted or cut depending on source and receiver placements in the room. As can be seen from Figure 3.1, there are some places where a specific frequency is cancelled out and areas where that frequency is boosted. This phenomenon is somewhat unavoidable, but it can be controlled with correct source and receiver placement and bass resonant structures.

At frequencies well above the Schroeder frequency, diffusiveness plays a significant role. Single strong reflections are considered disruptive and they decrease the sound fidelity. To avoid clear reflections but to keep the reverberation, diffusors are used [41]. A diffusor scatters the sound to different directions preventing strong peaks in room's impulse response. Figure 3.2 shows a common diffusor structure called quadratic residue diffusor.

Some reverberation is essential to natural sound perception. It affects our perception of space and the clarity of the source. Reverberation time can be controlled by adding absorptive materials like carpets and curtains. Reverberation time is widely used to describe the characteristics

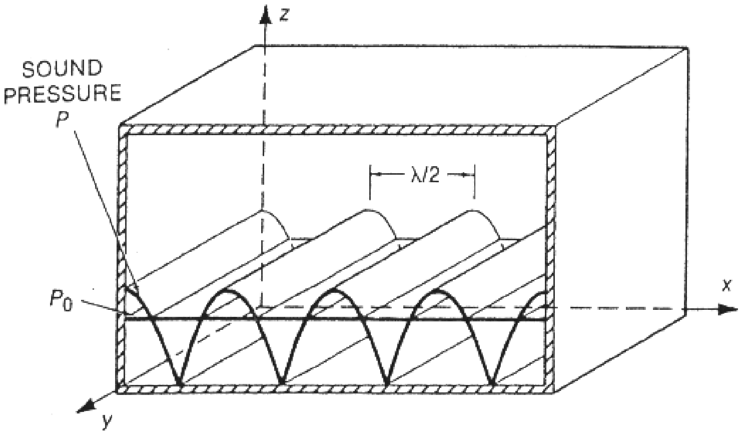


Figure 3.1: Sound pressure pattern of mode $n_l = 4$, $n_w, n_h = 0$. Adopted from [41].

of small rooms, although sound field in small rooms does not fulfill the definition of RT$_{60}$. The sound field should be random and well-mixed, but in small rooms there is only a series of reflected energy [41].

3.1.1 Listening Room Standards and Recommendations

International Telecommunications Union (ITU) and International Electrotechnical Commission (IEC) have specified listening rooms to use in critical listening tests [42][43]. Both recommendations specify geometric and acoustical properties of the listening room. Also loudspeaker and listening positions are defined. Since most of the rooms used for loudspeaker evaluation are built according to these specifications, the main goal of the virtual evaluation method proposed in this thesis is to imitate rooms like these.

The IEC report [43] gives exact dimensions that are considered ideal. Listening room volume should be 80 m^3 and the corresponding dimensions $(w \cdot l \cdot h)$ 4.2 m · 6.7 m · 2.8 m. This gives the total floor area of 28 m^2. The ITU recommendation [42] gives more freedom, because it specifies only proportions for width, length and height. Recommended floor area for stereophonic reproduction is given as 30 to 70 m^2. Note that this is somewhat larger than the corresponding

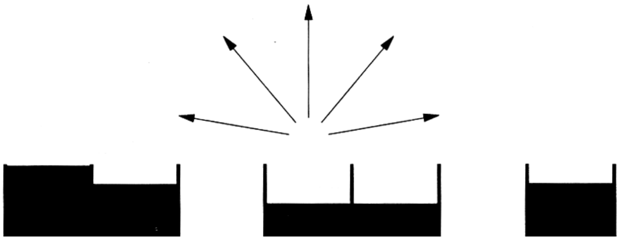


Figure 3.2: Profile of quadratic residue diffusor. Adopted from [37].

IEC criterion. Room dimension ratios should follow Eq. (3.3)

$$1.1w/h \leq l/h \leq 4.5w/h - 4, \tag{3.3}$$

where w, l and h are width, length and height, respectively. Additionally, the conditions $l/h < 3$ and $w/h < 3$ should apply.

According to IEC publication, the reverberation time, RT_{60}, should fall between 0.3 s and 0.6 s in frequency range from 250 Hz to 4000 Hz and less than 25% deviation from average value is allowed. Below 250 Hz and above 4000 Hz greater than 25% deviations from middle-frequency average are allowed, but at low frequencies the RT_{60} should not exceed 0.8 s. ITU gives middle-frequency reverberation time proportional to the volume of the listening room as follows:

$$T_m = 0.25\sqrt[3]{\frac{V}{V_0}}, \tag{3.4}$$

where T_m is the reverberation time between 200 Hz and 4000 Hz, V is the volume of the room and V_0 is reference volume of 100 m^3. T_m is allowed to deviate only 0.05 s in middle-frequencies, 0.3 s below 200 Hz and 0.1 s above 4000 Hz.

ITU recommends that early reflections during a time interval of 15 ms after the direct sound should be attenuated at least 10 dB in the range 1-8 kHz. IEC paper does not speak about early reflections, but recommends that the wall behind the listening position should be diffusive while the walls behind the loudspeakers and immediately to the sides of the loudspeakers should be reflecting.

According to ITU recommendation, the continuous background noise in the listening room should not exceed ISO NR10 curve, meaning that at 1 kHz the noise sound pressure level should be below 10 dB relative to 20 μPa.

Both ITU and IEC recommendations define the loudspeaker and listener placement for stereophonic listening in similar way. The distance between loudspeakers should be 2 meters at least. The ideal reference listening point is in the third corner of an equilateral triangle if the two loudspeakers define the two other corners of the triangle. Slight variations are allowed, but the angle between the loudspeakers form the listeners point of view should be between 55° and 65°. The reference axes of the loudspeakers should point towards the ideal listening position. Additionally, ITU recommendation defines that all loudspeakers should be at least one meter away from surrounding walls.

Listener and loudspeaker positioning is discussed further in Section 3.3.

3.2 Loudspeakers in a Room

A listening room is a highly complex transmission path from a loudspeaker to a listener. Loudspeakers are often measured in free field conditions where only the direct sound radiating from the loudspeaker towards the receiving point counts. This is somehow weird, since loudspeakers are always listened in rooms, where room modes and loudspeaker radiation to directions other than the receiving point makes significant difference.

Room boundaries give a remarkable boost to sound pressure level created by a loudspeaker. Theoretically, if an omni-directional loudspeaker is positioned to the corner of the room where it is surrounded by two walls and a floor, the power response of the loudspeaker is boosted by 9 dB [21]. Since loudspeakers are omni-directional only at low frequencies, only low frequencies are boosted in practice. Similarly, with two boundaries 6 dB increase in the power response can be found. Boost can be thought to be a result of summing image sources created by the boundaries.

A nasty property of loudspeaker placement is that the reflection from the wall behind the loudspeaker creates a strong dip to the magnitude response, if the wall is not perfectly absorptive. If loudspeakers were directional at the full frequency range, this would not be a problem, but unfortunately almost all loudspeakers are omni-directional at low frequencies. If the distance to the wall is l, a strong attenuation is experienced at the frequency corresponding to the wavelength of $4l$.

A loudspeaker excites different room modes depending on the placement of the loudspeaker. A monopole source will excite a mode fully, if it is placed at an anti-node of the mode. Respectively, when a loudspeaker is placed at a node, the corresponding mode will not be excited. According to Geddes, to get the best frequency response at low frequencies, as many modes as possible should be excited [31]. With a single subwoofer in a rectangular room, the best location is in a corner.

Different properties of a loudspeaker are emphasized depending on room acoustics and receiver distance. The first sound that arrives to the listener is always the direct sound with frequency response corresponding to free-field response of the loudspeaker. If the listener is located very close to the loudspeakers or the acoustics of the room is very dry, only the free field response matters. However, if the listener is far from the loudspeakers or there is a lot of reverberation, it is the power response, meaning the total energy radiated by the loudspeaker to all directions at different frequency bands, that determines the experienced response.

Geddes shows in [31] that the room response above the Schroeder frequency is a random variable and it can be analyzed only statistically. Computer simulations reveal that even slightest modification of loudspeaker or receiver placement, room shape, reverberation time or sound wave speed will result enormous changes in frequency response above Schroeder frequency. In Chapter 5, we try to freeze some of these chaotic variables by taking a snapshot of the room.

3.3 Stereo Reproduction over Loudspeakers

As described in Section 3.1.1, a standard stereophonic listening setup consists of two loudspeakers and a listener with the same distance between all components as in Figure 3.3. This formation provides the possibility to create phantom sources between the loudspeakers meaning that perceived sound doesn’t localize to the loudspeakers but between them.

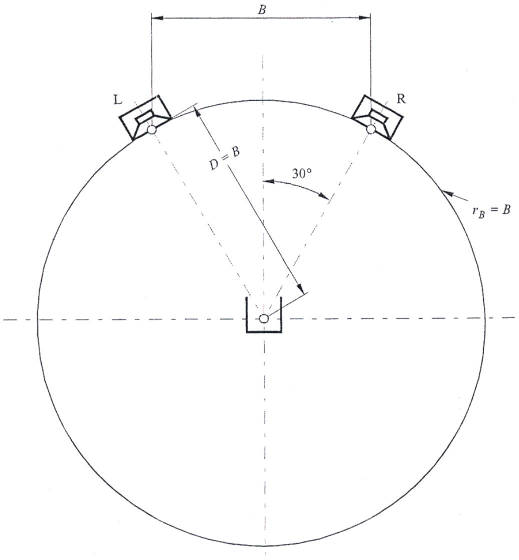


Figure 3.3: Standard stereophonic listening setup. Modified from [42].

It has been found quite early that level and time differences can be used to control the sound localization in stereophonic listening. According to Blauert, there are three things that may happen to the auditory event when the same signal is produced from two loudspeakers and time and level differences are varied [9].

First, if the time and level differences are small enough, the auditory event appears somewhere between the loudspeakers. Location is specified by levels and delays of both of the loudspeakers. Secondly, the auditory event can be localized based on only one of the loudspeakers, even if the other is radiating significantly too. Third case is that two auditory events appear. This happens usually when the time delay between the loudspeakers is sufficiently large.

Usually, level control i.e. *amplitude panning* is used if the auditory event is wanted to be between the loudspeakers. That is because of its better behavior at different frequencies. Figure 3.4 shows that while amplitude panning gives very similar curves at frequencies from 103 Hz to 1030 Hz, the hearing system responds quite differently to time delays at low frequencies compared to high frequencies. At low frequencies the auditory event moves almost linearly between the loudspeakers when time delay is varied between ±1 ms. At higher frequencies, curves differ significantly, especially with pure sine tones. As discussed in Section 2.5, this could be expected since phase information becomes arbitrary above 1000 Hz.

To some extent, time differences can be compensated with level differences and vice versa. Figure 3.5 reveals the trade-off with time delay and level differences. This is called *summing localization* meaning that localization is the sum of localizations given by time and level differences.

If the time delay between the loudspeakers is more than 1 ms but less than 30 ms, the auditory event appears to be in the loudspeaker that radiates the sound first. This is because of the

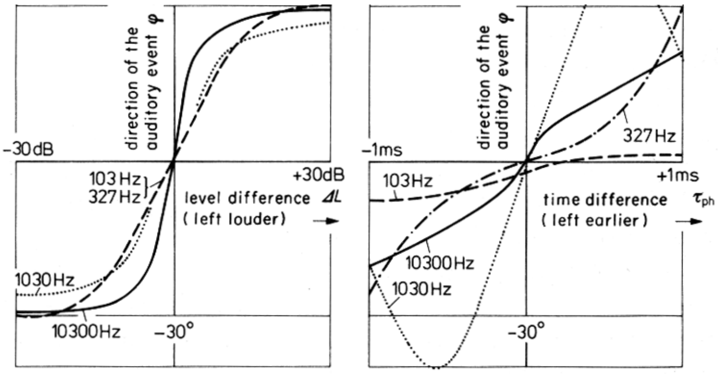


Figure 3.4: Localization with level differences and time differences. Adopted from [9].

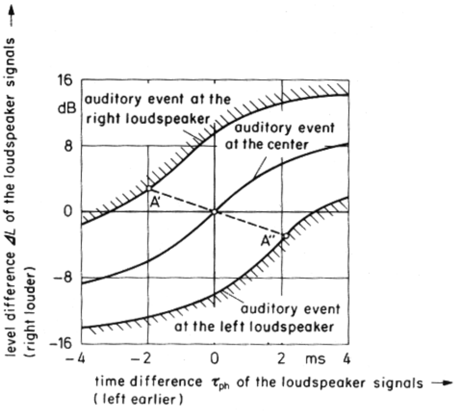


Figure 3.5: The trade-off between time and amplitude differences in stereophonic listening. Adopted from [9].

precedence effect described in Subsection 2.5.4.

Standard listening positions suggested in [42] and [43] are based on the information above. If the listener is closer to one loudspeaker than another, auditory event appears only in one loudspeaker. ITU recommends in [42] that listening area used should be in a radius of 0.7 m from the reference listening point described earlier.

3.4 Loudspeaker Listening Tests

In Section 3.1.1, recommended properties of a room and loudspeaker placement were shown. In this section, the rest of the IEC loudspeaker listening test standard [43] is summarized and problems of traditional loudspeaker listening tests are discussed.

Preferably only two pairs of loudspeakers should be listened to at a time if fast, silent and invisible mechanical substitution device is not available. Otherwise differences caused by different loudspeaker locations are easily greater than differences between the loudspeakers.

Program material should cover at least a speech sample recorded in anechoic conditions, classical music performed by full symphony orchestra and by smaller number of instruments and commercial rock or pop recordings. Classical music and speech are used because the test subject can compare them to real-life experiences. Commercial pop and or music should be used since it puts different demands on loudspeaker performance.

Sound levels of each loudspeaker should be balanced to equal level. Sound level can be measured with A-weighted, slow response measurement device at the listening position. However, this might not result in equal loudness levels, if loudspeakers differ considerably. Then, subjective methods should be used to equalize the loudness.

IEC recommends two different test procedures: single stimulus ratings and paired comparisons. In single stimulus rating, the test subject gives rating after every stimulus in scale from 0 to 10, 0 meaning worst imaginable reproduction and 10 being an ideal or true-to-life reproduction. In paired comparison method, stimuli are presented in paired sequences and ratings are given after two consecutive stimuli. According to IEC recommendation, paired comparison test is preferred by most of the listeners.

According to Toole, ideal loudspeaker listening test results should be repeatable, meaning that the same results should be achievable at different places. Results should reflect only the audible properties of a loudspeaker and give the magnitude of the differences found.[44] It is clear that the first point cannot be perfectly fulfilled when loudspeakers are evaluated in room acoustics. As discussed earlier, even a slightest change in listening room can make a great difference, and it is not possible to copy acoustics of one room to another. Loudspeakers can be hidden behind acoustically invisible curtains, but even then it can not be guaranteed that visual or emotional issues don't affect the results. Finding exact values for differences is not trivial since test subjects use given scales in different ways. Given verbal anchor points don't necessarily mean the same things to everyone.

As Toole states in [44], selecting the test material is a non-trivial task. The sound engineer has already made many aesthetic choices like microphone placement and type. Comparison to real life experiences is not possible in loudspeaker listening since there are no recordings that would present the audio as it was. Test material should be selected over a large number of recordings, since the test material should represent kind of an average of recordings available. Using only one randomly selected source can lead to biased results.

3.5 Reproduction over Headphones

Headphones produce pressure changes to a leaky chamber that consists of the ear canal, concha, outer ear and headphone cushion. It is like loudspeaker and a room in a much smaller scale. Although the differences between loudspeaker and headphone reproduction seem to be huge, one can ease himself with the fact that the ear is merely a pressure detector [2]. The ear does not care how the sound pressure is produced.

Compared to loudspeaker reproduction, headphones remove two important factors from the transmission chain. The listening room acoustics has no effect on headphone listening while in loudspeaker listening room reflections play a significant role. Also, there is no cross-talk between channels. In loudspeaker listening, both ears hear both loudspeakers in contrary to headphone listening where one ear hears only one signal.

Headphones can not provide true high-fidelity sound in its traditional sense since there is always some uncertainty in reproduction because of unique ear shapes and, especially with supra-aural and circumaural headphones, the effect of the headphone placement, which varies greatly. The effect of headphone placement on frequency response is studied further in Section 4.3. In general it can be stated that headphone responses are not repeatable above 8 kHz.

In the next two subsections, the basic headphone structures are shown and headphone design goals are investigated. The perceptual side of headphone reproduction is discussed and headphone sound localization is studied.

3.5.1 Headphones

This section is based more or less on Poldy's article in [21]. Thus, references are used only in case of direct quotation or if the reference is not [21]. Electrical equivalences are omitted but basic designs and transducer types are explained.

Headphones can be divided into three categories by design: circumaural, supra-aural and intra-aural. Circumaural headphones cover the whole pinna with quite large earshell. The earshell consists of a cup, which creates the volume around the ear, and a transducer that creates the varying sound pressure. In circumaural headphones the coupling volume of the cup and the ear can be as large as 30 cm^3.

A supra-aural headphone covers only the concha and some of the pinna. Supra-aural headphones provide less defined bass response because of the less determined leakage and cushion placement. The coupling volume consists of only the concha and ear canal. The frequency response of supra-aural headphones is less repeatable than for circumaural ones, although frequency response is not very stable variable in circumaural headphones either.

Intra-aural headphones don't cover the pinna at all but are inserted into the concha or into the ear canal. The frequency response of these headphones is more repeatable since there is not much room to move inside the concha or the ear canal, but the low frequency response is often poor because of low coupling volume. Measurement of the frequency response of intra-aural headphones is quite tricky compared to other designs.

According to Poldy, circumaural headphones are closest to high fidelity because of controlled bass response. However, there is always some amount of leakage in circumaural headphones that affects the bass response. The effect of the leakage can be noticed in a larger scale also: if you open windows and doors in a small room while playing music from loudspeakers, you will notice the lack of bass. The leakage is usually controlled by designing some intentional leakage that determines the response instead of random leakage.

There are five types of transducers used in headphones: isodynamic, moving-coil i.e. dynamic, electrostatic, electret, and electromagnetic. The electromagnetic transducers are not discussed here because they have no application in high-fidelity headphones. The isodynamic transducer consists of very light-weight conducting diaphragm between arrays of magnetic rods. Changes in current make the diaphragm to move. The drawback of this design is the low efficiency compared to the dynamic design.

The dynamic transducer is based on the same mechanism as dynamic loudspeakers or microphones: a diaphragm is attached to a coil in a static magnetic field and an alternating signal is led through coil. The advantage is high output efficiency compared to the isodynamic transducer, but the higher mass leads to a slow transient response.

The electrostatic transducer is not based on magnetic fields but electric fields. A statically charged membrane is driven with an alternating voltage creating a varying electrical field. According to Poldy, an electrostatic transducer gives the best transient response and transparent sound image. A disadvantage is the need of an extra high-voltage d.c. source for membrane polarization. The electret transducer uses the same idea but membrane is permanently polarized. Nothing comes free: a high signal voltage is needed since the permanent polarization cannot be very high.

3.5.2 Headphone Transfer Functions

To understand the equalization required for correct binaural reproduction, the transmission path from headphone to ear has to be examined in detail. This subsection follows the description given by Møller in [2].

In free field, Møller describes the transmission path from a source to the ear drum as in Eq. (3.5).

$$\frac{P_4}{P_1} = \frac{P_4}{P_3}\frac{P_3}{P_2}\frac{P_2}{P_1} \tag{3.5}$$

Here P_4 is the sound pressure at ear drum and P_1 is the sound pressure in the middle of the head while the listener is absent. P_2 is the sound pressure at blocked ear canal entrance and P_3 is the sound pressure at the entrance to the open ear canal, as in Figure 3.6. Møller calls the ratio P_3/P_2 *the pressure division ratio*, PDR. All ratios are direction and distance dependent with respect to P_1 but direction independent with respect to each other since, according to Møller, one-dimensional i.e. direction independent transmission starts even a few millimeters before the ear canal entrance.

Transmission from a headphone to the ear drum can be split up respectively. Figure 3.7 shows an anatomical sketch and its analogue model. Here E_{h} is the voltage at the headphone terminals, P_5 is the open circuit pressure at the ear canal entrance meaning the sound pressure with head absent, P_6 denotes the pressure at the entrance to an open ear canal and P_7 the pressure at the

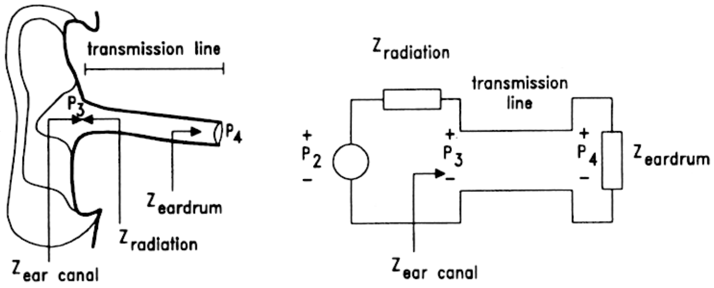


Figure 3.6: Sound transmission through external ear. Adopted from [2].

ear drum. Now we can split the transmission as follows:

$$\frac{P_7}{E_\mathrm{h}} = \frac{P_7}{P_6}\frac{P_6}{P_5}\frac{P_5}{E_\mathrm{h}}, \tag{3.6}$$

where P_6/P_5 and P_5/E_h depend on the headphone used, the test subject and the side of the head while P_7/P_6 depends only on the test subject and the side of the head.

It is noteworthy that free-field and headphone transmissions share some variables. Since transmission from an open ear canal to the ear drum does not depend on the sound source, we can write

$$\frac{P_7}{P_6} = \frac{P_4}{P_3}. \tag{3.7}$$

This equivalence is used in the next section when binaural recording and reproduction is reviewed.

The term open headphones refers different things depending on the source. In commercial

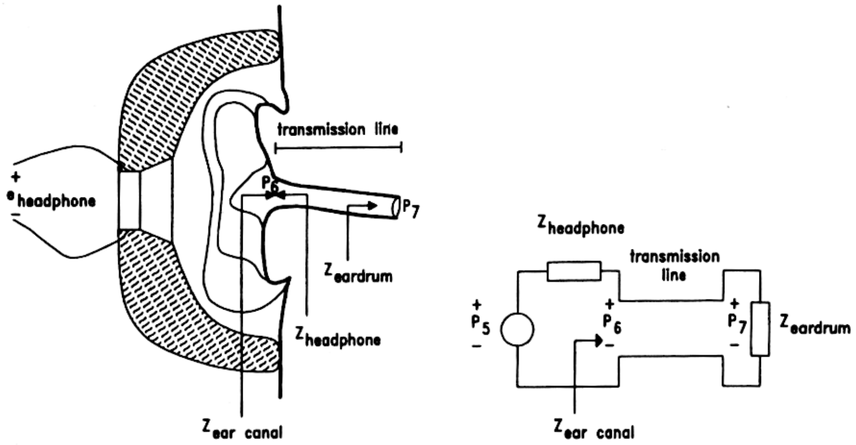


Figure 3.7: Sound transmission from headphone to ear drum. Adopted from [2].

area, openness refers to the property that headphones don't exclude outside sounds. Poldy uses the term to describe how the leakage is controlled in headphones. Open headphones have some intentional leakage while closed type headphones try to prevent all leakage. Møller uses the term in [2] to describe headphones that don't alter the radiation impedance seen from the ear canal at the ear canal opening. To be more precise, with open headphones

$$\frac{P_3}{P_2} \approx \frac{P_6}{P_5} \tag{3.8}$$

In his later papers Møller uses the term *free-air equivalent coupling*, FEC, to avoid confusions.

3.5.3 Magnitude Responses of Headphones

Diffraction and reflections from pinna unavoidably change the frequency response of incoming sound. Since headphones leave out some of the filtering caused by the head and ear, it is clear that flat frequency response at the ear canal entrance should not be the design target if any other source material than binaural recordings are used. With binaural recordings, flat response is desirable because pinna filtering is already recorded.

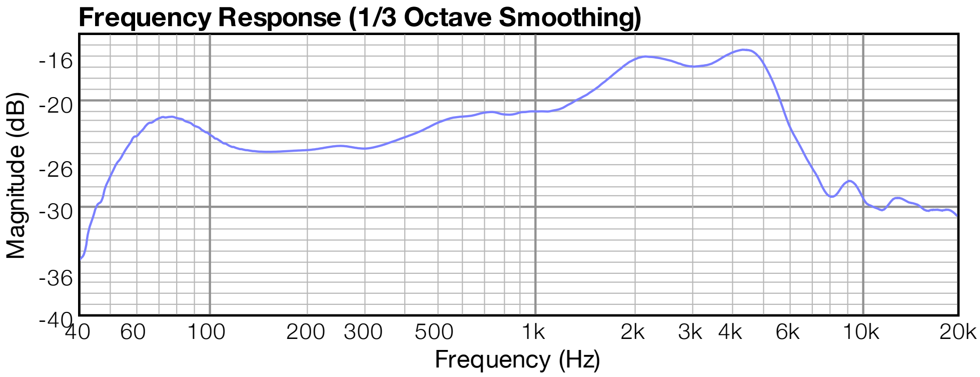


Figure 3.8: Left ear head-related transfer function averaged over 36 directions around the test subject. Measured at open ear canal entrance. Sound source was a Genelec 8030A loudspeaker.

Because the common use of headphones is to listen to material that is meant for loudspeakers, most of the headphone designs try to imitate the pinna filtering that would occur in loudspeaker listening. Figure 3.8 shows the magnitude response of a head-related transfer function measured in anechoic conditions and averaged over 36 directions around the test subject.

Certain similarity can be seen in the magnitude response of Sennheiser HD590 dynamic headphones. The response shown in Figure 3.9 is measured at open ear canal entrance with a miniature microphone. Both responses have quite remarkable boost at frequencies around 4 kHz and magnitude response decreases steeply after 5 kHz.

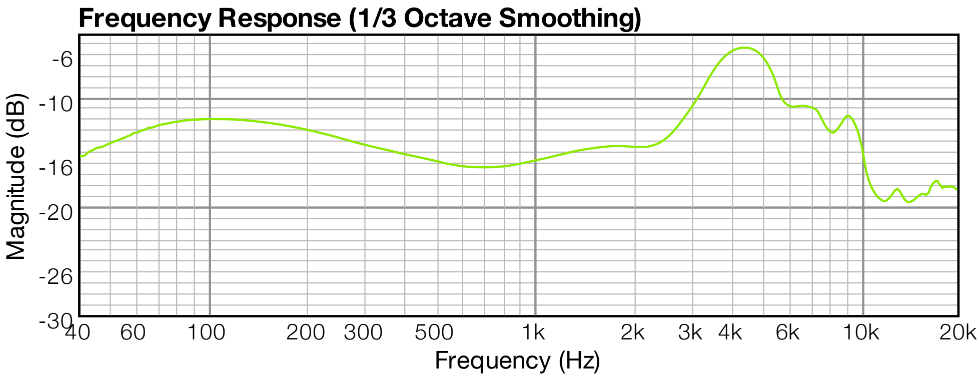


Figure 3.9: Sennheiser HD590 headphone magnitude response measured at open ear canal entrance.

The design goal of headphone magnitude response is either received by measuring ear transfer functions in anechoic conditions and averaging them or by making measurements in a reference sound field which could be for instance a standardized listening room. Møller et al. review earlier methods and propose a new one in [45].

An important point to be stressed out is that headphones are currently designed for reproduction of sound material that is usually reproduced over loudspeakers. This makes proper binaural reproduction even harder since the headphone design has to be undone to get a useful frequency response for binaural recordings.

3.5.4 Localization

Listening to audio material intended for loudspeaker reproduction results in a very unnatural sound image. Amplitude panned sources appear to be in a straight line between the left and right ears. The phenomenon is called lateralization or inside-the-head localization. The originally intended sound image diminishes to a group of point sources between the ears.

When thinking about monophonic sources that are positioned into a sound image by amplitude differences, it is no wonder that human hearing gets confused. The relation between ILD and ITD is missing and thus proper localization is impossible.

In addition, the absence of inter-channel cross-talk prevents a loudspeaker-like listening experience. In loudspeaker listening, some of the right channel material is always leaked to the left ear and vice versa. Headphone listening rules out this natural leakage.

Although headphones try to imitate a diffuse field magnitude response curve, it does not fix the problem with localization. It may result in quite natural timbre but not spatial image since in real life the reflections coming from different directions are all filtered separately with different HRIR filters.

Binaural techniques ease the problems mentioned here, but a few remain. Subsection 3.6.2 deals with the rest.

3.6 Binaural Recording and Reproduction

According to Møller, the idea behind binaural techniques is that the input to our hearing system consists of only two signals: sound pressures at the eardrums. If these signals are reproduced correctly, all aspects of a hearing event are repeated perfectly [2]. Headphones are the most practical reproduction device since they offer an almost complete channel separation. The method proposed in this thesis is based on the binaural theory and thus the theoretical recording and reproduction chain is investigated here.

The recording of binaural signals can be done either with HATS or with true-head techniques using miniature microphones. Different simulators are built, starting from spheres with two microphones to full scale replicas of average human upper body. The measurement procedure with HATS is more straightforward since microphone positions are fixed and HATS does not make unintentional movements or get tired. However, localization equal to true life is usually not achieved with simulators because of the individual character of HRTFs.

Møller et al. show in [46] that in terms of localization, best results are always achieved with individual recordings. HATS is only an approximation and can not provide good localization and timbre to everyone. For some people it works just fine while for others it does not work at all.

Individual recordings can be made at three positions without compromising the reproduction of spatial information: at the ear drum, at the entrance to the open ear canal or at the entrance to the closed ear canal [2]. Theoretical consideration for ear drum position is omitted here since it is not practical for the method proposed later.

Following the notation used in [2] and Subsection 3.5.2, the basic recording – reproduction chain for recordings made at the entrance to the blocked ear canal can be written as follows.

Let the microphone used have a transfer function M_1, which is frequency dependent. The electrical transmission path from microphone to the terminals of headphone is described by gain term G_b. To find out dependences for G_b we must put the transfer function of the reproduction chain equal to transfer function directly to the ear drum.

$$\frac{P_4}{P_1} = \frac{P_2}{P_1} \cdot M_1 \cdot G_\mathrm{b} \cdot \frac{P_7}{E_\mathrm{h}} \tag{3.9}$$

Solving G_b from Eq. (3.9) gives

$$G_\mathrm{b} = \frac{\frac{P_4}{P_1}}{\frac{P_2}{P_1} \cdot M_1 \cdot \frac{P_7}{E_\mathrm{h}}} \tag{3.10}$$

Now, using Eq. (3.6) and reducing we can rewrite

$$G_\mathrm{b} = \frac{\frac{P_4}{P_3}}{\frac{P_7}{P_6}} \cdot \frac{\frac{P_3}{P_2}}{\frac{P_6}{P_5}} \cdot \frac{1}{M_1 \cdot \frac{P_5}{E_\mathrm{h}}} \tag{3.11}$$

Substituting Eq. (3.7) and assuming FEC headphones described by Eq. (3.8) we can approximate

$$G_{\mathrm{b}} \approx \frac{1}{M_1 \cdot \frac{P_5}{E_{\mathrm{h}}}} \tag{3.12}$$

From Eq. (3.12) it can be seen that the electrical circuit must compensate only for the microphone transfer function and the headphone transfer function from headphone terminals to the blocked ear canal entrance. However, use of ideal FEC headphones is required to achieve correct reproduction. Since ideal FEC headphones do not exist, some error is present.

Corresponding derivation can be made to measurements with open ear canal. If the electrical transfer function is G_o, equation corresponding to Eq. (3.9) is

$$\frac{P_4}{P_1} = \frac{P_3}{P_1} \cdot M_1 \cdot G_{\mathrm{o}} \cdot \frac{P_7}{E_{\mathrm{h}}} \tag{3.13}$$

Solving G_{o} gives

$$G_{\mathrm{o}} = \frac{\frac{P_4}{P_1}}{\frac{P_3}{P_1} \cdot M_1 \cdot \frac{P_7}{E_{\mathrm{h}}}}, \tag{3.14}$$

which can be rewritten as

$$G_{\mathrm{o}} = \frac{\frac{P_4}{P_3}}{\frac{P_7}{P_6}} \cdot \frac{1}{M_1 \cdot \frac{P_6}{E_{\mathrm{h}}}} \tag{3.15}$$

The first term in Eq. (3.15) is unity because of Eq. (3.7). Now, we can finally write

$$G_{\mathrm{o}} = \frac{1}{M_1 \cdot \frac{P_6}{E_{\mathrm{h}}}} \tag{3.16}$$

Again, it is seen that electrical compensation for microphone transfer function M_1 and headphone transfer function from headphone terminals to open ear canal entrance is needed. However, microphone compensation can be avoided by using the same microphone for recording and headphone calibration.

Let M_1 be the transfer function of the microphone used for recording and M_2 the same thing for the microphone used to calibrate the headphones. In addition, let E_{m} be the voltage at the microphone output terminals. P_i in Equations (3.9) and (3.13) can be replaced with a measurement made with microphone producing

$$G_i = \frac{1}{M_1 \cdot (\frac{E_{\mathrm{m}}}{M_2} \cdot \frac{1}{E_{\mathrm{h}}})}, \tag{3.17}$$

which can be reduced to

$$G_i = \frac{M_2}{M_1} \cdot \frac{E_{\mathrm{h}}}{E_{\mathrm{m}}} \tag{3.18}$$

If the microphone transfer functions are equal and electrical transmission is expected to be ideal, Eq. (3.18) reduces to unity. Thus, microphone compensation is unnecessary.

Recording binaural signals at the entrance to an open ear canal is attractive, because it does not presume anything about the headphones used. Theoretically, any circumaural headphones that allow the measurement of headphone transfer functions will do. Yet, the repeatability of recordings made at an open ear canal entrance is not very good as will be seen in Section 4.3. Close to FEC headphones are recommended for reproduction since the microphone used will alter the sound field and produce a mismatch similar to recording at a closed ear canal entrance.

3.6.1 Transaural Reproduction

Unprocessed reproduction of binaural signals over loudspeakers leads to an unnatural sound image. Inter-channel crosstalk destroys the excellent localization characteristics of a binaural signals. However, it is possible to prevent the crosstalk in loudspeaker listening. Such systems are sometimes referred to as transaural systems [2].

The idea of crosstalk canceling is that a signal leaking from a left loudspeaker to the right ear is canceled by a signal coming from the right loudspeaker and vice versa. If symmetry is assumed, binaural recordings can be converted for loudspeaker reproduction using the suffler structure shown in Figure 3.10 [47]. H_i refers to the response from a loudspeaker to the ear at the same side (ipsilateral side) and H_c refers to the response from alternate side (contralateral side). According to Cooper et al. and Huopaniemi, H_i and H_c are joint minimum phase, meaning that they have common excess phase which is close to frequency-independent delay [47][48]. Cancelling filters can be then designed using minimum phase techniques and added delay.

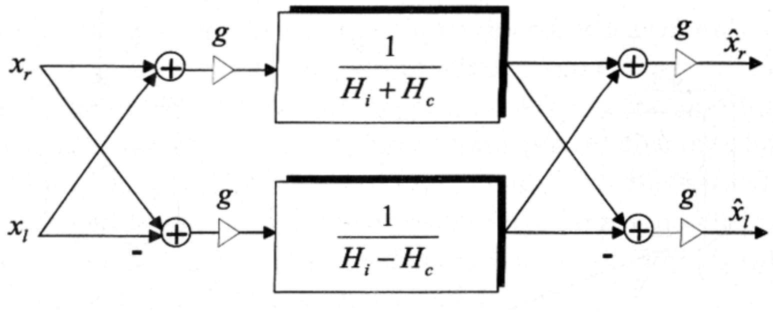


Figure 3.10: Suffler implementation of crosstalk canceling filters in a symmetric listening arrangement. Adopted from [48].

Transaural stereo reproduction suffers from similar problems as the headphone reproduction of binaural signals, although frontal localization may be easier to achieve if loudspeakers are visible to a listener. Transaural systems must be listened to in anechoic or at least close-to-anechoic conditions since reverberation prevents the correct crosstalk cancellation. An additional problem in transaural listening is a very small good listening area. Listeners outside the sweet spot can not localize sound sources correctly.

3.6.2 Problems in Binaural Listening

According to Møller, the worst problem in binaural listening is the poor frontal localization [2]. Sound sources directly in front of the listener tend to localize wrongly: behind or inside the head. Especially, recordings made with an artificial head can lead to front-back confusions and even lateralization. On the other hand, recordings made with a true-head are reported to work very well. Møller et al. found that human localization performance was comparable to real life localization with recordings made at the entrance to the blocked ear canal and using FEC headphones [46].

Toole lists some problems related to binaural techniques in [3]. Static localization cues which don't match with real cues are caused by artificial head or bad measurements made with a true-head. Transduction and coupling errors are caused by imperfections in headphone reproduction. Mismatches between auditory and visual cues may lead to wrong localization or inside-the-head localization. Visual cues often override auditory cues. Lack of dynamic localization cues causes a very unnatural sound image which rotates with the head. Again, the auditory image easily collapses inside the head. Body vibrations caused directly by high-level sounds or indirectly by floor and furniture are missing in headphone listening.

Some of the problems Toole mentions, like errors in static localization cues and transduction and coupling errors, can be minimized by careful measurements and accurate headphone correction, but some are difficult to overcome. Dynamic localization cues can be created with a full HRTF or HRRTF set and head tracking, but it requires lots of computing power and hundreds of accurate measurements. Visual cues could be created with a screen but it cannot correspond to full three-dimensional reality without special equipment. Special devices to create vibrations to furniture exist, but their usefulness in headphone listening can be questioned.

Chapter 4

Measurements

Accurate measurements are an integral part of many research areas and applications. Bad measurements can lead to false assumptions and decisions while good measurements can provide invaluable information about a test object. It is obvious that in acoustics, measurements are the most important source of new data.

Good measurements should be repeatable at different times and similar results should be achievable by independent studies. However, room responses can be considered as random variables at higher frequencies [31]. Even the slightest change in room conditions like air temperature and humidity or placing of furniture can result perceivable differences in direct comparison. Freezing all of the variables is impossible, but with careful measurement device placement and keeping all possible variables static, measurements can be repeatable to some extent.

In this thesis, accurate measurements play a significant role when the proposed method is implemented. If the effect of measurement errors grows greater than small impairments of loudspeakers under study, no reliable comparisons can be made. Thus, in the following sections the measurement techniques, equipment, accuracy and repeatability of the binaural room responses and headphone responses are under study.

4.1 Impulse Response Measurement Techniques

As known, acoustic space can be considered as an LTI system and it is completely characterized by its impulse response $h(t)$. The output of the system is convolution of the input signal and the impulse response, as in Eq. (2.3). Impulse response is important also for weakly nonlinear systems like transducers, since it reveals all linear properties of the system.

Impulse response can be achieved by a number of techniques which are based on the fact that if the input and output of the system are known, impulse response can be found using deconvolution. Mehods can be categorized based on the input signal type: MLS (maximum-length sequence) and IRS (inverse repeated sequence) use pseudorandom noise as input signal while time-streched pulses and sine-sweep techniques use time-varying signals [49].

Stan et al. have studied the pros and cons of different techniques in [49]. MLS technique is the oldest technique used to obtain LTI system response with deterministic input signal. The advantage of the MLS technique is that it is immune to impulsive disturbances. Instead of clear impulsive peaks in the system response, the MLS technique spreads the disturbance effects along the deconvolved impulse response as random noise. This is desirable, since the noise can be averaged out with multiple measurements. The disadvantage of the MLS technique are residues called distortion peaks. If the system under measurement is not linear, strong peaks appear in the time domain impulse response. The peaks can be heard as crackling noise when impulse response is convolved with an anechoic signal.

The IRS method is closely related to the MLS technique and the deconvolution method is exactly the same [49]. The IRS method has the ability to diminish the distortion peaks compared to the MLS method, although the peaks are not completely removed. Using time stretched pulses as input signal can give even better results, but it does not fully remove the peaks either.

The logarithmic sine-sweep technique proposed by Farina in [50] overcomes the problems related to earlier methods. The sine-sweep technique separates the harmonic distortions produced by a loudspeaker or another nonlinear device. After deconvolution, distortion components appear before the linear part in system response and can be removed from the final result. Another advantage is a better signal-to-noise ratio. According to Stan et al., in optimal conditions the swept-sine technique can produce even 20 dB better signal to noise ratios than the MLS technique.

In this thesis, all measurements are made using software called *FuzzMeasure* by Christopher Liscio. FuzzMeasure uses the swept-sine technique and allows the user to control several of the parameters.

4.2 Artificial Head and Torso Measurements

Binaural measurements can be made with a true-head, meaning that small microphones are attached to test subject's ears, or with artificial head and torso simulator (HATS), which is built to represent average human upper body. HATS has properties that make it superior to true-head measurements. HATS can be located accurately and it stays where it is put. Due to the sensitivity of the room responses to placement differences, the exact placement is essential for comparable results. The microphones of the HATS are also mounted permanently which removes the variances caused by microphones. Hair and clothing styles of the HATS are stationary in contrast to real test subjects. Also, all human factors are removed since the HATS does not get tired.

The disadvantage is, as stated in the previous chapter, the averaged nature of its responses. Although ITD and ILD cues might be very close to true-head responses, individual pinna reflections are not included and thus timbre doesn't necessarily correspond very well to subject's own responses.

The HATS used in this thesis is Manikin MK1 manufactured by 01dB-Metravib. Manikin is

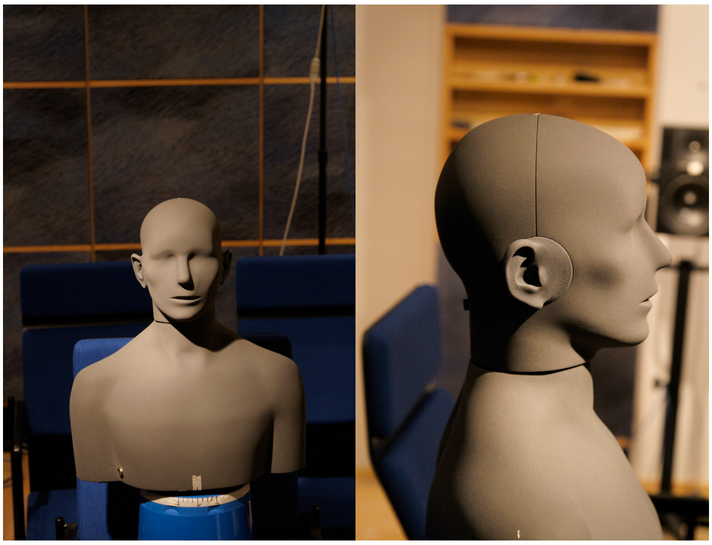


Figure 4.1: 01dB-Metravib Manikin MK1 in typical measurement position.

made of polyurethane with Nextel coating. The ear shape is in accordance with IEC 959 standard as well as DIN V 45608. Microphones are 1/2 inch condenser microphones positioned at the end of an ear canal which is 20 mm long. A separate preamplifier unit provides the polarization voltage for the microphones and all output connectors are located in the preamplifier. AD conversion is made in the preamplifier unit and audio data is transferred through an AES/EBU connection at sampling rate of 48 kHz. Figure 4.1 shows the manikin in typical measurement setup and Figure 4.2 shows a typical magnitude response measured in a standard listening room from a Genelec 1030A active loudspeaker at $\varphi = -30°$ angle to the left ear.

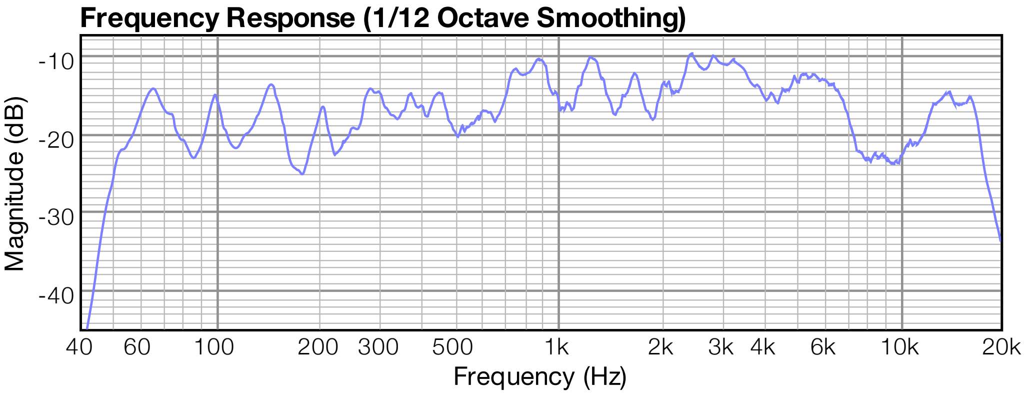


Figure 4.2: Typical magnitude response of the manikin measured in listening room from a Genelec 1030 active loudspeaker at -30° angle to the left ear.

Measurements were made in a listening room which is in accordance with the ITU-R BS.1116 recommendation reviewed in Subsection 3.1.1. A stereophonic setup was used meaning that the angle between the loudspeakers was 60°. The distance between the loudspeakers as well as the distance between each individual loudspeaker and the center point of the head of the manikin was measured to be 240 cm. Precise and repeatable positioning was confirmed with a plumb line hanging from the roof. The manikin was placed on a chair, head raised to the level of true-head listening.

All other equipment except the preamplifier unit were located in a separate control room. The AES/EBU signal from preamplifier was recorded to a computer hard-disk with MOTU Traveler external firewire audio interface connected to a MacBook computer. All signals ran along fixed cablings through the walls and thus all doors were closed.

4.2.1 Repeatability of HATS Loudspeaker-Room Responses

According to the complexity of the loudspeaker-room responses, it is questionable if repeatable measurements can be made at all. It is clear that the tolerances must be very small. Impulse response measurements were performed to find out the placement accuracy needed.

The measured impulse responses were convolved with stereophonic commercial rock music (Porcupine Tree: Trains from the record In Absentia) and monophonic pink noise. Four channels of convolution results were summed to two channels (left loudspeaker to left ear + right loudspeaker to left ear and left loudspeaker to right ear + right loudspeaker to right ear). The results were listened to with Sennheiser HD590 dynamic headphones. A program was made using Pure Data programming environment [51] to enable fast and seamless switching between different versions.

All measurements were made with HATS positioned as described in the earlier section. First, HATS was moved towards the line between the loudspeakers and binaural responses were measured for every two centimeters from each loudspeaker. Beyond 10 cm, only one measurement was made at 15 cm. Secondly, HATS was moved to the left parallel to the line between the loudspeakers one centimeter at a time and measurements were made. Thirdly, HATS was rotated 2.5° at a time from 0° to 10° and measurements were made as earlier.

Moving HATS forward was found to cause less perceivable differences than moving sideways. With music, a 15 cm movement provides a difference that is just noticeable. With pink noise, a 10 cm movement is noticeable. A placement change to side direction causes perceivable differences much faster. A one centimeter sideways movement is noticeable when listening to pink noise while a place change of three to four centimeters is perceivable with music.

Sensitivity to rotation depends highly on material. With pink noise, rotation of 2.5° made an audible difference, which was expected since earlier studies have shown that human localization blur in horizontal plane can be less than 2.5° [9]. However, sometimes even change of a 10° was found difficult to notice with music signal.

According to the results given here, it seems that to get comparable results, HATS should be placed very accurately. The sideways accuracy should be ±1 cm at least and the forward direction should be well specified. Variance of placement in frontal direction is not as critical as rotation and sideways placement but it should not be overlooked. It must be stressed out that these results were achieved only by informal listening by the author and thus should not be taken as an objective fact. In spite of that, the results give an idea how accurate the placement of the HATS should be. To some extent, the results can be applied to reproduction devices also.

To explore the overall repeatability of measurements, the following procedure was done. First, HATS was placed in the room as described earlier and first the measurement was made. Then, loudspeakers with stands were removed form the room and then carried back and positioned as they were. After measurements, HATS was removed and put back and the final measurements were made.

Similar informal listening as earlier was performed and it was confirmed that equipment can be located accurately enough to achieve comparable results. No difference was heard with music neither with pink noise.

Riederer has noted the excellent repeatability of HATS responses in anechoic conditions [52]. He measured HRTFs during a three-week interval and only ±0.2 dB variations were found. Figure 4.3 illustrates the repeatability in room acoustics in terms of magnitude response: only minimal variations are present even if the loudspeaker and the HATS are relocated.

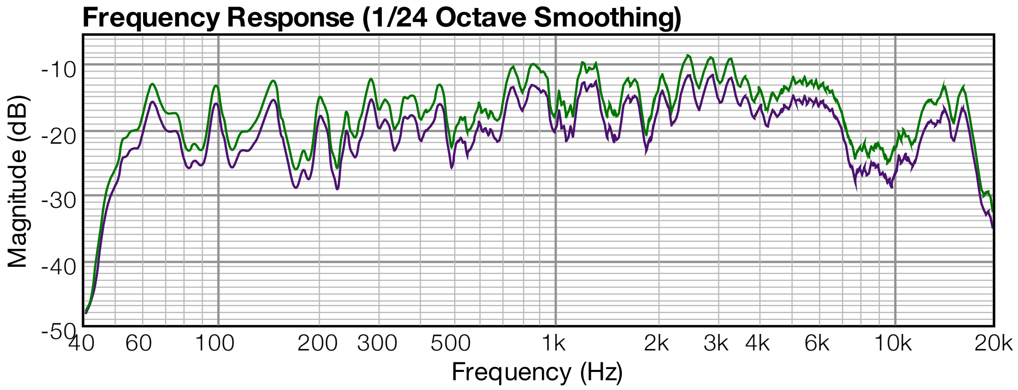
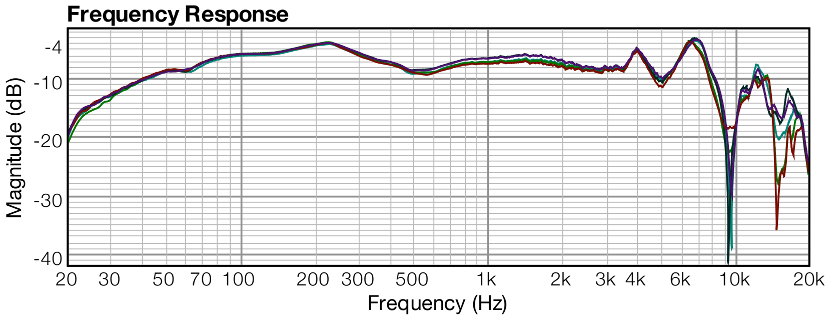


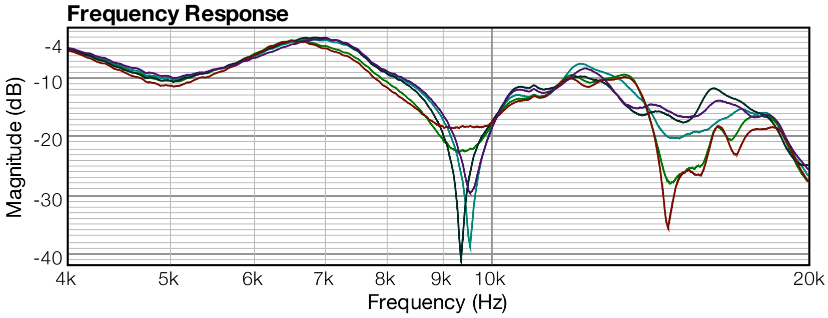
Figure 4.3: Responses from a Genelec 1030A loudspeaker at $\varphi = -30°$ to the left ear of HATS. The manikin and the loudspeaker were relocated between measurements. Curves are separated by 3 dB on purpose.

4.2.2 HATS Headphone Responses

In theory, measurement of headphone responses (or headphone transfer functions, PTFs) using HATS is simple. Microphones are fixed and measurement can be done anywhere if noise level is sufficient low. All you need is plug the cables in and put the headphones on the manikin. Yet, the last thing is the hardest. Even smallest differences in headphone placement cause large changes in the frequency response above 8 kHz.



(a)



(b)

Figure 4.4: Five consecutive measurements of headphone transfer functions with HATS. Fig. (a) shows the full audible range and Fig. (b) is zoomed to frequency range 4 kHz–20 kHz.

Figure 4.4 demonstrates the PTF repeatability using Sennheiser HD590 headphones. Responses have been measured five times consecutively. Headphones were taken off and put back between the measurements. Albeit effort was made to place the headphones equally, over 10 dB differences can be seen at frequencies above 7 kHz.

Møller et al. have studied headphone responses with human subjects and came to conclusion that the responses are reliable only up to 7 kHz [53]. Riederer investigated the repeatability of dummy head responses in [52] and noted that below 7 kHz responses agree very well. He achieved ±3 dB repeatability up to 13 kHz with Sennheiser HD580 dynamic headphones.

4.3 True-Head Measurements

Measuring room and headphone responses with true-head techniques brings lots of new variables to the system. Microphone placement can not be guaranteed to be exact and firm attachment is difficult. Head location can vary between measurements as well as during a measurement. As seen in Section 4.2, one centimeter can make a difference.

For binaural purposes, three measurement positions are of special interest: at the ear drum, at the opening to an open ear canal entrance and at opening to a closed ear canal entrance. Here, recording position at the open ear canal entrance is used, since it gives the best comfort to the test subject, measurements are easier to perform, and use of FEC headphones is unnecessary.

Room and placement related conditions are kept similar to HATS measurements discussed in Section 4.2. The only piece of equipment left in the measurement room was the preamplifier for the measurement microphones.

4.3.1 Measurement Equipment

Sennheiser KE 4-211-2 electret microphone capsules were used in the measurements. The diameter of the capsules is 4.75 mm and height 4.2 mm and the manufacturer promises almost perfectly flat frequency response from 40 Hz to 20 kHz. Capsules were soldered to cables and at the microphone end of the cable a thin and solid wire was wrapped around to give support and shape.

Two channel preamplifier Unides UD-MPA10e was used to amplify the microphone signals and provide polarization voltage. The preamplifier has two inputs, two outputs with coaxial connectors and shared gain controller. Figure 4.5 shows the microphones connected to the preamplifier.

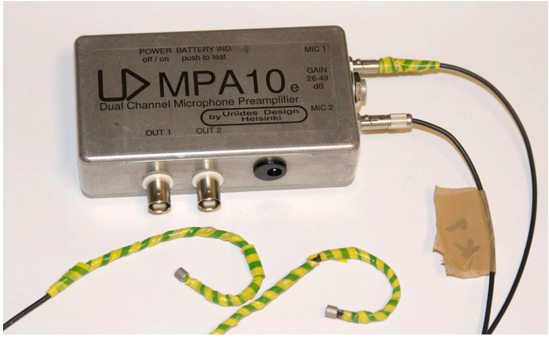


Figure 4.5: Sennheiser KE 4-211-2 microphones connected to UD-MPA10e preamplifier.

The frequency responses of the microphones were compared against a DPA 4191 free-field measurement microphone capsule with DPA 2669 preamplifier connected to a B&K Nexus amplifier. Figure 4.6 shows the results when a Genelec 1029A loudspeaker was the signal source. As can be seen, the frequency responses don't agree perfectly. The frequency response of the Sennheiser capsules rises towards high frequencies. However, the frequency response seems to be accurate enough for our purposes. The microphone response has not been corrected in the measurements presented in the next sections.

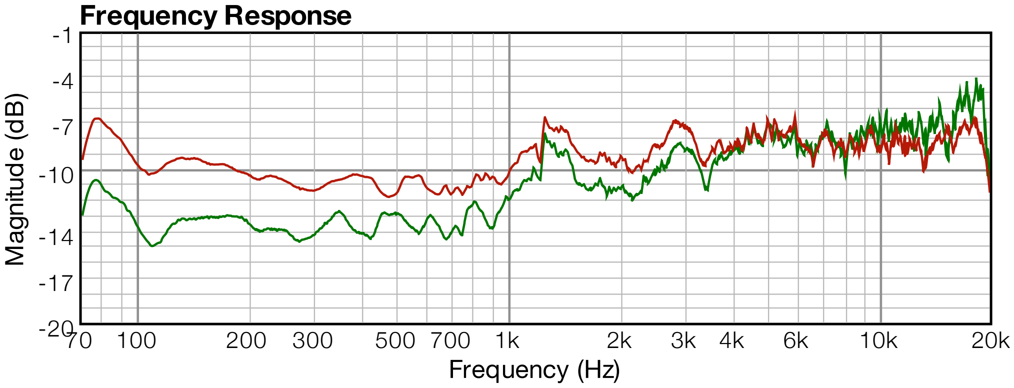


Figure 4.6: The magnitude response of Sennheiser KE 4-211-2 microphone capsule and UD-MPA10e preamplifier (green line) compared to DPA 4191 reference microphone (red line) in anechoic conditions. Signal source was a Genelec 1029A loudspeaker.

4.3.2 Repeatability of True-Head Loudspeaker-Room Responses

The microphones were attached to test subject's head as shown in Figure 4.7. The wire was twisted to fit behind the ear as an ear handle and sticky tape was used to relief strain and keep the microphones still.

To test the repeatability of the test arrangement, three consecutive measurements were made. Microphones were taken off and subject was allowed to walk for a while between measurements. Pictures were taken from the microphone attachments and special care was taken to place the microphones every time as similarly as possible.

The placement of the test subject's head was controlled with a plumb line hanging above the

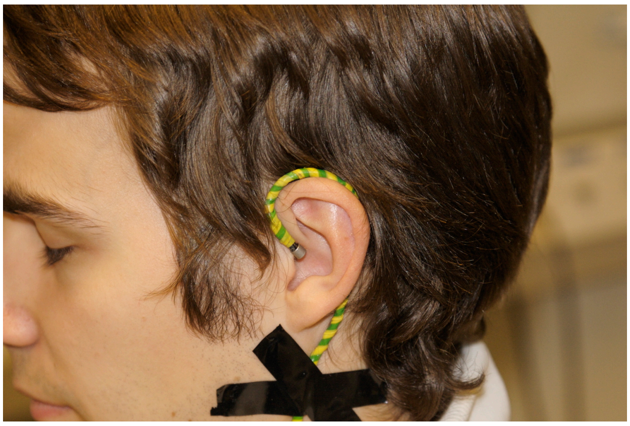
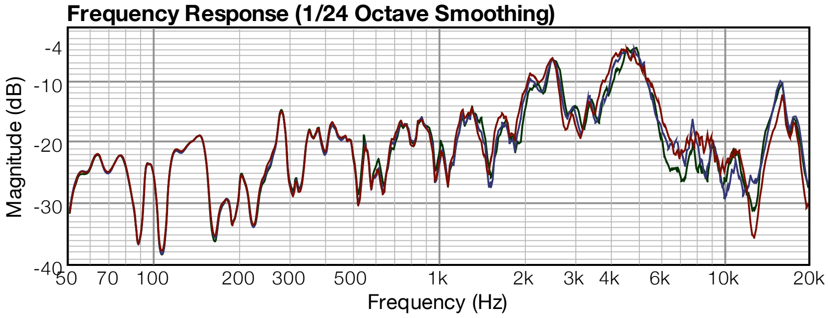


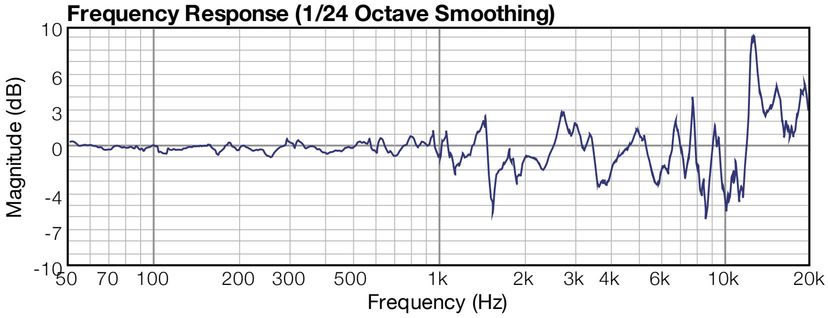
Figure 4.7: Microphone attached to subject's head.

head. It was quite easy to check if the plumb line was pointing to the center of the head. The test subject was asked to look at a black dot drawn in the front wall and to keep the head still.

As can be seen from Figure 4.8, frequency responses match very well up to 1 kHz but above that, different resonances appear. As could be expected, differences are very audible in direct comparison.



(a)



(b)

Figure 4.8: Comparison of three true-head measurements. Microphones were removed between measurements and the test subject was allowed to move. In Fig. (a), all magnitude responses are plotted in the same figure. (b) is the difference between measurements 2 and 3.

What is causing the differences can not be explicitly known, but a few guesses can be made. The microphone locations are not probably exact causing variance to the measurements. The head of the test subject cannot be located as accurately as the HATS and it may move during the measurement. Finally, human body is a time-varying noise source: blood circulation, breathing and gulps cause interferences.

Riederer noted similar sources of disturbances during the HRTF measurements. In addition, he pointed out that changes in subject's clothing, hair style, or spectacles could have effect on the repeatability of the measurements. [52]

4.3.3 True-Head Headphone Responses

True-head headphone response measurements have the same difficulty as HATS measurements: placing of the headphones makes the frequency response varying at high frequencies. In addition, microphone placement brings new variables to the game.

Here, similar techniques were used as in loudspeaker-room measurements. Microphones were attached to subject's head by experimenter. The headphones, Sennheiser HD590, were placed by the test subject, since Møller et al. have noted that it gives better repeatability [45]. Figure 4.9 shows that repeatability seems to be indeed better than with HATS. Headphones were taken off and put on a table between measurements. According to Figure 4.9, frequency responses are within 3 dB up to 13 kHz, which is a bit unexpected. Variation is almost constant with respect to frequency in contrast to HATS measurements, where much less variation was present at low frequencies.

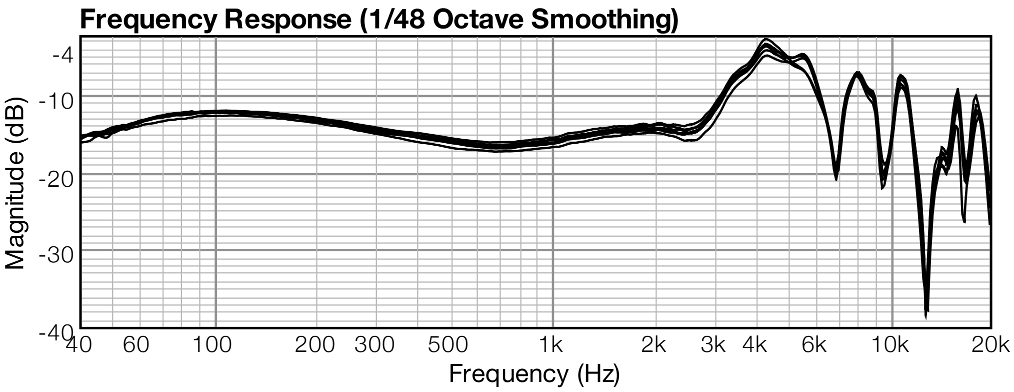


Figure 4.9: Repeated true-head headphone response measurements.

Headphone responses seem to be as individual as HRTFs and HRRTFs. Strong differences are present at high frequencies between ears and between test subjects. This would suggest that individual equalization has to be used to get proper binaural reproduction. Møller et al. have come to the same conclusion in [53] and [45]. Figure 4.10 demonstrates the differences between the left and right ear of the same test subject. As can be seen, frequency responses agree very well up to 2 kHz, but after that magnitudes and positions of high-frequency resonances vary greatly.

The microphone placement seems to have an significant effect on headphone responses when measuring at the entrance to an open ear canal. Figure 4.11 shows transfer functions when the microphones are removed and remounted between the measurements. As can be seen, variation between the measurements is greater than in Figure 4.9 where the microphones are kept in place and only the headphones are replaced. The effect is not necessarily caused only by microphone position. There was longer pause between the measurements and it may be that the test subject

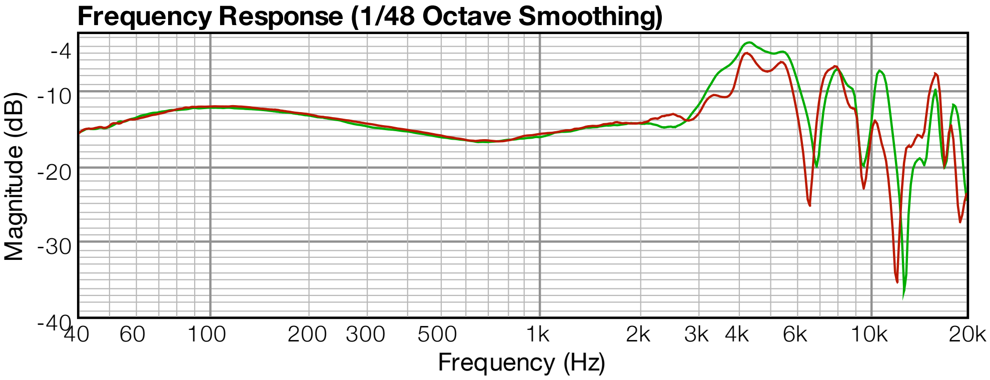


Figure 4.10: Headphone responses of left ear (green line) and right ear (red line) plotted in the same figure.

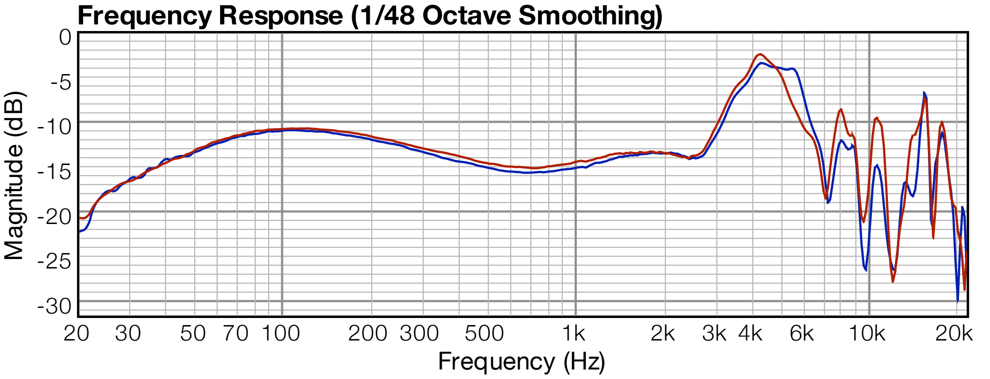


Figure 4.11: Headphone responses of two consecutive measurements when the microphones were replaced between the measurements.

could not place the headphones as similarly as in earlier test.

True-head headphone measurements seem to be surprisingly repeatable. However, measurements were all made in one session and effort was made to place the headphones similarly. Much greater variations would be seen if longer pauses had been kept, microphones were replaced or the headphones were just put on carelessly. Riederer has reported a bit larger deviations for measurements performed at open ear canal entrance. He found less than ±5 dB deviations up to 10 kHz, above which there are over 10 dB variations. [52]

Chapter 5

Method and Implementation

As seen in the earlier sections, a specific room and a loudspeaker form a complex system which can not be directly recreated in another place. The response of a room is unique and the placement of loudspeakers has significant effect on the perceived response. Thus it is not possible to compare one loudspeaker to another one, which has been listened to in a different room. The comparison is also limited by the auditory memory, which cannot provide solid references.

The listening position is critical when small impairments are investigated. As discussed in Chapter 4, very small movements in loudspeaker or receiver placements can cause perceivable differences in responses and prevent reliable comparison.

A method for loudspeaker comparison using binaural techniques is proposed in the next sections. The method eases some of the problems in the loudspeaker evaluation process and gives the possibility to compare a number of loudspeakers instantly without heavy equipment or a perfect listening room. The method consists of true-head and artificial head measurements in a standard listening room, a filter design method for artificial head to true-head equalization, measurements of true-head headphone transfer functions, and a filter design method for the headphone equalization.

From now on, we expect the stereophonic listening setup described in Section 3.3. In this work, effort is made to optimize the method for standard stereophonic listening but it could be easily extended to multichannel systems. The loudspeakers which are reproduced through headphones using the binaural technique are referred to as *virtual loudspeakers*.

In the following sections, all processing and measurements are done at 44.1 kHz sampling frequency if not mentioned otherwise.

In Section 5.1 the method is discussed in general level. Formal signal processing issues are shown. Section 5.2 explains what measurements are needed and the measurement setup is revealed. Sections 5.3 and 5.4 are reserved for signal processing issues related to measurements, and in Section 5.5 issues related to the listening arrangement and the convolution engine are discussed.

5.1 General Description

According to the binaural theory [2], an auditory event should be perfectly repeated if the same pressure signals are reproduced to the ear drums. To imitate the stereophonic loudspeaker listening setup with headphones, transfer functions from each loudspeaker to each ear, as in Figure 5.1, have to be measured. In addition, transfer functions from the headphone terminals to the ears are needed. Now, the listening experience in a specific room with a specific pair of loudspeakers can be repeated to a specific listener.

Any audio material can be listened to by convolving it with the binaural responses. If binaural responses for several loudspeaker pairs are known, the loudspeakers can be switched instantly without a delay and as the loudspeakers are measured at the same positions, the effect of loudspeaker positioning is ruled out.

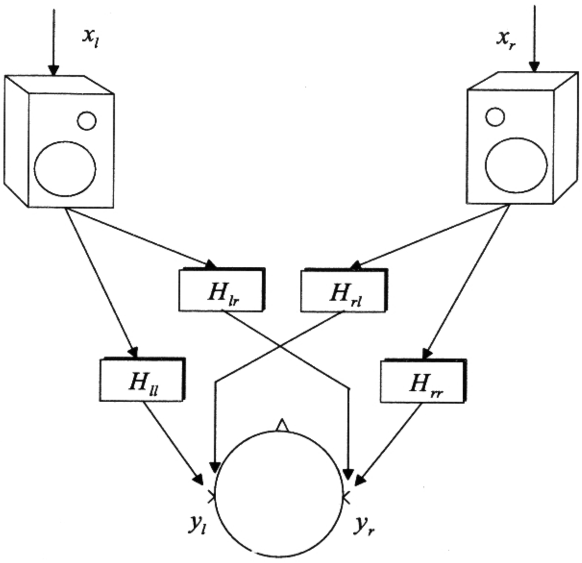


Figure 5.1: Transmission paths in the stereophonic listening setup. Adopted from [48].

5.1.1 Technique Using True-Head Responses

Signals in each ear consists of two signals: a signal coming from the ipsilateral loudspeaker and a signal coming from the contralateral loudspeaker. To achieve proper signals for the binaural reproduction, altogether four convolutions are needed. The left channel of a stereophonic signal, X_{l}, is convolved with the transfer functions H_{ll} and H_{lr} and the right channel, X_{r}, is convolved with the transfer functions of the opposite side. Finally, signals are summed as in Eqs. (5.1) and (5.2) and the headphone responses, P_{l} and P_{r}, are used to get the signals Y_{l} and Y_{r} for the headphone reproduction.

$$Y_\mathrm{l} = (X_\mathrm{l} H_\mathrm{ll} + X_\mathrm{r} H_\mathrm{rl}) / P_\mathrm{l} \tag{5.1}$$

$$Y_\mathrm{r} = (X_\mathrm{l} H_\mathrm{lr} + X_\mathrm{r} H_\mathrm{rr}) / P_\mathrm{r} \tag{5.2}$$

The transfer functions H can be achieved by true-head measurements. It must be stressed out that the transfer functions in Eqs. (5.1) and (5.2) are individual. To get proper reproduction, transfer functions have to be measured separately for every loudspeaker model and every listener. Also, headphone transfer functions P_l and P_r are individual and must be measured separately for every listener. Luckily, if the headphone responses are measured directly after the loudspeaker-room measurements using the same equipment, the effect of the microphones and other equipment is removed from Y_l and Y_r as discussed in Section 3.6.

In general, the exact inverse filter of the headphone transfer function, $1/P$, does not exist if P is not a minimum phase transfer function which it never fully is. Fortunately, it can be considered as near minimum phase, and minimum phase approximations can be used when designing the inverse.

Although the use of the true-head responses could give the best localization and timbre in a single case, they cannot be used for multiple loudspeaker comparison task directly. As seen in subsections 4.3.2 and 4.3.3, the repeatability of the true-head responses is not good enough. Variations between measurements could be greater than the differences between the loudspeakers under evaluation. Well comparable results could be achieved if all loudspeakers were measured in the same session keeping the microphones untouched and the test subject unmoved. However, the measurement session would become very long and adding loudspeakers later would be very difficult since the true-head measurements are difficult to repeat exactly.

From now on, the auralization using only true-head responses is referred to as *the true-head method*.

5.1.2 Technique Using HATS Responses

Since HATS measurements are found to be much more repeatable than true-head measurements, it would be advantageous to use the HATS to measure loudspeaker-room responses for each loudspeaker pair. Unfortunately, HATS responses correspond poorly to individual responses. As stated in the earlier chapters, the individual nature of HRTFs and HRRTFs prevents the use of the HATS responses directly.

To use the HATS responses instead of individual true-head measurements, the HATS responses must be equalized to match with the individual true-head responses. In theory, true-head responses and the HATS responses could be used together as in Eqs. (5.3) and (5.4)

$$Y_\mathrm{l} = \left(X_\mathrm{l} \frac{H_\mathrm{ll}^\mathrm{ref} G_\mathrm{ll}}{G_\mathrm{ll}^\mathrm{ref}} + X_\mathrm{r} \frac{H_\mathrm{rl}^\mathrm{ref} G_\mathrm{rl}}{G_\mathrm{rl}^\mathrm{ref}} \right) \cdot \frac{1}{P_\mathrm{l}} \tag{5.3}$$

$$Y_\mathrm{r} = \left(X_\mathrm{l} \frac{H_\mathrm{lr}^\mathrm{ref} G_\mathrm{lr}}{G_\mathrm{lr}^\mathrm{ref}} + X_\mathrm{r} \frac{H_\mathrm{rr}^\mathrm{ref} G_\mathrm{rr}}{G_\mathrm{rr}^\mathrm{ref}} \right) \cdot \frac{1}{P_\mathrm{r}} \tag{5.4}$$

where H^{ref} and G^{ref} refer to true-head and HATS measurements of a reference loudspeaker, G refers to a HATS measurement of a new loudspeaker, P refers to headphone responses and Y, X and indexes are as in Figure 5.1.

The problem is that the correction filters, $H^{\mathrm{ref}}/G^{\mathrm{ref}}$, don't exist since the loudspeaker-room transfer functions are mixed-phase transfer functions. Since the human hearing is quite inaccurate to phase changes, one possible solution is to perform the equalizations in the sense of magnitude responses and create minimum phases afterwards. Also, different loudspeakers may excite different room modes which makes the equalization more difficult.

To sum up, the method consists of the following steps:

1. Measurement of true-head loudspeaker-room responses with the reference loudspeaker pair at the entrance to an open ear canal.

2. Measurement of headphone responses with the same microphone placement. The calculation of individual inverse headphone filters.

3. Measurement of the reference loudspeaker pair with HATS in the same position where the true-head measurements were made.

4. Measurement of all needed loudspeakers with HATS in similar manner. Loudspeakers can be added later, if the measurement position as well as the loudspeaker positions are well documented.

5. Calculation of four individual filters for HATS to true-head equalization and the equalization of the HATS responses using the filters.

6. Preliminary loudness normalization of the responses.

7. Convolutions between the corrected HATS responses and test signals, headphone equalization using the pre-calculated filters.

The final stage can be done in realtime or the files can be processed off-line. However, computing power limits the number of loudspeakers which can be compared, if the realtime method is used. Four channels of convolution with typical response length of 14000 samples (about 0.3 seconds at 44.1 kHz sampling frequency) and the headphone equalization are needed for single loudspeaker setup, setting high demands on the computing power.

The HATS technique trades the problems of accurate receiver positioning and microphone placement to problems related to the correction of general HATS responses to the individual true-head responses. HATS offers superior repeatability compared to a true-head since its microphones are fixed and positioning is much more accurate. Adding new loudspeakers later on is easy if receiver and loudspeaker positions are well specified.

In the next sections, the auralization using HATS responses together with true-head responses is referred to as *the HATS method*.

5.2 Measurements and Equipment

The measurement equipment and setup used in the method proposed here are as described in Sections 4.2 and 4.3. As discussed in Section 3.6, different recording positions can be selected for the recording of binaural responses. The method described here is not limited to one position but recording at the entrance to an open ear canal is preferred for the following reasons.

If the microphone used to measure the binaural responses is small enough not to disturb significantly the sound field at the entrance to an open ear canal, only the measured headphone response needs to be compensated. In case of measurements done at the entrance to a closed ear canal, headphones with the FEC properties would be required, or the effect of the imperfect pressure division ratio (PDR, see Eq. (3.8) in Section 3.5) should be taken into account when designing the inverse headphone filters. Since the measurement uncertainty of the PDRs is high at frequencies above 7 kHz [53], it is better to avoid inaccurate corrections by measuring the responses at the entrance to an open ear canal.

Measurements at the entrance to an open ear canal give the maximum comfort to test subjects. There is no need for individual ear molds since the microphone setup described in Section 4.3 does not block sound and it is relatively easy to reshape for anyone.

Only one artificial head was tested in this study (01dB-Metravib Manikin MK1), but probably many others could be used. The shape of HATS responses should be as close as possible to true-head responses. It implies that a HATS with ear canals is preferred to closed type if the measurements are done at the entrance to an open ear canal. Unfortunately, the responses of the HATS used here do not match very well with true-head responses, as can be seen from Figure 5.2. This could be expected since the microphones of the HATS are located at the ear drum position. Also, the ITD of the HATS should match well to true-head ITDs. No ITD

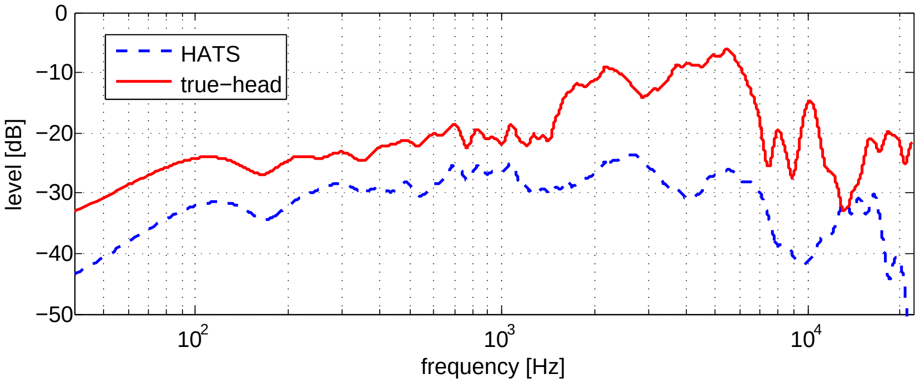


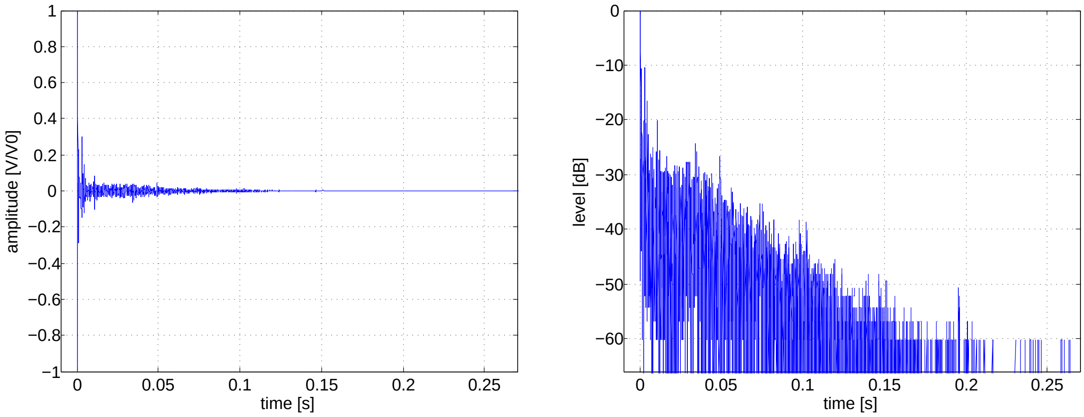
Figure 5.2: The magnitude responses of the 01dB-Metravib Manikin MK1 and a test subject in anechoic conditions. The measurements are from right ear, sound source (Genelec 8030A) being at $\varphi = +30°$ angle. Different resonances can be seen at high frequencies.

correction was implemented to the method since the differences were found to be smaller or equal to measurement inaccuracies. The true-head ITDs were from 10 to 13 samples at sampling frequency of 44.1 kHz, sound source being at $\varphi = \pm 30^\circ$ angle, while HATS provided ITD of 11 samples.

5.3 Processing of Loudspeaker-Room Responses

A low signal-to-noise ratio (SNR) could be a problem in true-head measurements. A head shadows effectively the signal arriving to the contralateral ear decreasing the SNR. Excellent SNRs are difficult to achieve even at the ipsilateral side because of noisy room conditions and the internal noise of the small microphones. To get usable results for the method, measured binaural responses are truncated and the HATS to true-head correction is calculated using the magnitude responses of the HATS and the true-head measurements. Two different correction techniques were tested and a minimum phase technique was chosen to the listening tests.

Figure 5.3 shows a typical true-head response of the ipsilateral ear. It can be seen that the response is strongly divided to early reflections and reverberation. The reverberation starts after few strong reflections, approximately 20 ms after the arrival of the first sound. The sound level has fallen almost 30 dB compared to the direct sound when the reverberation starts. It must be concluded that the effect of the early reflections is much greater than the effect of the reverberation.

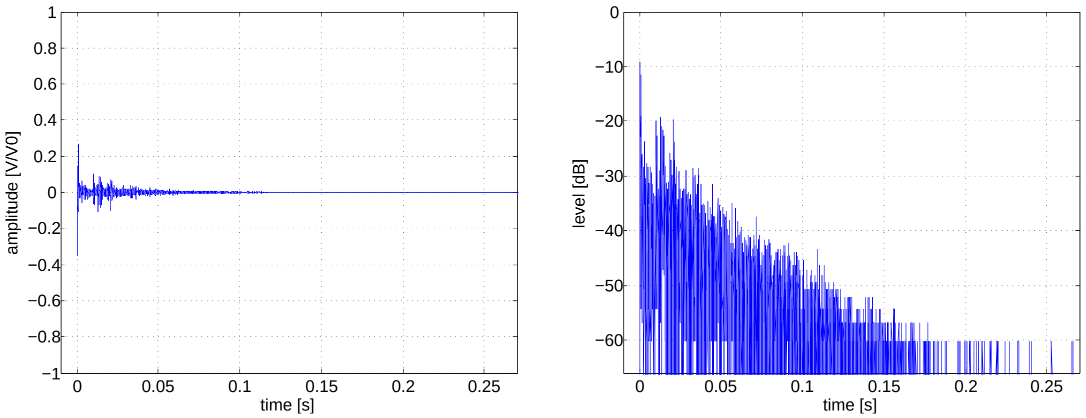


(a)

(b)

Figure 5.3: Typical true-head response of the ipsilateral ear. In Fig. (a), 12000 first samples are plotted. Fig. (b) is the same response squared and plotted on the logarithmic scale.

Referring to Figure 5.3(b), the noise floor seems to be around −60 dB. This is satisfying, since the SNR is limited not only by the environment but also by the measurement program, which allows only 16 bit recording.



(a)

(b)

Figure 5.4: A typical true-head response of the contralateral ear. In Fig. (a), 12000 first samples are plotted. Fig. (b) is the same response squared and plotted in the logarithmic scale.

The SNR in the contralateral ear is naturally less than the SNR of the ipsilateral response because of the shadowing effect of the head. Figure 5.4(a) shows a typical true-head response measured at the contralateral side. The SNR is 9 dB less than at the ipsilateral side.

The truncation point of the responses is determined iteratively. First, a response is truncated to the length of 3 times the expected reverberation time, squared and mapped to the logarithmic scale. Secondly, the response is smoothed with a rectangular window of 800 samples. Then, a preliminary truncation point is decided. The maximum point of the smoothed response was selected. A new estimation for the truncation point is achieved by calculating the average of the response between the estimated truncation point and the end of the signal. When the estimate and the average of the response tail are equal or inside a specified threshold, iteration is stopped. The response is faded linearly to zero after the truncation point. Figure 5.5 shows the effect of the truncation. Extra noise is removed and the exponential decay continues below the noise floor. The start point of the response is decided using a constant threshold.

The proposed truncation method is fast, but it may not lead to optimal truncation in all cases. The truncation method is considered to be a minor issue if the SNR is good. Karjalainen et al. review more advanced decay estimation methods in [54].

All responses are truncated as described before the convolutions or other processing.

5.3.1 Equalization of HATS Responses

As discussed in the earlier sections, HATS responses can not be used directly. Two different approaches to correct the HATS responses for the HATS method were tried. In the Kautz method, *a Kautz filter* is designed. A Kautz filter can be considered as a generalization of a finite impulse

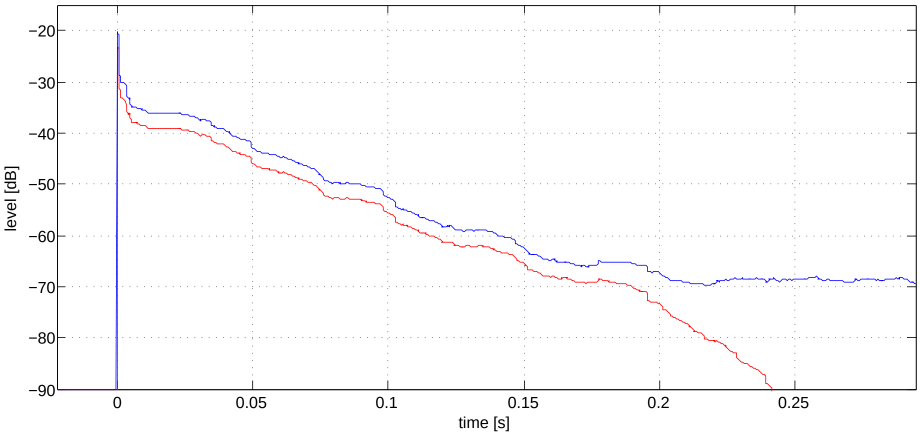


Figure 5.5: The effect of the truncation of a room response. Upper curve is the original response, averaged with moving rectangular window and mapped to logarithmic scale. Lower curve is the truncated response, averaged and mapped as previous. The curves are separated by 3 dB on purpose.

response (FIR) filter [55][56]. Ready-made Matlab functions for Kautz design are available in the Internet [57] and they were used without modifications. In the second method, the minimum phase method, a minimum phase FIR filter is designed based on smoothed loudspeaker-room responses. The problem was to find four filters which would fulfill the $H^{\mathrm{ref}}/G^{\mathrm{ref}}$ parts of the Eqs. (5.3) and (5.4).

The generic form of a Kautz filter is described by a transfer function

$$H(z) = \sum_{i=0}^{N} w_i G_i(z) = \sum_{i=0}^{N} w_i \left(\frac{\sqrt{1 - z_i z_i^*}}{1 - z_i z^{-1}} \prod_{j=0}^{i-1} \frac{z^{-1} - z_j^*}{1 - z_j z^{-1}} \right), \tag{5.5}$$

where $w_i, i = 0, \ldots, N$, are somehow assigned tap-output weights. Its time-domain counterpart is

$$h(n) = \sum_{i=0}^{N} w_i g_i(n), \tag{5.6}$$

where functions $\{g_i(n)\}_{i=0}^{N}$ are impulse responses of functions $\{G_i(z)\}_{i=0}^{N}$. Here, we omit further mathematical descriptions but one: the task of approximating a target response by a Kautz filter. Let $g_i(n)$ be a basis function formed from a selected pole set. Now, the approximation of the target response $h_{\mathrm{TE}}(n)$ is composed as

$$h_{\mathrm{EQ}}(n) = \sum_{i=0}^{N} c_i g_i(n), c_i = \langle h_{\mathrm{TE}}, g_i \rangle, \tag{5.7}$$

where $\langle h_{\mathrm{TE}}, g_i \rangle$ is the inner product of h_{TE} and g_i. The point is that the pole set, from which the basis functions are formed, can be freely selected. The model can be tuned by adding, removing or changing the poles.[56]

In the Kautz method, the correction of individual resonances of the loudspeaker-room-head system was tried. The 32768 point magnitude response of a measured true-head response is divided by the magnitude response of the HATS and a minimum phase is created for the resulting magnitude response using Hilbert transformation. Mathematical examination is out of the scope of this thesis but more about the issue can be found from [58] and [59].

The achieved impulse response is modeled using a Kautz filter. Based on preliminary informal listening tests, a logarithmic pole set was chosen. Figure 5.6 shows every second pole of the selected pole set which has 230 poles. As can be seen, there are more poles at low frequencies and the radius of the poles is decreasing towards high frequencies. The pole radius was decreased towards high frequencies because it decreases the chance of getting strong peaks in the magnitude response of the correction filter. Other pole sets, like linearly distributed in frequency, were tried but the logarithmic one was found to be optimal since it gives an impulse response of reasonable length and imitates the nature of the frequency resolution of the human hearing system. The amount of poles is based on preliminary informal listening tests. The number of poles was increased until the difference between a true-head response and an equalized HATS response was found imperceptible. The equalizer was designed using the same HATS response which was equalized.

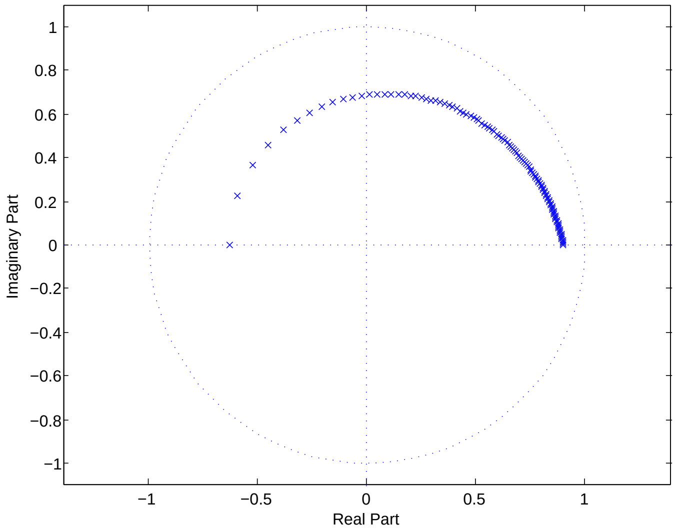
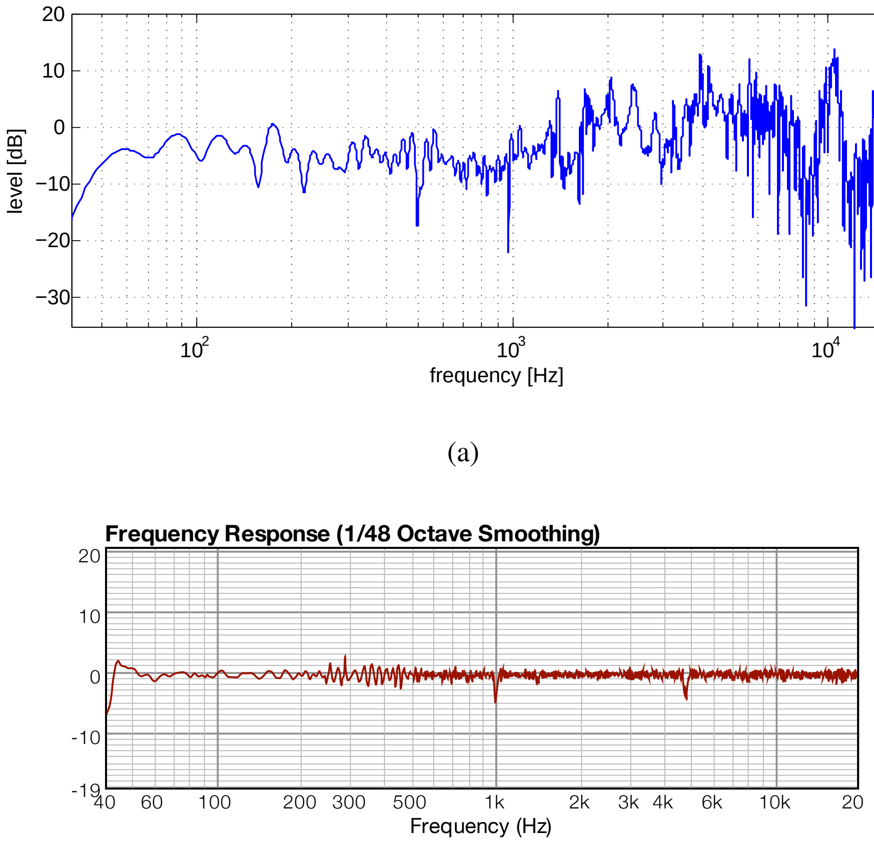


Figure 5.6: Every second pole of the pole set for the Kautz model.

Figure 5.7 shows the magnitude response of the designed Kautz filter and the difference between a true-head magnitude response and the equalization result when a HATS response is corrected using a filter which is created using the same HATS response. Considering the very slight smoothing in Figure 5.7(b), the filter seems to work very well. In 1/3 octave sense, the magnitude response is a straight line which implies very small perceptual differences. In informal listening it was found that the equalized HATS response was close to imperceptible from



(a)

(b)

Figure 5.7: (a) A magnitude response of a Kautz filter. (b) The difference between a true-head magnitude response and the equalization result when a HATS response is corrected using a filter which is created using the same HATS response.

the original true-head response. However, when equalizing a HATS response which is not used in the filter design, the results get worse. Figure 5.8 illustrates the performance of such equalization. A HATS response of Genelec 8030A is equalized using a filter which is designed from Genelec 1030A measurements. In terms of Eqs. (5.3) and (5.4), the signals to the ears are

$$Y_\mathrm{l} = \left(X_\mathrm{l} \frac{H_\mathrm{ll}^\mathrm{g1030} G_\mathrm{ll}^\mathrm{g8030}}{G_\mathrm{ll}^\mathrm{g1030}} + X_\mathrm{r} \frac{H_\mathrm{rl}^\mathrm{g1030} G_\mathrm{rl}^\mathrm{g8030}}{G_\mathrm{rl}^\mathrm{g1030}} \right) \cdot \frac{1}{P_\mathrm{l}} \tag{5.8}$$

and

$$Y_\mathrm{r} = \left(X_\mathrm{l} \frac{H_\mathrm{lr}^\mathrm{g1030} G_\mathrm{lr}^\mathrm{g8030}}{G_\mathrm{lr}^\mathrm{g1030}} + X_\mathrm{r} \frac{H_\mathrm{rr}^\mathrm{g1030} G_\mathrm{rr}^\mathrm{g8030}}{G_\mathrm{rr}^\mathrm{g1030}} \right) \cdot \frac{1}{P_\mathrm{r}}. \tag{5.9}$$

As seen from Figure 5.8, the filter does not perform very well. The difference around 2 kHz is very audible in direct comparison to the true-head response. The peaks in the equalized response

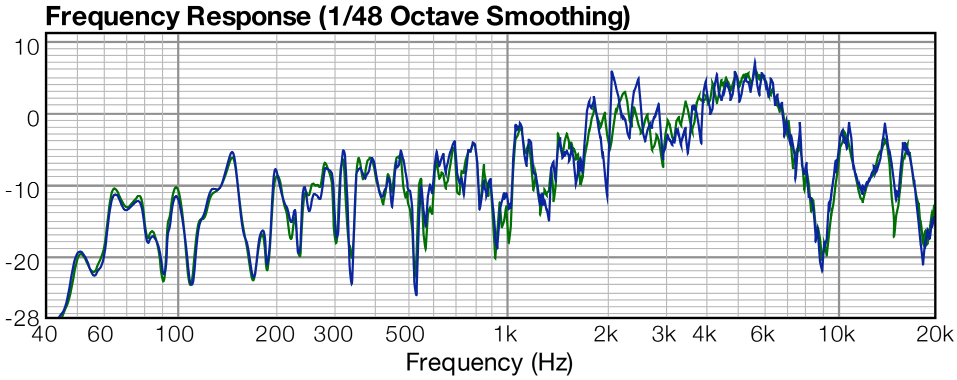


Figure 5.8: The magnitude response of a true-head measurement of a Genelec 8030A and a HATS response of a Genelec 8030A which is equalized using a Kautz correction filter designed from Genelec 1030A measurements.

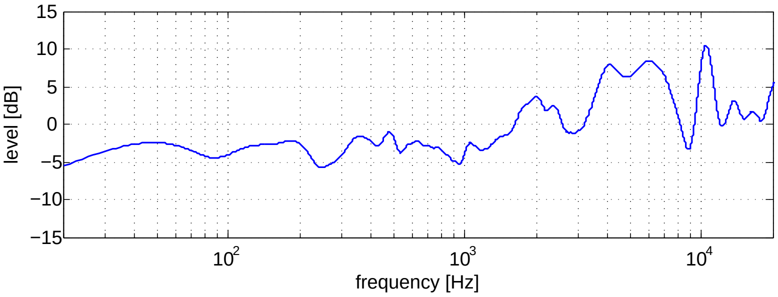


Figure 5.9: The magnitude response of a filter achieved by minimum phase method.

are generated when the peaks in the correction filter are not hitting the dips in the measured response and vice versa.

In the minimum phase method, equalization of individual resonances is not the target. Instead, only the general shape of the magnitude response is equalized. The performance of the equalizer depends on the shape of the HATS response. The closer the HATS response is to the true-head response, the better the result should be.

In the minimum phase method, the 32768 point magnitude responses of the true-head and the HATS responses are smoothed with moving hanning window. The smoothed true-head magnitude response is divided by the smoothed magnitude response of the HATS. In preliminary listening, $1/4$ octave smoothing was found to perform well. Minimum phase is created and the resulting impulse response is truncated. The truncation point is selected based on a constant threshold and absolute values of the impulse response. Figure 5.9 shows the magnitude response of a typical correction filter achieved by the minimum phase method.

Figure 5.10 illustrates a similar situation as in Figure 5.8 but with minimum phase equalization. As can be seen, peaks in the equalized HATS response around 2 kHz are gone. In that sense, the minimum phase method seems to perform better than the Kautz method. Yet, the minimum phase method does not correct the frequencies above 10 kHz as accurately as the Kautz method because of the smoothing which averages out the high frequency fluctuations.

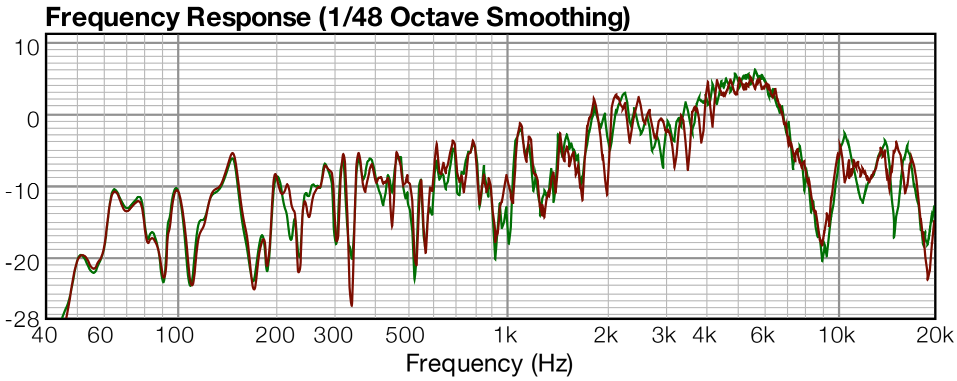


Figure 5.10: The magnitude response of a true-head measurement of a Genelec 8030A and a HATS response of a Genelec 8030A which is equalized using a minimum phase correction filter designed from Genelec 1030A measurements.

Another benefit of the minimum phase method is that the equalization performs equally well for all HATS responses while the Kautz method works very differently when the same HATS response, which is used in the filter design, is equalized. The difference between the methods is mainly caused by the target response. The Kautz method is used to correct the magnitude response resonance by resonance while the minimum phase method equalizes the general shape of the magnitude response. Kautz filters could be used in the latter case instead of minimum phase design.

5.4 Processing of Headphone Responses

The correction of the headphone transfer functions is the most critical issue related to the tone color of the binaural reproduction. As discussed earlier, headphones must be equalized to produce a flat frequency response at the point of the binaural measurement, in this case at the entrance to an open ear canal.

In theory, it would be enough to design an inverse filter, $1/P$, as in Eqs. (5.1), (5.2), (5.3) and (5.4). However, the direct inversion of the magnitude response does not provide optimal solution because of the variance in headphone placement. At high frequencies, resonances are moving slightly and the magnitudes of the resonances are changing from one measurement to another as in Figure 4.9. In the inversion, zeros of the headphone transfer function are changed to poles which create very high, narrow peaks to the inverted magnitude response. If the peak does not meet exactly a similar dip, it leads to a very audible and annoying resonance in reproduction.

According to Bücklein, peaks in magnitude response should be avoided at all costs. Single peak can devastate the reproduction while several equal dips may go undetected [60]. In addition, Toole and Olive state in [61] that wider resonances are detected easier than narrow peaks. Two guidelines can be now formulated for the headphone inverse filtering.

1. Avoid high peaks, especially the wide ones.

2. Do not widen the existing peaks and dips if possible.

In practice, the first one means that peaks in the inverted magnitude response of the headphone transfer function should be compressed somehow to ensure that there are no peaks above the average level in the equalized response. The second one means that the inverted magnitude response should not be smoothed too much since the smoothing widens the resonances and on the other hand it flattens the dips which are needed to compensate the peaks of the headphone transfer function in the reproduction phase.

Since the magnitude responses of the headphones need to be modified, the easiest way is to forget the phase information and generate minimum phase afterwards using the Hilbert transformation.

The proposed method for the headphone equalization is as follows. A measured headphone response is truncated to 512 samples. One-sided, 4096 point spectrum of the headphone response is calculated to get the magnitude response which is then smoothed slightly to get rid of the noise and unwanted small variations. The smoothing is done by averaging the magnitude response with a moving hann window. The width of the window is $1/48$ octave.

The smoothed magnitude response is inverted. In Figure 5.11, a typical result of the inversion of the magnitude response of Sennheiser HD590 headphones is shown. The dashed curve in Figure 5.11 represents the smoothed and inverted response. After smoothing, a level for peak reduction is decided based on the average level of the inverted response from 40 Hz to the frequency of the minimum magnitude value below 4000 Hz. Usually, the higher frequency limit hits the first resonance of the ear canal.

Now, all magnitude values exceeding the peak reduction level above the higher frequency value found for the peak reduction calculation are compressed. In the preliminary listening tests, $1/4$ compression ratio above the peak reduction level on linear amplitude scale was found to give satisfactory results. The effect of the peak reduction is shown in Figure 5.11. Since the peak reduction does not guarantee the absence of peaks at high frequencies, there is also a slight high frequency roll-off starting from 4000 Hz. The roll-off was designed to compensate the sharpness caused by unsuccessfully equalized resonances.

Finally, a minimum phase response is created and the impulse response of the filter is truncated when absolute values do not exceed 0.001. The filter design is done separately for each ear.

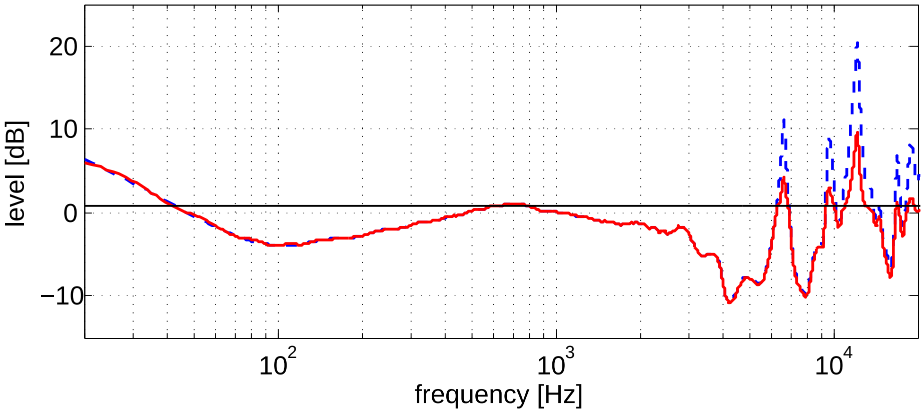


Figure 5.11: A typical inverted headphone response. Horizontal line indicates the peak reduction level. Dashed line indicates the original inverted response.

It is strongly recommended that the headphone response is measured in the same session where the binaural loudspeaker-room responses are measured. The remounting of the measurement microphones can move the resonances significantly resulting in improper headphone equalization for a specific set of binaural loudspeaker-room responses.

The effect of the headphone equalization can be studied by equalizing one headphone response with a filter created from another measurement. Figure 5.12 shows three different equalization cases. First, a headphone response is equalized using a filter which is created using the same response (the upmost curve). No peak reduction was used. As can be seen, the equalization seems to work very well. The effect of high frequency roll-off and impulse response truncation (low frequency ripple) are visible. Secondly, a headphone response is equalized with a filter created from another measurement, without the peak reduction (the middle curve). Equalization works fine up to 6 kHz, but after that random variations are present. The worst is the 4 dB peak around 11 kHz. The lowest curve in Figure 5.12 is the result of equalization using the peak reduction. The peak around 11 kHz is removed, but the dips above 6 kHz are much deeper. In preliminary listening tests, the peak reduction was found to be very important when reaching for natural sounding reproduction.

5.5 Convolutions, Listening Arrangement and Reproduction

To allow changing of responses and to improve flexibility, the impulse responses of equalized HATS responses and the impulse responses of the headphone equalization filters are stored separately. Pure Data graphical programming environment [51] was used to create a program which performs the realtime convolutions needed for the binaural reproduction of any audio material. On a MacBook computer with 2 GHz Intel Core 2 Duo processor, the program is capable of

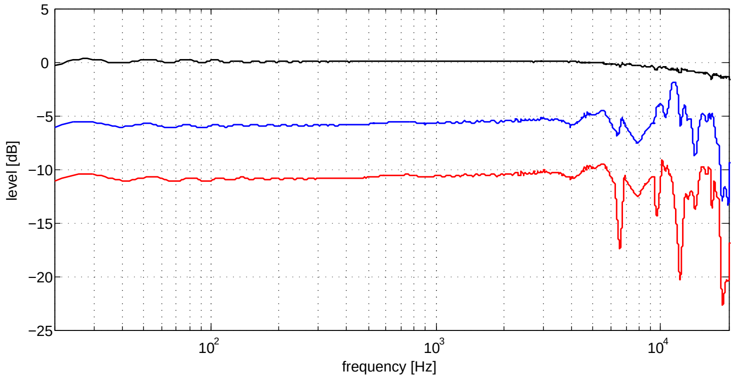


Figure 5.12: The upmost curve is the result of headphone response equalization with a filter designed using the same response and without peak reduction. The middle curve is the same but using different responses in the equalization. The lowest curve is as the middle one but with the peak reduction.

processing four sets of binaural responses, meaning 16 channels of convolution and the corresponding headphone equalization. Block size used in the FFT convolutions is 65536 samples which gives the maximum response length of 32768 samples when using the overlap-add real-time convolution method [11]. Parallel convolutions enable seamless switching between different responses. Four pairs of virtual loudspeakers can be compared instantly without any physical arrangements.

If possible, binaurally created virtual loudspeakers should be listened to in a listening room with visible loudspeakers. The ventriloquism effect seems to be very effective here: it was noticed that visible loudspeakers improve the externalization remarkably. Also, if the virtual loudspeakers are listened to in the same room where the measurements are made, real loudspeakers could be compared to the virtual ones. However, this should be done cautiously since the virtual loudspeakers are only a copy of the real loudspeakers measured in one position.

Without a head tracker, virtual loudspeakers rotate according to head movements. It is highly unnatural and may cause the sound image to collapse inside the head. To improve the externalization, a sound image can be constructed by switching on the virtual loudspeakers one by one. If the sound image is collapsed inside the head, the correct sound image can be restored by listening the virtual loudspeakers one at a time before switching them all on. The procedure stabilizes the loudspeaker positions and helps to keep frontal sources out of the head.

5.6 Summary and Discussion

In this chapter, a new method for subjective loudspeaker evaluation was proposed. The method trades the problem of loudspeaker placement and switching to the problems related to measurement accuracy and equalization.

The true-head method proposed here is meant to be only a reference for the HATS method. Although the true-head method might give the best localization and timbre, the uncertainties in the measurement procedure prevent adding new loudspeakers to the system. The HATS method tries to combine the good repeatability of the HATS measurements with the individual character of the true-head responses. The question is, if the HATS to true-head equalization is good enough to provide timbre and localization similar to the true-head method. Also, the reproduction of the virtual loudspeakers should be transparent and close enough to the reproduction over real loudspeakers allowing the listener to focus on the small impairments between the loudspeakers.

The usefulness of binaural methods in loudspeaker comparison can be questioned based on the averaging nature of the real listening situation. Loudspeakers are not listened to in a single position. Listener moves his head around which averages the listening experience across multiple points in time and space. On the contrary, binaural responses are merely a snapshot of a specific time instant at specific location representing only a fragment of the real listening experience.

The frequency range of the method proposed here is limited by the measurement devices and bad signal-to-noise ratios at low frequencies and measurement inaccuracies at the highest frequencies. There is a significant amount of low-frequency noise caused by the traffic, air conditioning and heating in most rooms. In practice, a high-pass filter at 22 Hz was used in the HATS measurements. The actual cut-off frequency depends on the loudspeaker used. It could be advantageous to digitally filter out frequencies below the cut-off frequency of the loudspeaker. At high frequencies, accurate reproduction of reality is limited by headphone placement and measurement inaccuracies related to that. However, the inaccuracies affect similarly all responses not necessarily decreasing the comparability between the virtual loudspeakers.

In spite of the mentioned defects, the proposed method overcomes the worst disadvantages of the traditional loudspeaker listening tests. The auditory memory is not the limiting factor since virtual loudspeakers can be switched seamlessly. Loudspeaker positioning doesn't have an effect on the evaluation since all loudspeakers can be measured on the same position. Although the reproduction over virtual loudspeakers may not be fully comparable to reproduction over real loudspeakers, the virtual loudspeakers are comparable to each other since the processing is the same for all loudspeakers.

Chapter 6

Evaluation

Although physical measurements and mathematical considerations give some idea of the sound quality of an audio system, there is no complete auditory model which could replace the subjective listening tests. However, a listening test gives only statistical information about the system under study. The results can be generalized to some extent, but individual responses may differ significantly from the average.

The performance of binaural recordings and the binaural synthesis has been evaluated in numerous studies [46][62][63][64][65] (for more see [10]). The previous research has mainly focused on the localization performance leaving issues related to the tone color intact.

In the previous chapter, a binaural method for loudspeaker listening and evaluation was proposed. The method was investigated using magnitude response graphs and informal listening. In the next sections, formal listening tests are conducted to find out, how well the virtual loudspeakers correspond to the real ones and are there perceptual differences between the HATS method and the true-head method described in the previous chapter. Spatial and spectral differences are evaluated using the verbally anchored ITU small impairment scale [42] and five different attributes.

Section 6.1 and its subsections explain the test arrangement, and in Section 6.2 the results of the listening test are shown. In Section 6.3 the results are analyzed and discussed.

6.1 Listening Test: Comparison to Reality

The idea of the test was that test subjects could compare the virtual loudspeakers reproduced by headphones to the real loudspeakers in the same room. As all environmental variables were static, only the differences in the reproduction would be significant. The loudspeakers were left visible to the test subjects to help the externalization.

The test had two goals. First, the test was performed to find out how well the virtual loudspeakers correspond to the real loudspeakers in terms of spatial properties and tone color. In the interest was, what properties the binaural reproduction would preserve and where the pitfalls would be. Secondly, the performance of the HATS method compared to the true-head method needed to be examined. The question was, if the performances of the methods are similar and if not, what are the differences.

The task was to evaluate the differences in the reproduction in terms of five attributes: *apparent source width, direction of events, distance to events, spaciousness, and tone color*. The three first ones are directly related to the localization performance. Apparent source width describes how the width of a sound source or sound sources is perceived. Is the source well defined or is it blurred somehow? Direction of events refers to the direction where the sound event appears to originate and distance to events is the distance from the listening position to the point where the sound event appears to be. Spaciousness describes the amount of space present in the listening. Tone color describes the spectral content of the sample. Written descriptions of the attributes given to the test subjects are in Appendix A.

The test subjects were asked to rate the difference between real and virtual loudspeakers using the previously described attributes and the ITU small impairment scale, which is a verbally anchored continuous scale from 1 to 5 [42]. The anchor points and the verbal descriptions with Finnish translations are in Table 6.1. The test subjects were able to adjust the difference rating in increments of 0.1 point.

Table 6.1: ITU small impairment scale.

Grade	Impairment	Ero
5	Imperceptible	Ei havaittavissa
4	Perceptible but not annoying	Havaittavissa, mutta ei häiritsevä
3	Slightly annoying	Hieman häiritsevä
2	Annoying	Häiritsevä
1	Very annoying	Erittäin häiritsevä

6.1.1 Listening Room, Equipment and Measurements

The listening test was arranged in the listening room of the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology. The listening room is in accordance with ITU standard [42].

Two pairs of active loudspeakers were used in the test, Genelec 1030A and Genelec 8030A. The loudspeakers were selected since they are both active and small enough to move around. Although the loudspeakers are very similar, differences in the tone color are easily heard with noise. The placement of the loudspeakers was controlled with plumb lines hanging from the ceiling.

The measurement equipment were as described in Chapter 4. The location of the test subject was controlled with a plumb line. The impulse responses were measured with swept-sine technique using 2000 ms logarithmic sweep and four times averaging. The measurements were repeated until consistent results were achieved. The quality of the responses was ensured with visual inspection. The HATS measurements were done 6 weeks before the listening tests, but rigging points for the plumb lines were not removed between the sessions.

Virtual loudspeakers were reproduced through the same Sennheiser HD590 headphones which were used in the earlier experiments.

6.1.2 Test Subjects

Altogether eight subjects, seven males and one female, participated in the test. None of the test subjects reported hearing damages except one who had continuous tinnitus. All participants were staff members or Master's thesis workers of the Laboratory of Acoustics and Audio Signal Processing. Although not all of the test subjects can be considered as experts in loudspeaker evaluation, everyone had at least some experience of participating in listening tests.

6.1.3 Samples and Processing

Three different audio sources were used in the test. *Anechoic male speech* which moves slowly from left to right and back gave an easy way to evaluate the directions and the tone color since the human hearing is specialized to analyze speech. A forty second extract of the song *Screen Play* from the record *Landmark* by *Mika Pohjola* was used since it has a wide spectrum and simultaneous sound sources located in different positions. *Pink noise*, meaning wide-band noise which has equal energy in each octave, is the most critical test signal when evaluating the tone color.

The three audio excerpts were auralized using the truncated true-head responses and the individually equalized HATS responses. The minimum phase method was chosen to equalize the HATS responses since it was found to give better results than the Kautz method in preliminary listening tests performed by the author. The responses of Genelec 1030A loudspeaker were used to design the equalizer for Genelec 8030A and vice versa. Altogether there were six different cases for one loudspeaker pair. Table 6.2 clarifies the different cases. The cases were repeated once and all attributes were rated in each case.

Table 6.2: The different samples in the test.

	Genelec 1030A		Genelec 8030A	
	method		method	
speech	true-head	HATS	true-head	HATS
music	true-head	HATS	true-head	HATS
noise	true-head	HATS	true-head	HATS

6.1.4 Test Procedure

The test was divided into four sections. First, the experimenter attached the microphones on the test subject's head and measured the binaural true-head impulse responses in stereophonic listening setup for each loudspeaker pair. Headphone responses were measured directly after the loudspeaker measurements. As the validity of the responses was ensured, the microphones were removed. The measurement phase took about 35 minutes including microphone positioning and changing of the loudspeakers.

While the audio files for the listening test were rendered, the test procedure was explained to the test subject. Written descriptions of the scale and attributes, available in Appendix A, were given. The test subject was advised not to pay attention to possible loudness differences or background noise. He or she was instructed to keep his/her head still and to look forward when listening to the virtual loudspeakers through the headphones. Also, the listening order of headphones first and real loudspeakers then was recommended but not forced. The test subject was allowed to familiarize himself with the material used in the test and to try to switch between the virtual and real loudspeakers. Signal processing and familiarization took approximately 25 to 30 minutes.

The evaluation phase was divided into two parts, one for each loudspeaker pair. A short break was kept between the parts and the loudspeakers were replaced.

In the evaluation phase, the test subject rated the difference of one virtual loudspeaker pair compared to the real loudspeakers using a computer mouse and the interface shown in Figure 6.1. As can be seen, the test subject could switch between the headphones and the loudspeakers

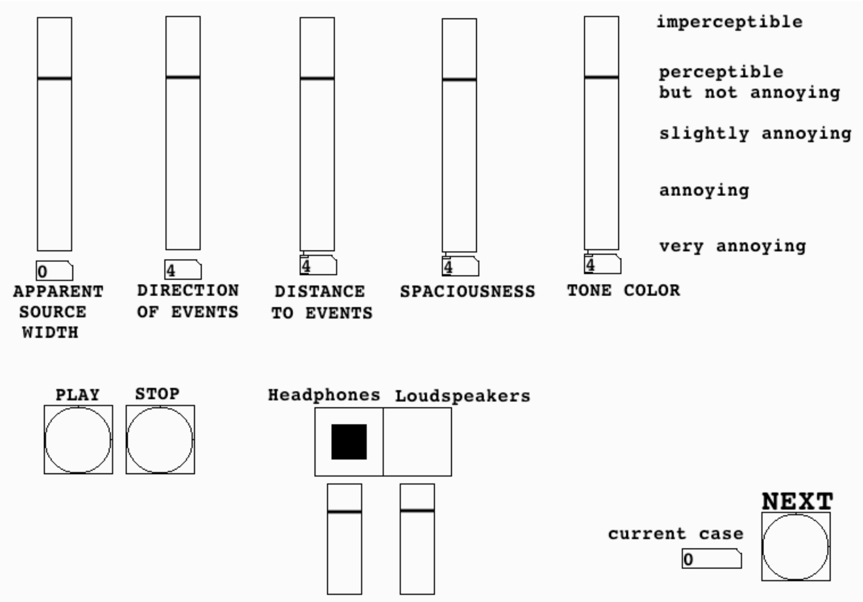


Figure 6.1: The user interface of the test.

at any time. Pressing the play button started the audio clip from the beginning but switching between the reproduction devices was instant. The reproduction continued from the switching point. The test subject was able to adjust the volumes of the reproduction devices to equal the loudness and he/she was advised to do so if the loudness of the loudspeakers did not match with the loudness of the headphones. There was no time limit. When one case was rated, the test subject could move on by pressing the next button. The order of the samples was randomized for each test subject. The first case, so-called zeroth case, was extra and it was not taken into account when analyzing the results. The zeroth case was only for test subject training.

The average duration of the evaluation phase was one hour including a pause between the two parts. After the second part, short verbal comments were asked.

6.2 Results

The received data was analyzed using the multi-way analysis of variances (ANOVA) and multiple comparison tests in the Matlab programming environment. The data was fitted to a normal distribution. Homogeneity of variances was tested using Levene's test and deviations from normal distribution were visually inspected. Although it was found that the data does not exactly fulfill the assumptions of ANOVA, ANOVA is known to be robust for small violations of the assumptions [66].

6.2.1 Means and 95% Confidence Intervals

Figure 6.2 shows the means an 95% confidence intervals for each attribute, test signal and processing method.

It seems that both methods work very well with speech signal. Apparent source width, direction of events and distance of events are all rated above 4.5, which corresponds to imperceptible in ITU small impairment scale. Spaciousness and tone color lie between 4 and 4.5 which corresponds to perceptible but not annoying. Although the means of the HATS method are slightly worse, the differences are small (< 0.1) and confidence intervals overlap.

All attributes get lower grades with music signal. The difference to the real loudspeakers is rated as perceptible but not annoying. The HATS method received worse grades than the true-head method in direction of events, distance to events, and spaciousness, but was rated equally good in terms of apparent source width and tone color. Again, confidence intervals overlap significantly. Distance to events decreases most when comparing the HATS method with the true-head method.

With pink noise signal, the means of all attributes except tone color are above 3.5 corresponding to perceptible but not annoying. The difference to the real loudspeakers in terms of tone color was rated as slightly annoying. The true-head method performs better on all attributes except apparent source width. The differences between the true-head method and the HATS method seem to be greater than with the two other test signals.

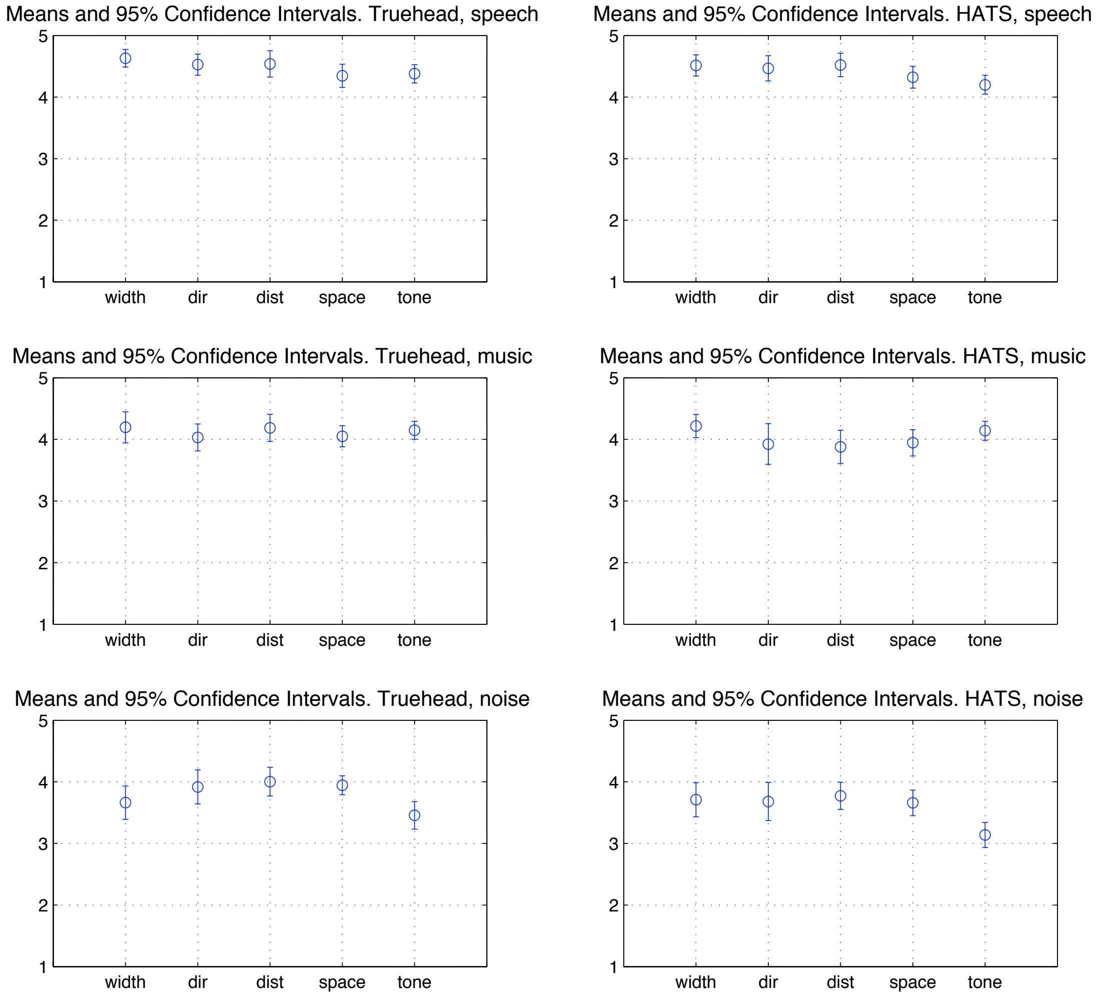


Figure 6.2: The means and 95% confidence intervals. The data from both loudspeaker pairs is combined. Y axis scale refers to the ITU small impairment scale.

6.2.2 ANOVA Main Effects

The six main effects in the analysis were the audio material used (sample), processing method of the binaural measurements (method), the attributes used (attrib), repetitions of the ratings (repet), the loudspeaker type (speaker) and a test subject (subj). All other main effects except the repetitions and the loudspeaker type were found significant ($p < 0.01$). There were also a few significant second and third order interactions but these are discussed in the next section. Full ANOVA table is presented in Appendix B.

The most significant effect was the audio sample. In further investigations with a multiple comparison test (Tukey's post-hoc test) it was found that the means of all three samples were significantly different. The difference between the reproduction methods was rated the highest with the noise sample while the jazz music sample and the speech sample got ratings closer to imperceptible, respectively.

The effect of the test subjects appeared to be significant which implies that the performance of the binaural method depends on the test subject. The multiple comparison test showed that one test subject gave significantly lower ratings while one of the eight subjects gave significantly higher ratings than others. This dependence could have been remove by normalizing the results. Since the scale had verbal anchors, the normalization was not performed.

The effect of the attributes is not interesting alone since it only implies that the attributes were graded differently, which was expected. Also the insignificance of the repetitions and loudspeaker type effects was expected. The test subjects were experienced and the method should work similarly regardless of the loudspeakers.

Although the effect of the method was found significant, the F value was low compared to the F values of the significant main effects. By visual inspection it was concluded that there is no perceptual difference between the true-head method and the HATS method or the difference is highly insignificant compared to other factors like the inter-subject variation. Of course, the conclusion is valid only in indirect comparison like the test described here.

6.3 Analysis and Discussion

The significant ($p < 0.01$) second-order interactions in the ANOVA table were sample*attrib, sample*subj, attrib*subj and repet*subj. Figure 6.2 confirms the sample*attrib interaction. The attributes are rated differently depending on the sample. The three other interactions are related to the test subjects, which confirms that either the performance of the binaural methods depends on the test subject or the subjects were not a very homogenous group. Most of the significant third-order interactions are also related to the test subjects. Sample*method*speaker interaction suggests that the loudspeaker might have some effect on the ratings. The conclusion is supported by the low p value of the main effect (0.08). In general, the F values of the interactions are low, indicating that the interactions are not as significant as the main effects.

The strong dependence of the ratings on the test signal (sample) is probably connected with the measurement and equalization inaccuracies at high frequencies. Most of the speech signal energy is below 4 kHz, above which the headphone equalization cannot be exact. There is more energy at high frequencies in the music and noise signals, which leads to the audible differences in direct comparison to real loudspeakers. One explanation can be found from the well-known problems of binaural techniques. The speech signal was moving and the movement started from the direction of a real sound source. The movement gave the feeling of the presence of dynamic localization cues helping the externalization remarkably. The noise signal was stationary and located in front of the listener where the performance of binaural techniques is the worst.

Inaccuracies in the measurement procedure may have increased the effect of the test subjects. If there were no time limitations in the testing, the test subjects could have subjectively selected the best measurement set instead of the visual selection by the experimenter. Also, it could be that the test subjects understood the scale or the attributes differently. However, the 95% confidence intervals in Figure 6.2 are small, indicating that most of the test subjects rated the attributes similarly.

The true-head method and the HATS method are easily distinguishable if compared directly to each other. In comparison to reality, the methods seem to perform equally well. The equal performance of the methods can be only an illusion since the test subjects could not compare the two methods directly. To examine the difference between the methods in detail, a new test should be organized.

Many of the test subjects were astonished by the perfectness of externalization of the male speech voice. In the familiarization phase, two of the test subjects were convinced that the sound came from the loudspeakers although it came from the headphones. All critical comments were related to high frequencies. Either the sound was too bright or the high frequencies were not located correctly. At least four of the test subjects reported that the localization of high frequencies was not correct in the music and noise signals. Instead of frontal localization, some of the high frequencies appeared to be behind or around the head.

Chapter 7

Conclusions, Discussion and Future Work

In this thesis, binaural techniques were investigated and their use in a loudspeaker comparison task was studied. A method for virtualized loudspeaker listening tests using the stereophonic listening setup was developed. Altogether eight test subjects participated in the formal listening tests which were conducted to find out differences between the virtual loudspeakers and the real loudspeakers.

In theory, binaural techniques have the potential to replace traditional listening tests, and improve the headphone listening experience by externalizing sound sources out of the head. Unfortunately, some unsolved problems related to the measurements and headphone reproduction remain.

Measurements at the entrance to an open ear canal were found repeatable only up to 6 kHz. Above 6 kHz, differences between the measurements are significant and clearly audible. The poor repeatability is annoying since it makes true-head responses unusable for the loudspeaker comparison task. Differences between two consecutive measurements are greater than differences between loudspeaker types.

Although the reproduction over headphones provides a perfect channel separation and thus the headphones are a natural choice for binaural reproduction, the problems related to headphone equalization are difficult to overcome. The use of intra-aural headphones instead of circumaural ones might ease the problems related to the headphone placement and repeatability, but at the same time the measurement procedure would become more difficult.

According to research done by Møller et al. [2][53][46], recording at the entrance to a closed ear canal could improve the repeatability of binaural measurements or at least diminish the high frequency resonances, which make the headphone equalization particularly hard. However, changing the recording position does not remove the problem related to headphone placement. The lack of headphones with ideal FEC properties lessens the usability of closed ear canal measurements.

The performance of the proposed method for loudspeaker evaluation was examined with a formal listening test. The test results and verbal comments from the test subjects confirm that binaural reproduction can be close to imperceptible from the reality if the audio source is conveniently chosen. Difference to reality increases if more demanding test signal is used. Some of the problems are definitely related to equalization at high frequencies, but according to author's own opinions, some of the problems are caused by lack of dynamic localization cues. This conclusion is supported by the listening test results. Directional properties of a moving source were found to be closer to reality than the same properties of a stationary source.

Since the difference between the virtual and real loudspeakers was found perceptible and the difference is dependent upon audio signal, virtual loudspeakers should not be compared to the real ones. Instead, all loudspeakers should be virtualized. Virtual loudspeakers are comparable to each other since the same processing is done to each loudspeaker pair. The difference between the virtual and real loudspeakers is perceptible or even slightly annoying, indicating that virtual listening tests cannot entirely replace traditional listening tests. Referring to Section 6.2, a limiting factor in the headphone reproduction is the tone color.

Now, the questions that were set in Chapter 1 can be answered. HATS responses can and they should be used together with true-head responses. The repeatability of true-head measurements alone is not good enough, but together with HATS measurements, the good sides of both techniques are combined. The tolerance between measurements is very small if imperceptible reproduction between the measurements is wanted. The required accuracy can be achieved with a HATS only. The repeatability of the headphone responses is rather poor at high frequencies and this must be taken into account when designing the equalizer. Finally, the inside-the-head localization can be defeated to some extent, but head tracking and dynamic localization cues are needed if input signal is randomly selected.

To improve the proposed method, the measurement position at the entrance to a open ear canal should be examined in more detail. Particularly, it would be interesting to know how the sound field changes when a microphone is moved around the ear canal entrance, and how the measurement microphone disturbs the sound field near the ear canal entrance. Measurements using a simplified ear model could be useful in the investigation. Obvious improvement would be a head tracker. Even a limited amount of dynamic localization cues would probably improve the listening experience remarkably. More measurements per loudspeaker pair would be needed if the head tracker was implemented. The headphone equalization is the most significant factor affecting the tone color. To improve the reproduction, headphones which could be placed exactly similarly every time are required.

In the future, advanced binaural techniques are needed in all kinds of virtual reality applications. It remains to see if the audio side of the virtual reality can ever reach true-to-life quality.

Bibliography

[1] E. Bruce Goldstein. *Sensation and Perception*. Wadsworth-Thomson Learning, 6th edition, 2002.

[2] Henrik Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4):171–218, 1992.

[3] Floyd E. Toole. Binaural record/reproduction systems and their use in psychoacoustic investigations. *the 91st Convention of the Audio Engineering Society (AES), preprint no. 3179*, October 1991.

[4] Toni Hirvonen. Headphone listening test methods. Master's thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2002.

[5] Sean E. Olive, Peter L. Schuck, Sharon L. Sally, and Marc E. Bonneville. The effects of loudspeaker placement on listener preference ratings. *Journal of the Audio Engineering Society*, 42(9):651–669, September 1994.

[6] Robert H. Gilkey and Timothy R. Anderson, editors. *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 28. Lawrence Erlbaum Associates, 1997.

[7] Witold Mickiewicz and Jerzy Sawicki. Headphone processor based on individualized head-related transfer functions measured in listening room. *the 116th Convention of the Audio Engineering Society (AES), preprint no. 6067*, May 2004.

[8] Victor B. Ganjian and Douglas Preis. Reproduction of loudspeaker listening room sound through headphones: Measured coherence analysis, cross spectra and digital filter impulse responses. *the 105th Convention of Audio Engineering Society (AES), preprint 4807*, 1998.

[9] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, revised edition, 1997.

[10] Jens Blauert, editor. *Communication Acoustics*. Springer, 2005.

[11] Sanjit K. Mitra. *Digital Signal Processing - A Computer-Based Approach*. McGraw-Hill, 2nd edition, 2001.

[12] Sanjit K. Mitra and James F. Kaiser, editors. *Handbook for Digital Signal Processing*. John Wiley & Sons Ltd, 1993.

[13] Thomas D. Rossing, Paul A. Wheeler, and F. Richard Moore. *The Science of Sound*. Addison Wesley, 3rd edition, 2002.

[14] E. Zwicker. Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33(2):248–248, February 1961.

[15] Matti Karjalainen. Kommunikaatioakustiikka. Technical Report 51, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 1998.

[16] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74 (3):750–753, September 1983.

[17] Brian C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 4th edition, 1997.

[18] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5:82–108, 1933.

[19] E. Zwicker. Ein verfahren zur berechnung der lautstärke. *Acustica*, 10:304–308, 1960.

[20] Alf Gabrielsson, Björn Hagerman, Tommy Bech-Kristensen, and Göran Lundberg. Perceived sound quality of reproductions with different frequency responses and sound levels. *Journal of the Acoustical Society of America*, 88(3):1359–1366, 1990.

[21] John Borwick, editor. *Loudspeaker and Headphone Handbook*, pages 493–574. Focal Press, 3rd edition, 2001.

[22] David M. Green. *An Introduction to Hearing*. John Wiley & Sons Ltd, 1976.

[23] Douglas Preis. Linear distortions: Measurement, methods and audible effects - a survey of existing knowledge. *Convention 2i of the Audio Engineering Society, preprint C1005*, 1984.

[24] Wikipedia contributors. *Group Delay and Phase Delay*. Wikipedia, The Free Encyclopedia, 2007. URL `http://en.wikipedia.org/w/index.php?title=Group_delay_and_phase_delay&oldid=141159633`.

[25] Eilif Bitsch Jensen and Henrik Møller. On the audibility of phase distortion in audio systems. *the 47th Convention of the Audio Engineering Society (AES), Copenhagen, Denmark*, March 1974.

[26] Jens Blauert and P. Laws. Group delay distortions in electroacoustical systems. *Journal of the Acoustical Society of America*, 63(5):1478–1483, 1978.

[27] J. A. Deer, P. J. Bloom, and Douglas Preis. Perception of phase distortion in all-pass filters. *Journal of the Audio Engineering Society*, 33(10), October 1985.

[28] Hideo Suzuki, Shigeru Morita, and Takeo Shindo. On the perception of phase distortion. *Journal of the Audio Engineering Society*, 28(9), 1980.

[29] Wolfgang Klippel. Tutorial: Loudspeaker nonlinearities - causes, parameters, symptoms. *Journal of the Audio Engineering Society*, 54(10):907–939, October 2006.

[30] Martin Colloms. *High Performance Loudspeakers*. Pentech Press Ltd, 1985.

[31] Earl R. Geddes and Lidia W. Lee. *Audio Transducers*. GedLee Associates, LLC, 2002.

[32] J. Moir. Doppler distortion in loudspeakers. *Wireless World*, page 65 et seq., April 1974.

[33] P. A. Fryer. Intermodulation distortion listening tests. *the 50th Convention of the Audio Engineering Society, preprint L-10*, 1975.

[34] Matti Karjalainen. A new auditory model for the evaluation of sound quality of audio systems. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85*, 1985.

[35] B. G. Haustein and W. Schirmer. Messeinrichtung zur untersuchung des richtungslokalisationsvermögens. *Hochfrequenztechnologie und Elektroakustik*, 79:96–101, 1970.

[36] Lord Rayleigh. On our perception of sound direction. *Philos Mag*, 13:214–232, 1907.

[37] M. Barron. *Auditorium Acoustics and Architectural Design*. London, Spon, 1993.

[38] Durand R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society*, 40(11):895–904, November 1992.

[39] H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *American Journal of Psychology*, 62:315–336, 1949.

[40] Heinrich Kuttruff. *Room Acoustics*. Spon Press, 4th edition, 2000.

[41] Don Davis and Eugene Patronis. *Sound Systems Engineering*. Focal Press, 3rd edition, 2006.

[42] ITU. *Recommendation BS.1116-1: Methods for the Subjective Assesment of Small Impairments in Audio Systems Including Multichannel Sound Systems*. International Telecommunications Union (ITU), 1997.

[43] IEC. *Sound system equipment - Part 13: Listening tests on loudspeakers*. International Electrotechnical Commission (IEC), 1985.

[44] Floyd E. Toole. Listening tests - turning opinion into fact. *Journal of the Audio Engineering Society*, 30(6), June 1982.

[45] Henrik Møller, Clemen Boje Jensen, Dorte Hammershøi, and Michael Friis Sørensen. Design criteria for headphones. *Journal of the Audio Engineering Society*, 43(4):218–232, April 1995.

[46] Henrik Møller, Michael Friis Sørensen, Clemen Boje Jensen, and Dorte Hammershøi. Binaural technique: Do we need individual recordings. *Journal of the Audio Engineering Society*, 44(6):451–469, June 1996.

[47] J. L. Bauck D. H. Cooper. Prospects for transaural recording. *Journal of the Audio Engineering Society*, 37(1), January 1989.

[48] Jyri Huopaniemi. *Virtual Acoustics and 3-D Sound in Multimedia Signal Processing*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 1999.

[49] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262, April 2002.

[50] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swepsine technique. *the 108th Convention of Audio Engineering Society (AES), preprint 5093*, 2000.

[51] 2007. URL `http://puredata.info/`.

[52] Klaus A J Riederer. *HRTF Analysis: Objective and Subjective Evaluation of Measured Head-Related Transfer Functions*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2005.

[53] Henrik Møller, Dorte Hammershøi, Clemen Boje Jensen, and Michael Friis Sørensen. Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society*, 43(4):203–217, April 1995.

[54] Matti Karjalainen, Poju Antsalo, Aki Mäkivirta, Timo Peltonen, and Vesa Välimäki. Estimation of modal decay parameters from noisy response measurements. *Journal of the Audio Engineering Society*, 50(11):867–878, November 2002.

[55] T. Paatero and Matti Karjalainen. Kautz filters and generalized frequency resolution: theory and audio applications. *Journal of the Audio Engineering Society*, 51(1-2):27–44, 2003.

[56] Matti Karjalainen and T. Paatero. Equalization of loudspeaker and room responses using kautz filters: Direct least squares design. *EURASIP Journal on Advances in Signal Processing*, 2007.

[57] T. Paatero. Kautz filters – matlab scripts and demos. URL `http://www.acoustics.hut.fi/software/kautz/kautz.htm`.

[58] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1975.

[59] Thomas F. Quatieri and A. V. Oppenheim. Iterative techniques for minimum phase signal reconstruction from phase or magnitude. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(6):1187–1193, December 1981.

[60] R. Bucklein. The audibility of frequency response irregularities. *Journal of the Audio Engineering Society*, 29(3):126–131, March 1981.

[61] Floyd E. Toole and Sean E. Olive. The modification of timbre by resonances: Perception and measurement. *Journal of the Audio Engineering Society*, 36(3):122–142, March 1988.

[62] P. Minnaar, S.K. Olesen, F. Christensen, and Henrik Møller. Localization with binaural recordings from artificial and human heads. *Journal of the Audio Engineering Society*, 49: 323–336, 2000.

[63] Henrik Møller, Dorte Hammershøi, C. B. Jensen, and M. F. Sørensen. Evaluation of artificial heads in listening tests. *Journal of the Audio Engineering Society*, 47:83–100, 1999.

[64] J. Kawaura, Y. Suzuki, F. Asano, and T. Sone. Sound localization in headphone reproduction by simulating transfer functions from the sound source to the external ear. *Journal of the Acoustical Society of Japan*, 12(5):203–216, 1991.

[65] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94:111–123, 1993.

[66] Andrew Rutherford. *Introducing Anova and Ancova: a GLM approach*, chapter 7. Sage Publications Ltd., 2000.

Appendix A

Written Instructions for the Test Subjects

A.1 General Description

The purpose of the test is to find out how well the spatial properties and the timbre of the virtual loudspeaker technique correspond with the reality and what is the difference between a true-head method and a method where head and torso simulator (HATS) is used with true-head responses to create individual binaural responses.

The difference is evaluated in ITU small impairments scale. It is important to note, that it is the difference to real loudspeakers not the fidelity of the headphone reproduction that is evaluated.

A.2 Scale

ITU small impairments scale is a continuous scale from 1 to 5. Verbal anchor points, shown in table A.1, are used.

Grade	Impairment	Ero
5	Imperceptible	Ei havaittavissa
4	Perceptible but not annoying	Havaittavissa, mutta ei häiritsevä
3	Slightly annoying	Hieman häiritsevä
2	Annoying	Häiritsevä
1	Very annoying	Erittäin häiritsevä

Table A.1: ITU small impairments scale.

The five attributes evaluated in the test are *Apparent Source Width*, *Direction of Events*, *Distance to Events*, *Spaciousness* and *Tone Color*. Following questions are related to the first attribute. How wide is the sound source or how wide are the sound sources in the case of multiple sources? How wide is the overall sound image? The listener should decide how well the width of the sound sources reproduced over headphones correspond with the loudspeaker reproduction. The second attribute, Direction of Events, describes the actual directions where the sound events appear to originate. The question is, do the directions in the headphone listening match with the directions in the loudspeaker reproduction. Distance to Events describes the actual distance from where the sound events appear to originate. The listener should decide if the distances in the headphone listening correspond with the distances in the loudspeaker listening. Spaciousness describes the amount of space present in the listening. The listener should decide how similar the perception of space in the headphone listening is compared to the reproduction over loudspeakers. Does the listener feel being in a similar space in both cases? The last attribute, Tone Color, describes the spectral content of the perceived audio sample. The listener should decide how much Tone Color differs from the loudspeaker reproduction.

Attribute	Related questions
Apparent Source Width	How wide is the sound source? Is the width well defined?
Direction of Events	From which directions the sound events originate?
Distance to Events	What is the actual distance to the sound events?
Spaciousness	How the listening space together with recorded space is perceived?
Tone Color	Describes the spectral content of the audio sample.

Table A.2: The list of attributes used in the listening test.

Appendix B

The ANOVA Table

Table B.1: The ANOVA Table

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
sample	90.2274	2	45.1137	266.6122	0.0
method	3.9466	1	3.9466	23.3236	1.7366e-06
attrib	7.7543	4	1.9386	11.4566	5.7029e-09
repet	0.081018	1	0.081018	0.4788	0.48923
speaker	0.53384	1	0.53384	3.1549	0.076204
subj	58.9291	7	8.4184	49.7511	0.0
sample*method	0.71889	2	0.35944	2.1242	0.12041
sample*attrib	12.7395	8	1.5924	9.4109	2.6985e-12
sample*repet	0.10874	2	0.05437	0.32131	0.72532
sample*speaker	0.72308	2	0.36154	2.1366	0.11894
sample*subj	43.1955	14	3.0854	18.234	0.0
method*attrib	0.84039	4	0.2101	1.2416	0.29201
method*repet	0.082316	1	0.082316	0.48647	0.48578
method*speaker	0.73415	1	0.73415	4.3387	0.037676
method*subj	2.9358	7	0.4194	2.4786	0.016294
attrib*repet	0.2901	4	0.072524	0.4286	0.78803
attrib*speaker	0.45237	4	0.11309	0.66836	0.61413
attrib*subj	41.7445	28	1.4909	8.8107	0.0
repet*speaker	0.20431	1	0.20431	1.2074	0.27228
repet*subj	3.9146	7	0.55922	3.3049	0.0018496
speaker*subj	2.9021	7	0.41459	2.4501	0.017514
sample*method*attrib	1.8972	8	0.23716	1.4015	0.19255
sample*method*repet	0.25038	2	0.12519	0.73984	0.47762
sample*method*speaker	2.5479	2	1.2739	7.5288	0.00058929
sample*method*subj	3.8825	14	0.27732	1.6389	0.064591
sample*attrib*repet	1.0023	8	0.12529	0.74041	0.65578
sample*attrib*speaker	0.72041	8	0.090052	0.53219	0.83259
sample*attrib*subj	33.1641	56	0.59222	3.4999	1.7097e-14

sample*repet*speaker	0.16377	2	0.081887	0.48393	0.61659
sample*repet*subj	3.0638	14	0.21884	1.2933	0.20603
sample*speaker*subj	6.3387	14	0.45276	2.6757	0.00081775
method*attrib*repet	0.29973	4	0.074932	0.44283	0.77766
method*attrib*speaker	0.46801	4	0.117	0.69146	0.59804
method*attrib*subj	3.442	28	0.12293	0.72647	0.84814
method*repet*speaker	0.0014278	1	0.0014278	0.0084378	0.92684
method*repet*subj	0.85831	7	0.12262	0.72463	0.65116
method*speaker*subj	3.6014	7	0.51448	3.0405	0.0037646
attrib*repet*speaker	0.25481	4	0.063704	0.37648	0.82549
attrib*repet*subj	6.2541	28	0.22336	1.32	0.12703
attrib*speaker*subj	6.7536	28	0.2412	1.4254	0.073599
repet*speaker*subj	0.77578	7	0.11083	0.65495	0.71031