AALBORG UNIVERSITY

Binaural auditory model for audio quality assessment

Master of Science in Engineering
Acoustics and Audio Technology

Master Thesis

Zuzanna Podwińska

AALBORG UNIVERSITY

Department of Electronic Systems
Frederik Bajers Vej 7B
9220 Aalborg Ø
Phone (+45)9940 8600
Fax (+45) 9940 9840
http://www.es.aau.dk/

Title:
Binaural auditory model
for audio quality assessment

Semester theme:
Master Thesis

Project period:
Master Program in Acoustics and Audio Technology
4th semester, Spring 2012

Project group:
Group 1060

Group members:
Zuzanna Podwińska

Supervisor:
Christian Sejer Pedersen

Secondary supervisor:
Woo-Keun Song, Brüel & Kjær

No. printed Copies: 5

No. of Pages: 57

No. of Appendix Pages: 5

Total no. of pages: 72

Completed:
May 31, 2012

Abstract:

Since audio quality is an area of growing concern for many users of audio equipment, so is the area of audio quality assessment for its manufacturers. This assessment can be done by listening tests, however, there is a need for a cheaper, yet equally accurate method. Most of the objective computational models, which predict audio quality, concentrate only on the monophonic perception, which might lead to underestimation of spatial degradations to audio.
The aim of the project was to develop - based on perceptual models which are already available - a binaural model of auditory perception, which can be used to assess audio quality degradation, in both its spatial, and non-spatial character. One monophonic model (CASP) and three different binaural processors were considered.
A listening test was conducted, in order to validate the combined objective models, as well as to adjust some of their parameters and optimise their predictions.
Finally, a combination of models and parameters, which seemed optimal, was chosen. However, if a truly optimal and easy to use model should be developed, there are still some areas which need further investigation.

Contents

1 Introduction and background **1**
- 1.1 Binaural hearing . . . 1
- 1.2 Audio quality assessment . . . 2
- 1.3 The aim of the project . . . 3

2 CASP model **5**
- 2.1 Introduction . . . 5
- 2.2 Outer- and middle-ear transformations . . . 7
- 2.3 DRNL filterbank . . . 8
- 2.4 Mechanical-to-neural transduction and adaptation . . . 11
- 2.5 Modulation filterbank . . . 12

3 Binaural models **15**
- 3.1 Introduction . . . 15
- 3.2 Lindemann . . . 16
- 3.3 Breebaart . . . 18
- 3.4 Dietz . . . 20
- 3.5 Combined models . . . 21
  - 3.5.1 CASP-L . . . 23
  - 3.5.2 CASP-B . . . 25
  - 3.5.3 CASP-D . . . 28
- 3.6 Summary . . . 29

4 Listening test **31**
- 4.1 Introduction . . . 31
- 4.2 Test method . . . 31
- 4.3 Experimental set-up . . . 33
- 4.4 Experimental procedure . . . 35
- 4.5 Results . . . 36
- 4.6 Comments from participants . . . 39
- 4.7 Summary . . . 39

5 Quality prediction with binaural models **41**
- 5.1 Introduction . . . 41
- 5.2 Sound samples pre-processing . . . 41
  - 5.2.1 Binaural room impulse response measurements . . . 42
  - 5.2.2 Gain adjustment . . . 43

5.2.3 Sound pressure at the blocked ear canal . . . 45
5.2.4 Ear canal transfer function . . . 45
5.3 Decision device for audio quality assessment . . . 46
5.4 Simulation results . . . 48
5.5 Summary . . . 52

6 Discussion and Conclusions **53**
6.1 Discussion . . . 53
6.1.1 Listening test . . . 53
6.1.2 Objective quality prediction . . . 54
6.1.3 Areas of potential future work . . . 55
6.2 Conclusions . . . 56

A Responses from the listening test **59**

B Enclosed DVD contents **63**

Bibliography **65**

CHAPTER 1

Introduction and background

Contents

1.1 Binaural hearing 1
1.2 Audio quality assessment 2
1.3 The aim of the project 3

Among the most important areas which meet within this project, are audio quality assessment, binaural hearing and perceptual modelling. A brief introduction to those is presented in the following two sections, in hope that it will not only give the necessary background on the topics, but also help to understand motivation behind this project. In section 1.3, the goal of the project is described.

1.1 Binaural hearing

Human auditory system is a complex system which transforms small changes in air pressure to mechanical vibrations and to neural impulses, which are analysed in the brain to construct an auditory "image" of the world around the listener. It is not within the scope of this project to give comprehensive description of ways in which hearing functions. A detailed description of the auditory system can be found in literature - e.g. Moore (2003).

To mention it very briefly, the acoustic pressure arriving at a human ear is transmitted through the ear canal to the eardrum, where it is transformed into vibrations of auditory ossicles (the middle ear). Those movements are transmitted to the cochlea, where the vibrations of the basilar membrane are translated into neural impulses, which, in turn, are then sent to the central nervous system.

Neural firings containing information about the sound reaching each ear are sent to the superior olivary complex in the brainstem, and although much of what is happening in the neural domain is not completely understood, it is believed that the signals from the left and the right ear are compared at this point, mainly by the means of cues such as interaural time or phase differences (ITD, IPD) and interaural level

differences (ILD).

The advantage of hearing with two ears, rather than one, can not be overstated. One obvious benefit is the ability to localize sound sources. Based on ITDs and ILDs between sound pressure reaching each ear, humans are able to localize sounds with precision up to 1-2 degrees for some directions. Those skills, complemented by vision, proved very useful in the course of evolution, when being able to hear from which direction a sound comes from – before the source was even noticed – might have quite literally been a matter of life and death.

However, sound localisation is not the only benefit which comes with the ability to hear binaurally. The fact that the brain is able to compare information obtained from each ear, improves some general aspects of hearing, such as signal detection in noise. It has been shown that if the same, but phase-shifted, sinusoidal signal is presented to each ear, while the masking noise remains the same for both ears, detection of the signal can be improved by as much as 15 dB compared to a monoaural presentation. This phenomenon is often referred to as binaural masking level differences (BMLD).

Another effect is that understanding of speech in noisy environments is enhanced while listening with two ears. Humans are able to concentrate on only one among many speakers present at the same time, as long as their locations are different.

1.2 Audio quality assessment

With the availability of high standard equipment increasing, and audio – in the form of music, radio, cinema – being a large part of people's lives today, it is natural that the interest in audio quality and means of assessing it is growing.

This assessment can be done in two ways. First of all, through listening tests. Those involve asking real subjects, trained or untrained, to give their opinions on the quality of a particular system under test. A listening experiment, although a direct way of finding perceived quality, is rather expensive and time-consuming. Moreover, many considerations need to be made when designing such a test, and knowledge from many different fields, such as psychophysics or statistics, is needed in order to design and analyse such an experiment correctly, and thus avoid basing conclusions on biased results. For a guide on how to conduct such a listening experiment on audio quality, the reader is referred to Bech and Zacharov (2006).

A different, cheaper and easier to use approach would therefore be desirable. Objective algorithms have been developed with the aim of computationally predicting the perceived audio quality of a system – one notable mention is the standardised PEAQ algorithm (ITU-R BS.1387-1, 2001). The most basic idea behind the majority those models is to compare a change in the signal, with respect to a reference, and relate that change to the impression of perceptual quality. However, most of the models which have been developed until now, concentrate on monophonic processing and do not take binaural perception into account.

This is seen in this project as an important area of potential improvement. As mentioned previously in this chapter, binaural hearing has a large significance for human sound perception. This also includes perceived audio quality. Some of the aspects of spatial audio quality, which have been identified, are source location, source width, source depth, envelpment, and others.

One model which tries to address those is Rumsey et al. (2008). Their approach is to find and extract those features from the audio signal, which correspond to certain perceptual impressions, associated with location, width and envelopment of a sound scene. In doing so, they do not intend to model the perceptual path itself in any way. It is more a model of the *effect*, than the *process*.

Gaining more and more knowledge about human auditory system, however, allows for creation of computational perceptual models, which aim at mimicking human sound perception. Those models can have different applications, one of them being audio quality assessment. Most of the binaural perceptual models so far have been made for other purposes, such as sound source localisation or signal detection (modelling BMLDs).

An interesting attempt to model the full, monophonic and binaural, auditory path and use it for assessment of codec audio, was made by Robinson (2002) in his PhD thesis. His work is, again, based on the idea of comparing a reference sound to a degraded sound (in his case, processed with a codec) to detect change in attributes, such as for example a shift in sound source location or stereo image width. A similar approach will be used in this project.

1.3 The aim of the project

The aim of the project is to attempt to develop, based on the knowledge that is already available (specifically, available perceptual models), a binaural model of auditory perception, which can be used to assess audio quality degradation, in both its spatial,

and non-spatial character. The model should take two audio signals as an input (test and reference), and process them both with the monophonic, as well as the monophonic+binaural parts. Outputs of those would be fed into a detector, which would give a prediction of perceived change in audio quality.

In chapter 2 on the facing page, a monophonic computational model of auditory perception is described. This model was used to obtain internal representations of each channel separately, as well as combined with binaural processors to obtain binaural information from the 2-channel signal.

In chapter 3 on page 15, three binaural models, considered for this project, are presented. A description of each binaural processor is given, and output of combining each with the CASP monaural part is discussed.

In chapter 4 on page 31, a listening test is described, which was conducted in order to validate the predictions obtained from combined models described in chapters 2 and 3.

In chapter 5 on page 41, the process of obtaining quality predictions from the models is described. Moreover, those predictions are compared to the subjective responses from the listening test, and the results are presented.

In chapter 6 on page 53, a discussion is given, concerning both the obtained results, as well as some other aspects of the project. Conclusions are also included in the chapter.

Additionally, **Appendix A** presents in more detail responses obtained from the listening test, and **Appendix B** lists the contents of the enclosed DVD.

CHAPTER 2

CASP model

Contents

2.1 Introduction 5
2.2 Outer- and middle-ear transformations 7
2.3 DRNL filterbank 8
2.4 Mechanical-to-neural transduction and adaptation 11
2.5 Modulation filterbank 12

2.1 Introduction

A monophonic model of computational auditory signal-processing and perception (CASP) was developed by Jepsen et al. (2008). It focuses on modelling perceptual masking phenomena, and was largely based on previous work by Dau et al. (1997a,b). Changes made to the original model by Dau et al. include a non-linear basilar membrane processing stage, as well as outer- and middle-ear transfer functions.

Overall structure of the model can be seen on figure 2.1.
In the following sections, stages of the model will be discussed in more detail.

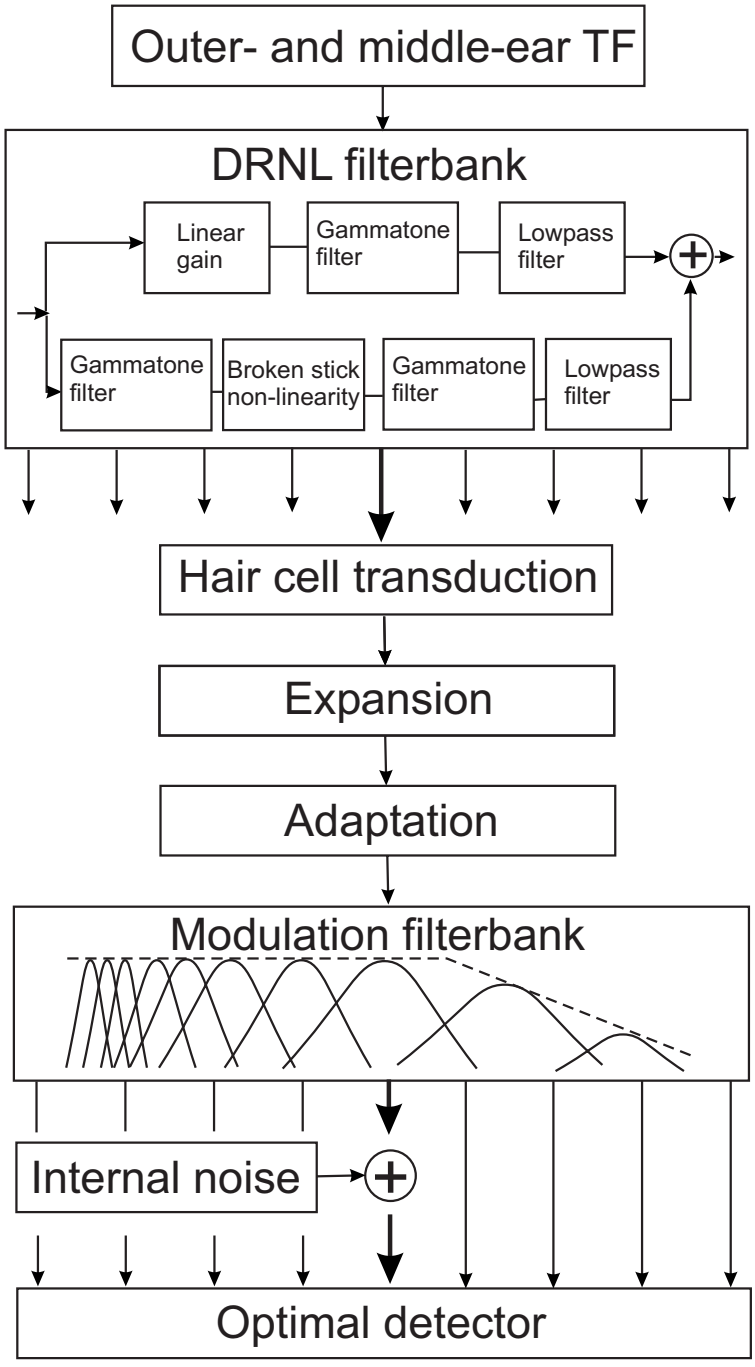


Figure 2.1: Structure of the CASP model, from Jepsen et al. (2008).

2.2 Outer- and middle-ear transformations

Firstly, he input to the model is scaled to be represented in pascal. Then, it is filtered with two transfer functions, to simulate the influence of the outer and middle ear. Those transfer functions are realized in Matlab by two linear phase FIR filters:

- the outer-ear filter is a headphone-to-eardrum transfer function for a specific pair of high quality headphones, which are circumaural, open and diffuse-field equalized (see figure 2.2);

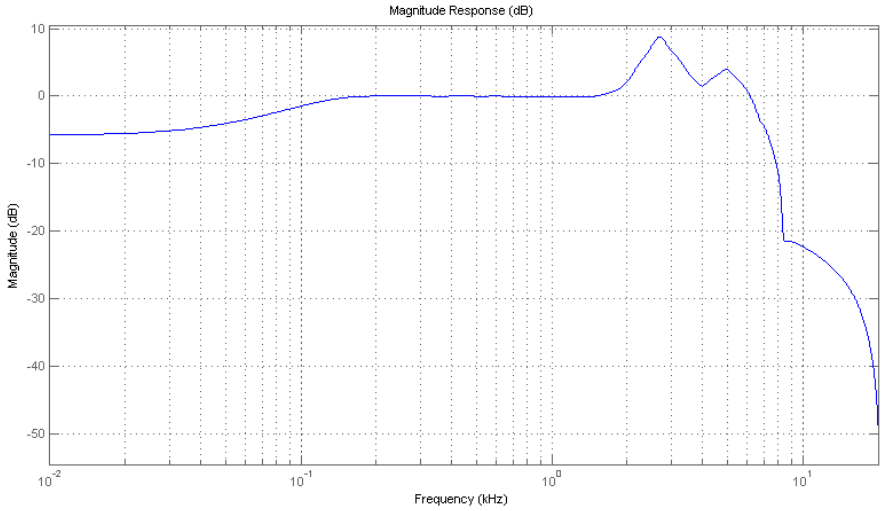


Figure 2.2: Outer ear transfer function, 512 taps FIR filter.

- the middle-ear filter was derived from human cadaver data (see figure 2.3).

The outer- and middle-ear transfer functions correspond to those described by Lopez-Poveda and Meddis (2001). Output of this stage represents peak velocity of vibration of the stapes.

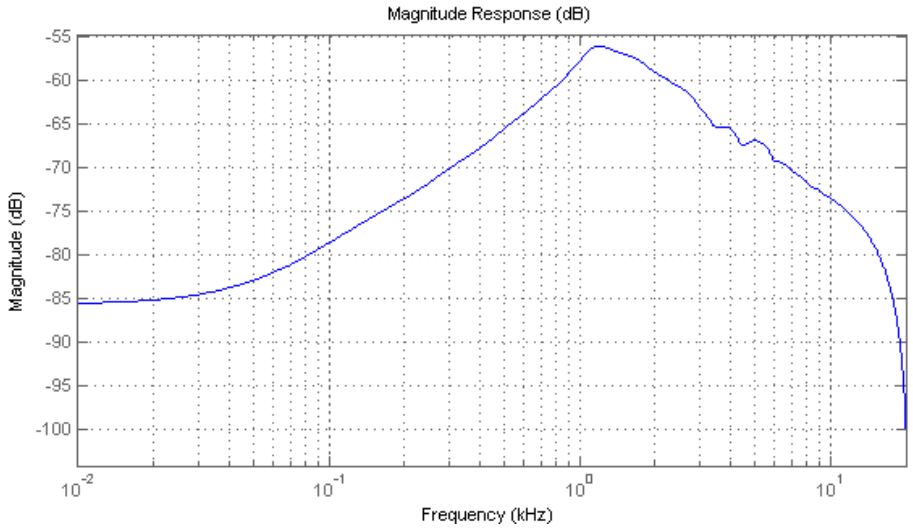


Figure 2.3: Outer ear transfer function, 512 taps order FIR filter.

2.3 DRNL filterbank

This part is taken (with some modifications) from the work of Lopez-Poveda and Meddis (2001), and is intended to simulate the properties of human cochlea (transmission of energy from stapes motion into basilar membrane vibration). The BM algorithm includes two parallel paths: a linear one, and a compressive nonlinear one, and its output is a sum of those two paths. The structure can be seen in figure 2.4 (numbers of cascade filters have been changed in the CASP model).

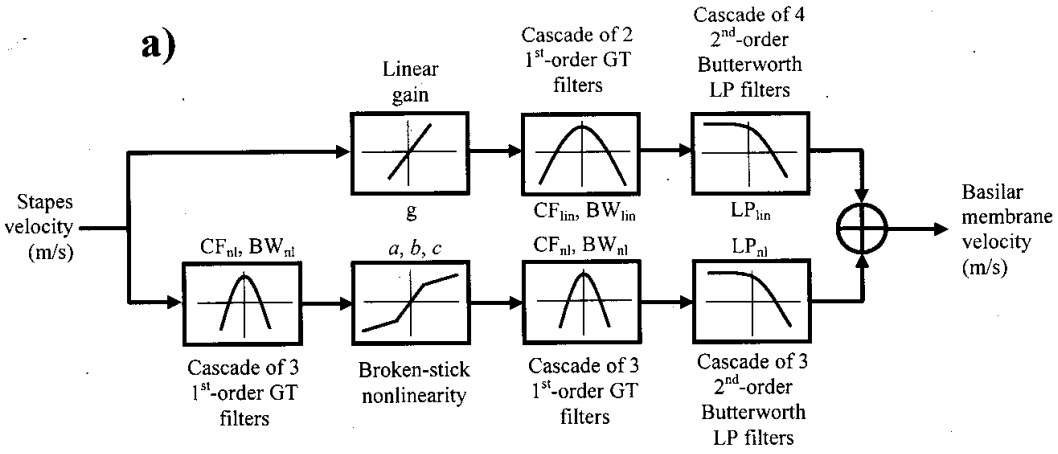


Figure 2.4: Structure of the DRNL filterbank, from Lopez-Poveda and Meddis (2001).

At low signal levels (below 30-40 dB SPL), the nonlinear part behaves linearly. At medium signal levels (40-70 dB SPL), the nonlinear part is compressive. At high signal levels (above 70-80 dB SPL), the output of the linear path dominates the sum. Parameters of the model were fitted to psychophysical data (Plack and Oxenham, 2000) to simulate the properties of human cochlea.

The model uses 60 separate and independent DRNL paths, each tuned to a different center frequency (CF). The 60 CFs are equidistantly spaced on the ERB scale, from 100 Hz up to 8 kHz. The signal obtained from the first stage of the model (outer- and middle-ear filtering) is fed to each of those parallel paths. The following steps are computed for each path (each CF).

In the **linear path**:

1. Linear gain

$$g = 10^{4.20405 - 0.47909 \log + 10\mathrm{CF}} \tag{2.1}$$

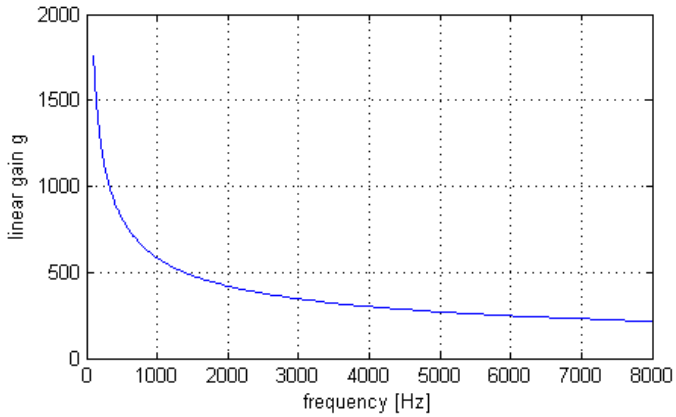


Figure 2.5: Linear gain.

2. Cascade of two gammatone filters, where CF_{lin} and BW_{lin}, the center frequency and the band width of the filters, are equal to:

$$CF_{lin} = 10^{-0.06762 + 1.01679 \log_{10} \mathrm{CF}} \tag{2.2}$$

$$BW_{lin} = 10^{0.03728 + 0.75 \log_{10} \mathrm{CF}} \tag{2.3}$$

3. Cascade of 4 low pass filters, where the filter cut-off frequency is:

$$LP_{lin} = 10^{-0.06762 + 1.01 \log_{10} \mathrm{CF}} \tag{2.4}$$

Figure 2.6 presents the two filters used for a linear path, gammatone and low pass, for CF = 1 kHz.

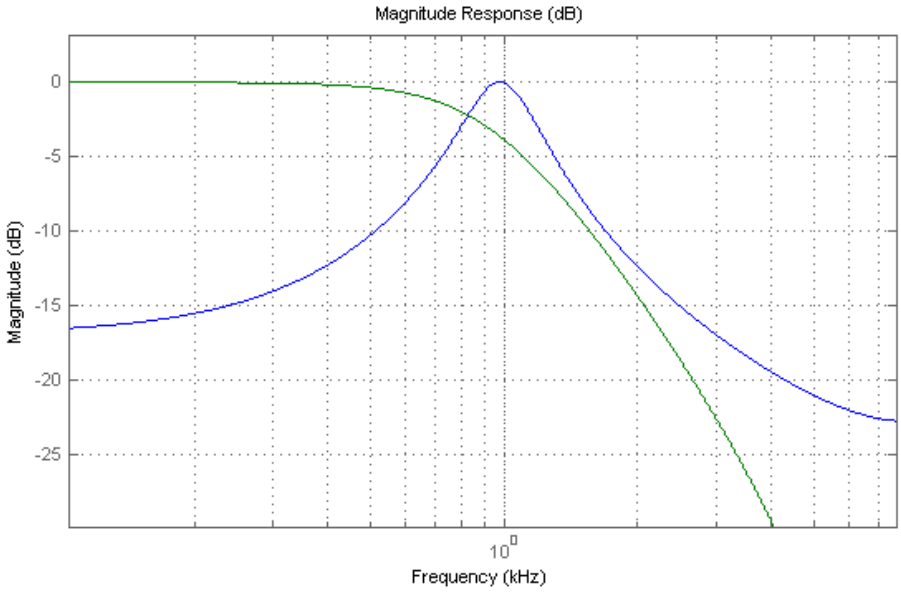


Figure 2.6: A gammatone (blue) and a low pass filter (green) for CF = 1 kHz.

In the **non-linear path**:

1. Cascade of 2 gammatone filters, where the center frequency and the band width of the filters are respectively:

$$CF_{nlin} = 10^{-0.05252+1.01650\log_{10}\mathrm{CF}} \tag{2.5}$$

$$BW_{nlin} = 10^{-0.03193+0.7\log_{10}\mathrm{CF}} \tag{2.6}$$

2. Non-linear gain function:

$$y(t) = \mathrm{sign}(x(t))\min(a|x(t)|, b|x(t)|^c) \tag{2.7}$$

where:

for CF $\leqslant$ 1500 Hz

$a = 10^{1.40298+0.81916\log_{10}\mathrm{CF}}$

$b = 10^{1.61912-0.81867\log_{10}\mathrm{CF}}$

for CF $>$ 1500 Hz

$a = 10^{1.40298+0.81916\log_{10}1500}$

$$b = 10^{1.61912-0.81867 \log_{10} 1500}$$

and

$$c = 10^{-0.60206}$$

3. One low pass filter, where:

$$LP_{nlin} = 10^{-0.05252+1.01 \log_{10} \mathrm{CF}} \tag{2.8}$$

Figure 2.7 illustrates an example of two filters used in the non-linear path (CF = 1 kHz).

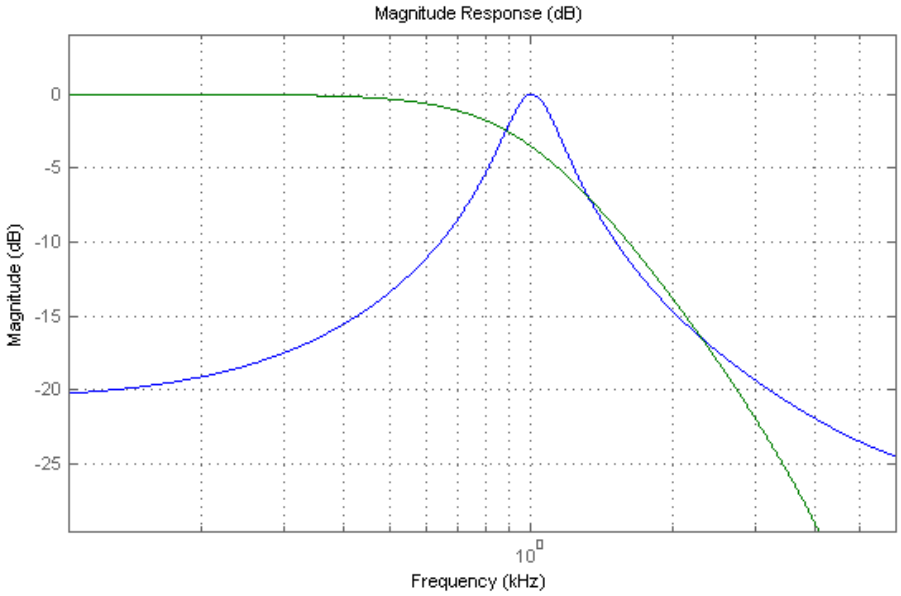


Figure 2.7: A gammatone (blue) and a low pass (green) filters used in the non-linear paths; CF = 1 kHz.

The output of this stage is a matrix of 60 frequency channels, containing filtered time signals. The output at this point corresponds to basilar membrane oscillations velocity. In the following stages, each channel will be processed independently.

2.4 Mechanical-to-neural transduction and adaptation

The hair-cell trasduction stage is roughly simulated in the model by half-wave rectification and a first order lowpass filter at 1kHz. Low pass filtering keeps the fine structure of the signal at low frequencies and extracts the envelope of the signal at high frequencies. Then, a squaring expansion is applied, and the lowest signal levels

are adjusted, depending on CF, according to a table of minimum values shown in figure 2.8.

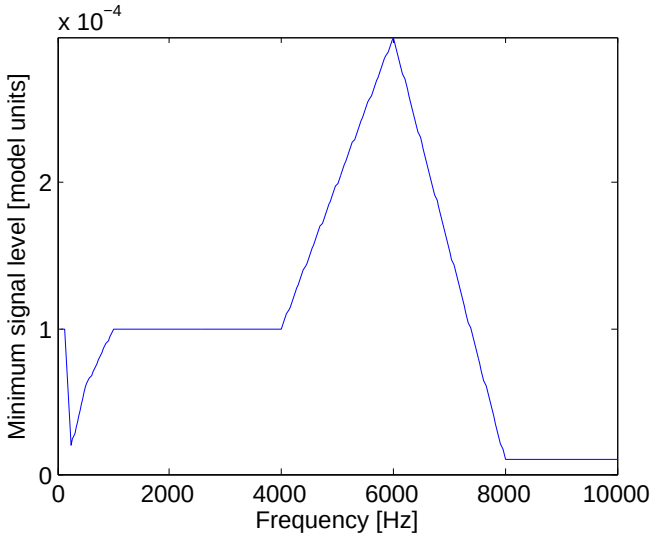


Figure 2.8: Minimum allowed signal values

Next step is the adaptation stage of the model, corresponding to changes in the gain of the system in response to changes in input level. In the model, it is realized by a chain of five feedback loops with different time constants. Each loop consists of a low pass filter and a division operation. The low pass filtered output is fed back to the denominator of the dividing element. The time constants, ranging between 5 and 500 ms, were chosen to account for perceptual forward-masking data. Maximum ratio of the onset response amplitude and steady-state response amplitude is set to be 10.

2.5 Modulation filterbank

In this part of the model, the signal is first low pass filtered at 150 Hz, which simulates a decreased sensitivity to modulation at lower modulation frequencies. Then, each channel is passed through a modulation filterbank. The lowest filter in the filterbank is a low pass filter with 2.5 Hz cut-off frequency. The highest modulation filter frequency is 1/4 of CF and not more than 1000 Hz. The modulation filters tuned to 5 and 10 Hz have a constant bandwidth of 5 Hz. Center frequencies of modulation filters above that are logarithmically scaled, their Q factor being always 2, and their transfer functions overlapping at -3 dB points.

The modulation filters are complex frequency-shifted first-order low pass filters. For filters above 10 Hz, the absolute value of the output is considered. For filters at and below 10 Hz, the real part of the output is considered. The output of the modulation filters above 10 Hz is attenuated by a factor of $\sqrt{2}$, to adjust the RMS value of all filters.

Example modulation filterbank, for CF = 1 kHz, is shown in figure 2.9.

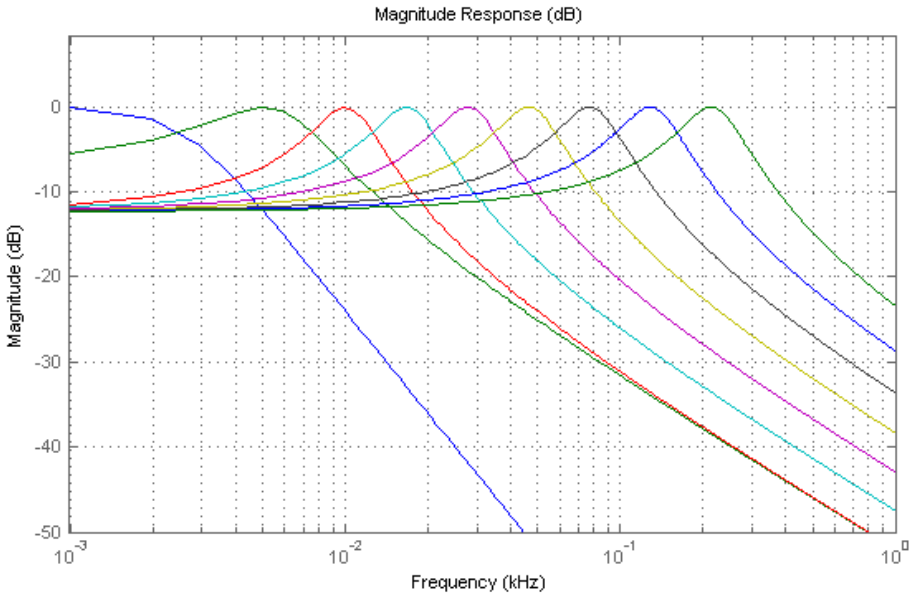


Figure 2.9: Moduation filterbank for CF = 1 kHz.

The output data at this stage is a 3D matrix, where one dimension corresponds to time, one to peripheral channels (60 CFs), and one to the modulation filters.

Figures 2.10 and 2.11 show 3 examples of output data obtained from the CASP model.

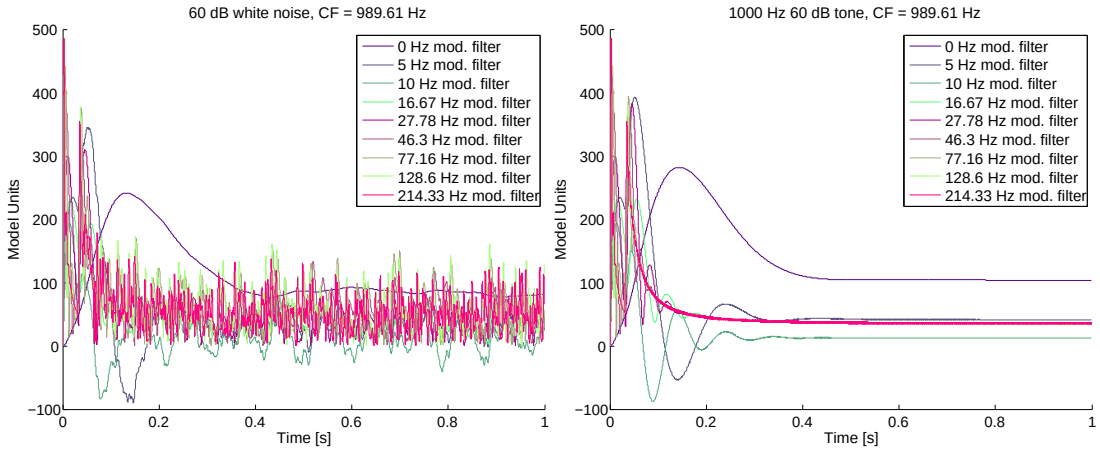


Figure 2.10: Results showing the output at (around) 1 kHz peripheral channel. Left: input to the model is 60 dB SPL RMS white noise; right: input to the model is 60 dB SPL 1 kHz tone. ('0 Hz mod filter' is actually a low pass filter with cut-off frequency of 2.5 Hz)

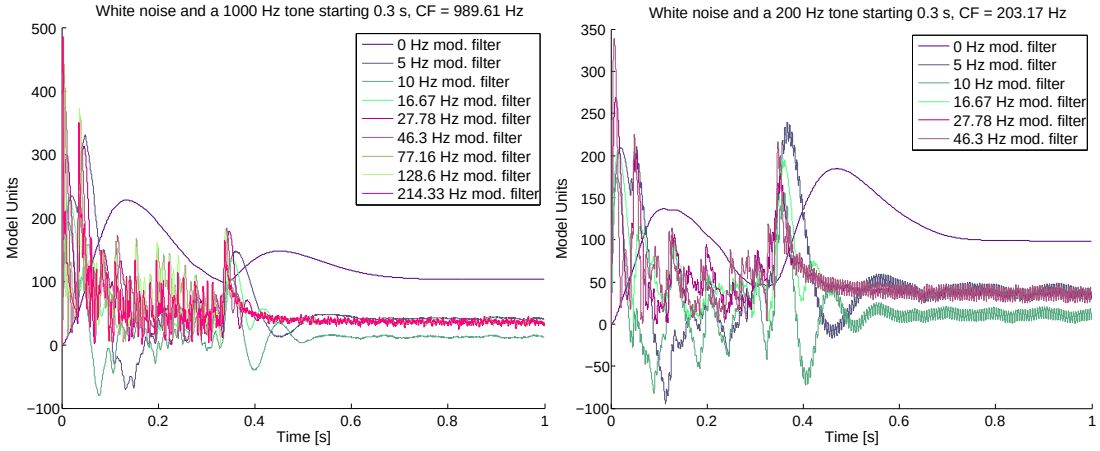


Figure 2.11: Results showing the output of the CASP model. Input to the model in both cases is 50 dB SPL RMS white noise, and a tone added at 0.3 s. Left: a 1 kHz 60 dB SPL tone, right: a 200 Hz 60 dB SPL tone. In both cases the peripheral channel corresponding to the input signal frequency is shown. ('0 Hz mod filter' is actually a low pass filter with cut-off frequency of 2.5 Hz)

CHAPTER 3

Binaural models

Contents

3.1 Introduction 15
3.2 Lindemann 16
3.3 Breebaart 18
3.4 Dietz 20
3.5 Combined models 21
3.5.1 CASP-L 23
3.5.2 CASP-B 25
3.5.3 CASP-D 28
3.6 Summary 29

3.1 Introduction

Three models have been chosen to be tested in an audio quality assessment task: models by Lindemann (1986), Breebaart et al. (2001) and Dietz et al. (2008). In the following chapter, those models are discussed. They are all available in the Auditory Modelling Toolbox (AMToolbox) for Matlab (Søndergaard et al., 2011), which is available to download from `http://amtoolbox.sourceforge.net/` under GNU General Public License. The models' implementations from AMToolbox were used in this project, without any modifications, and default parameter values were always used.

A more detailed description of each model is given in the following sections. It has to be noted, that originally, each of those models has its own peripheral processing stage, although they are all fairly similar in concept. In this project all those monaural stages will be replaced with corresponding parts of the CASP model described in chapter 2. Therefore only binaural parts of the models will be discussed here.

As mentioned before, in chapter 1, the idea behind using perceptual models for audio quality assessments is that they should be able to predict perceptual differences from a reference signal. Hence, the main task for the binaural models is to detect changes in spatial qualities of a sound, such a sound source position, source width, envelopment etc.

3.2 Lindemann

First model considered in the project is a lateralisation model by Lindemann (1986). It is based on the idea presented first by Jeffress (1948), which lied ground for many binaural processing models today.

Jeffress tries to explain sound localisation in the human auditory system by means of spatial summation of left- and right-ear signals reaching a "ladder" of tertiary fibers (see figure 3.1). Location of neural activity on the "ladder" is an indication of the interaural time difference, and thus, of the place on the horizontal plane, where the sound source is localised.

Lindemann built on that concept and extended it to include inhibition mechanisms and monaural detection. The model is based on two tap delay lines, one coming from each ear, going in opposite directions (see figure 3.2 on the next page), which is implemented as a running cross-correlation of the left- and right-ear signals. The monaural detection is designed to produce results in situations when signal at one ear is zero, and the binaural cross-correlation does not provide any localisation information. Inhibition, in turn, is introduced so that an offset of the first cross-correlation peak suppresses secondary peaks within a certain time interval. This allows for the model to make sure that delayed reflections do not contribute to sound localisation, thus taking care of the precedence effect.

Equations 3.1 and 3.2 illustrate the AMToolbox implementations of inhibition and monaural detection, correspondingly.

$$\begin{aligned} r(m+1,n+1) &= r(m,n) \cdot (1 - c_s l(m,n)) \\ l(m+1,n+1) &= l(m,n) \cdot (1 - c_s r(m,n)) \end{aligned} \tag{3.1}$$

$$\begin{aligned} R(m,n) &= r(m,n)[1 - w_l(m)] + w_l(m) \\ L(m,n) &= l(m,n)[1 - w_r(m)] + w_r(m) \\ w(m) &= w_f e^{-(m+M)/M_f} \end{aligned} \tag{3.2}$$

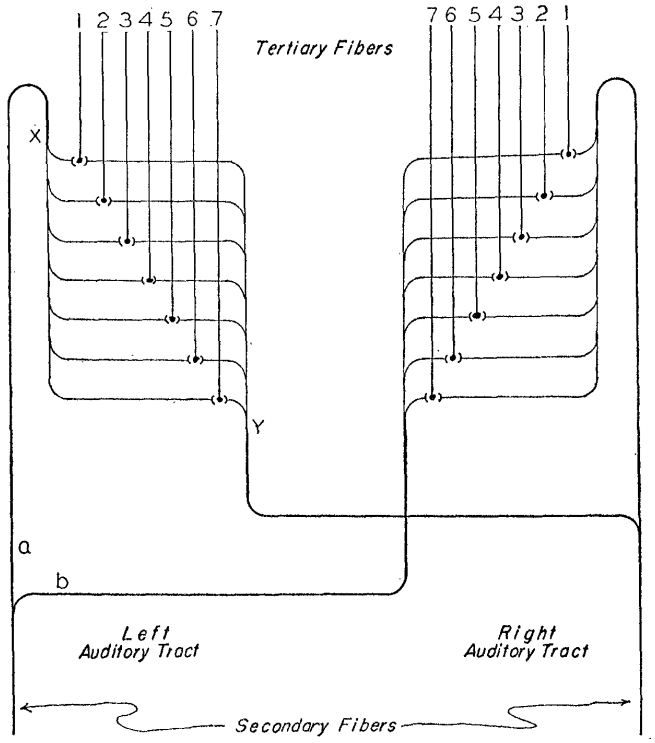


Figure 3.1: Localisation mechanism suggested by Jeffress, from Jeffress (1948).

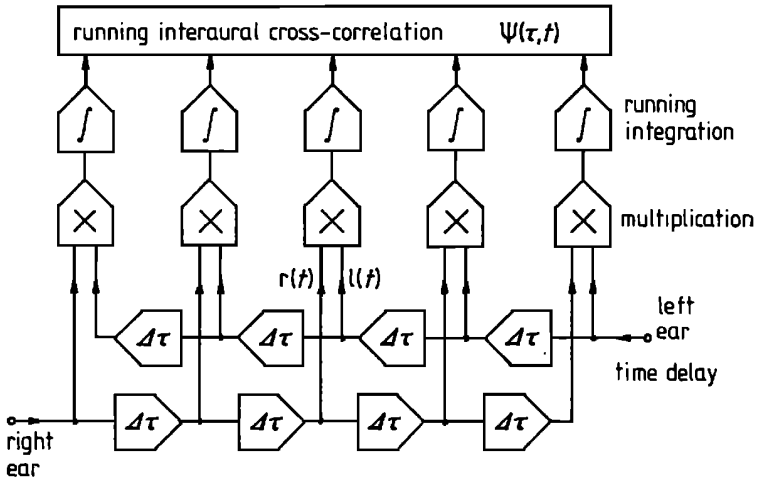


Figure 3.2: Basic structure of the Lindemann running cross-correlation model (without monaural detection and inhibition), from Lindemann (1986).

where r and l are the right and left signals going through the delay lines, m is a discrete tap delay step, n is the time sample number, c_s is stationary inhibition factor ($c_s = 0.3$), and w_f is the monaural sensitivity at the end of the delay line ($w_f = 0.035$).

Then, the running cross correlation is:

$$c = \sum_{n=N_1}^{N_2} R(m,n) \cdot L(m,n) \cdot e^{-(N_2-n)/T_{int}} \tag{3.3}$$

where R and L are the left and right signals, with monaural sensitivities and inhibition factors applied, and T_{int} is the integration time constant.

For a time-varying signal, the output of the model at a given time sample n is a vector of neural activity along the delay axis. Since this cross-correlation is performed for each peripheral frequency channel, and for each time sample, the output of the model is a 3-dimensional matrix: time vs. delay line vs. frequency channel.

3.3 Breebaart

Binaural model described in Breebaart et al. (2001) is based on equalisation-cancellation (EC) theory of binaural hearing (Durlach, 1963). It is hypothesised that first, in the equalisation step, signals coming from two ears are adjusted, so that the noise components are almost exactly the same in two ears (the process is not expected to be ideal). Then, in the cancellation part, the total signal in one ear is subtracted from the total signal in the other ear. Models based on this theory can account for BMLDs, as well as binaural pitch (see Breebaart et al. (2001) for examples). Breebaart notes, that in principle, his model could also be used to extract localisation information.

The principle of the Breebaart model is depicted in figure 3.3. Here, Jeffress model is extended with a chain of attenuation elements at each tap of the delay line. "EI" blocks are excitation-inhibition elements, corresponding to EI-type neurons in the lateral superior olive, which are excited by the ipsilateral, and inhibited by contralateral ear.

Each EI element corresponds to a certain characteristic interaural delay τ and a characteristic interaural attenuation α (which is expressed in dB). It is then possible to describe output from one element, excited by the left input and inhibited by the right one, as:

$$E_L(i,t,\tau,\alpha) = \left[10^{\alpha/40} L_i(t+\tau/2) - 10^{-\alpha/40} R_i(t-\tau/2)\right]^2 \tag{3.4}$$

and the opposite, excited by the right input and inhibited by the left, as:

$$E_R(i,t,\tau,\alpha) = \left[10^{-\alpha/40} R_i(t-\tau/2) - 10^{\alpha/40} L_i(t+\tau/2)\right]^2 \tag{3.5}$$

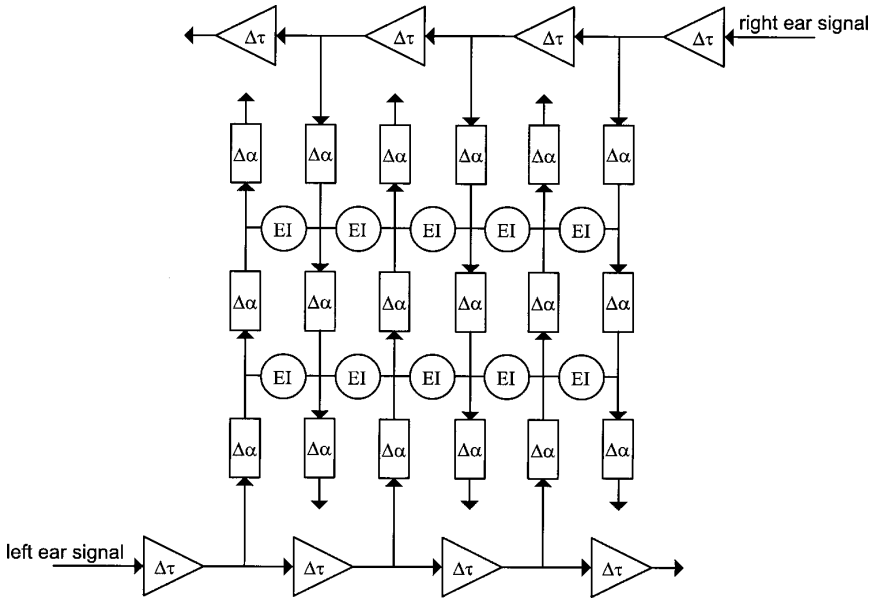


Figure 3.3: Structure of the Breebaart binaural processor, from Breebaart et al. (2001).

where $L_i(t)$ and $R_i(t)$ are the outputs of left and right channel peripheral processing at a time t, and i is the number of the frequency peripheral channel. $\lceil\rceil$ denotes half-wave rectification.

It can be shown that the summation of the output signal from equations 3.4 and 3.5, results in the following combined output E:

$$E(i,t,\tau,\alpha) = (10^{\alpha/40}L_i(t+\tau/2) - 10^{-\alpha/40}R_i(t-\tau/2))^2 \tag{3.6}$$

This output shows a minimum in its activity if the inputs' IID is the same as the characteristic IID of the EI element.

Then, in the model, a sliding temporal integrator is applied on the output signals to simulate limited temporal resolution of the system. The signal is weighted with a function $p(\tau)$, which decreases with τ, in order to account for the fact, that cells with larger interaural delays are less frequent than those with smaller delays. A compressive function simulating saturation effects of the EI cells is applied, and internal noise is added.

3.4 Dietz

The last binaural model used was Dietz et al. (2008), which is based on neural firing rate coding derived from the interaural phase difference.

After the peripheral stage, each frequency channel is filtered in parallel with 2 complex-valued band-pass gammatone filters: a fine-structure filter, tuned to the frequency of the peripheral channel, and a modulation filter, which extracts the envelope of the signal. The modulation filter is centered at 150 Hz for all peripheral channels.

The output of those filters is then a complex value:

$$g(t) = a(t) \cdot e^{i\phi(t)} \tag{3.7}$$

and the internal phase difference is determined from:

$$\mathrm{IPD} = \arg\left([\mathrm{ITF}]_{lp}\right) \tag{3.8}$$

where $[]_{lp}$ indicates low-pass filtering, and ITF - interaural transfer function - is:

$$\mathrm{ITF}(t) = g_l(t) \cdot \overline{g_r(t)} = a_l(t) \cdot a_r(t) \cdot e^{\phi_l(t) - \phi_r(t)} \tag{3.9}$$

where $\overline{g_r(t)}$ is the complex conjugate of $g_r(t)$.

IPD, then, represents change in phase between the left and right signal. The low-pass filter in equation 3.8 is employed to simulate a finite temporal resolution (smoothing of the signal in time).

In the Dietz model, based on IPD, the firing rate of neurons can be described by:

$$l(t) \propto sin(IPD(t)) \tag{3.10}$$

where $l(t) < 0$ denotes left, and $l(t) > 0$ - right lateralisation.

Additionally, as in Dietz et al. (2011), the interaural level difference (ILD) was derived from the energy ratio of the left and right signals, both filtered with a 30 Hz low-pass modulation filter:

$$\mathrm{ILD}(t) = \frac{20}{c} \cdot \log\left(\frac{h_r(t)}{h_l(t)}\right) \tag{3.11}$$

where h_r and h_l are the corresponding left and right low pass filtered signals. Compression factor c was 0.4, as default in AMToolbox.

The general structure of the model is summed up in figure 3.4 on the facing page.

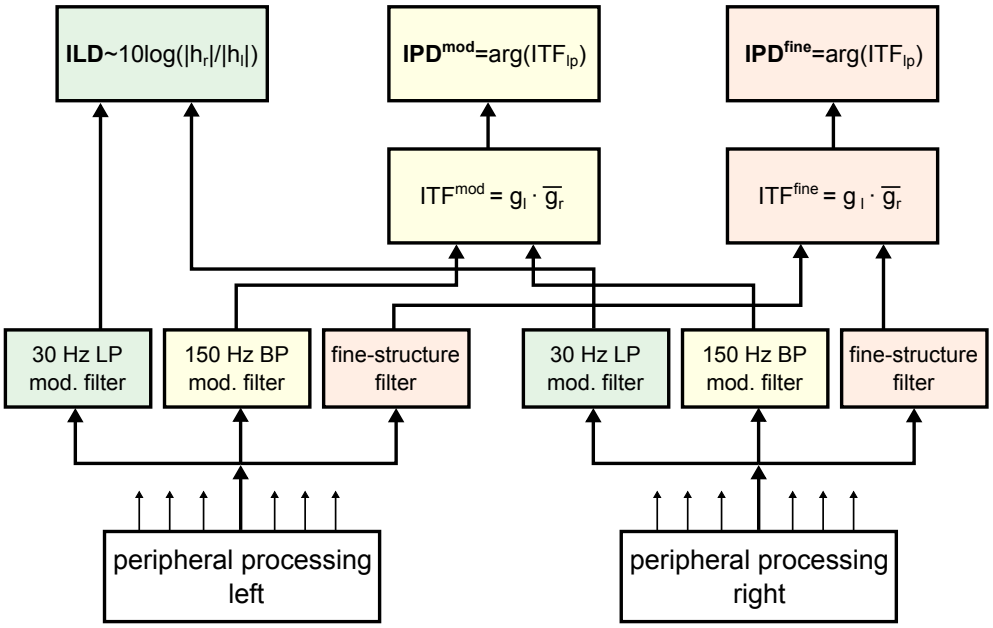


Figure 3.4: Basic structure of the Dietz model used in the project, adapted from Dietz et al. (2011).

3.5 Combined models

Binaural parts of the models described above were combined with the monaural CASP model. To ensure that no fundamental assumptions of the binaural parts are violated, only the parts of CASP up to the point which corresponds to the output of the original monaural stage were used. A comparison of the monaural parts, as well as the CASP model, is presented in table 3.1.

Thus, the models under test in the project from now on are:

- **CASP-L** (CASP + Lindemann): outer and middle ear filtering → DRNL filterbank → half wave rectification and low pass filter → running cross-correlation with inhibition and monaural detection;
- **CASP-B** (CASP + Breebaart): outer and middle ear filtering → DRNL filterbank → half wave rectification and low pass filter → adaptation loops → EI cell binaural model;
- **CASP-D** (CASP + Dietz): outer and middle ear filtering → DRNL filterbank → half wave rectification and low pass filter → IPD rate coding.

	Breebaart (2001)	**Lindemann (1986)**	**Dietz (2011)**	**CASP**
Monaural	outer and middle ear filtering as a band pass filter 1-4 kHz		middle ear filtering (500-2000 Hz 1st order band-pass)	outer and middle ear filtering
	gammatone filter bank with 1-ERB spaced filters	ERB filterbank (36 filters)	gammatone filterbank between 200 and 5000 Hz, 23 filters with 1 ERB spacing	DRNL filterbank (incl. cochlear compression)
			power-law compression with an exponent of 0.4 (cochlear compression)	
	half wave rectification followed by low-pass filtering to 770 Hz (5th order)	first order low pass filter at 800 Hz and haf-wave rectification	half wave rectification followed by filtering with a 770 Hz 5th order low pass filter	half wave rectification and low pass filter at 1 kHz
	cascade of 5 adaptation loops			cascade of 5 adaptation loops
				modulation filterbank
Binaural	an excitation-inhibition (EI) cell model	cross-correlation between the left and right channel	IPD rate coding, including fine-structure and modulation filters	

Table 3.1: A comparison of the 3 binaural models and the CASP model.

Example outputs of those three combined models will be presented in this section. Two types of simple stimuli were created. They were based on a 500 Hz tone (with a hann window), 2 s long, with 1 s of silence before and 1 s of silence after the tone. The test signals were stereo signals with that same tone in both channels, where:

- in the first signal, the tone in the left channel was presented 0.5 ms before the right one,
- in the second signal, the tone in the left channel was presented at a level 6 dB higher than the right one.

In both of these cases, the sound source should be perceived as being located on the left side of the listener. Additionally, corresponding signals with the sound source on the right were created, as well as a reference with the exact same signal coming from both channels.

3.5.1 CASP-L

As mentioned in section 3.2, output of the CASP-L model, based on Lindemann binaural processor, is a 3-dimensional matrix. Figure 3.5 illustrates one frequency channel from the output of this model, when the input is a 2-channel signal with the exact same 500 Hz tone in both channels. Figure 3.6, in turn, shows the output only for one frequency channel and one chosen time frame (corresponding to around 1.5 s). The inputs in this case are three different signals, one of them with no interaural difference, one with ITD and one with ILD introduced. Taking into account the actual interchannel attenuation introduced, it seems that the model responds to level changes worse than it does to time delay.

Lindemann suggests two ways of determinig the sound lateralisation from this output: the location fo the centroid along the delay line, or the location of the maxima of the inhibited cross-correlation function. Figure 3.7 shows the calculated centroid over time for the same signals as before. For the purpose of this project, a relevant problem is to find the most accurate representation of *change* in sound lateralisation, for more than one sound source.

Since the first method, centroid, is meant for one sound source only, it is not expected to give good predictions in the real-life-situation of detecting change in audio spatiality. However, if a change is big enough, e.g. almost all the signal moves to the left or right side, a centroid could perhaps also prove useful. In the case of detecting the maxima, the challenge is in identifying relevant peaks and their displacement.

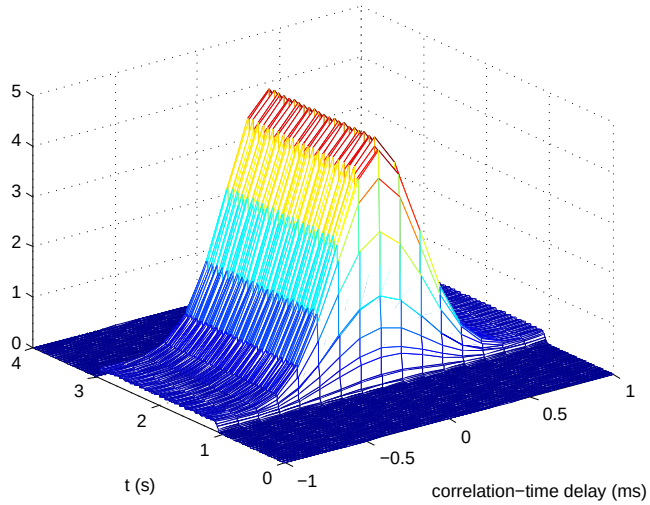


Figure 3.5: A reference signal output from the CASP-L model. The signal is exactly the same 500 Hz tone played through both channels. The figure only shows the frequency channel corresponding to the stimuli.

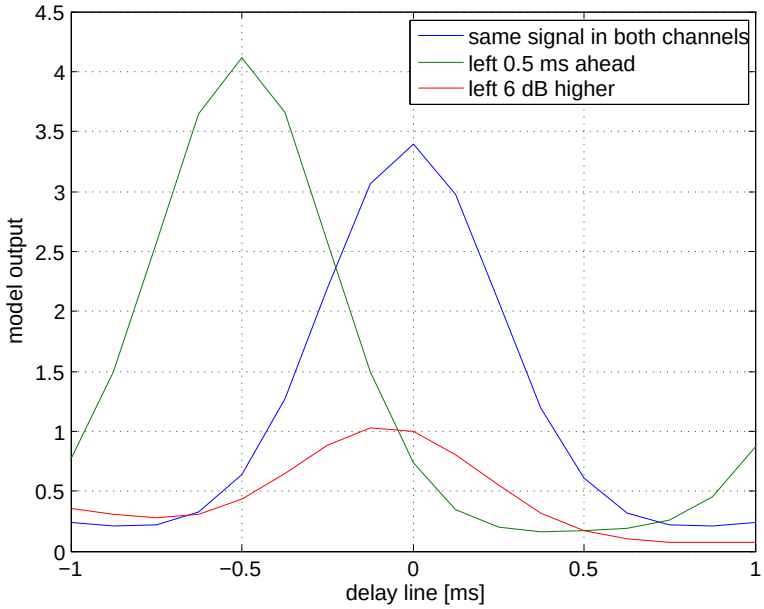


Figure 3.6: Output of the CASP-L model. The signal in all three cases is a 500 Hz tone, and the figure only shows the frequency channel corresponding to the stimuli, and a time frame corresponding to 1.5 s.

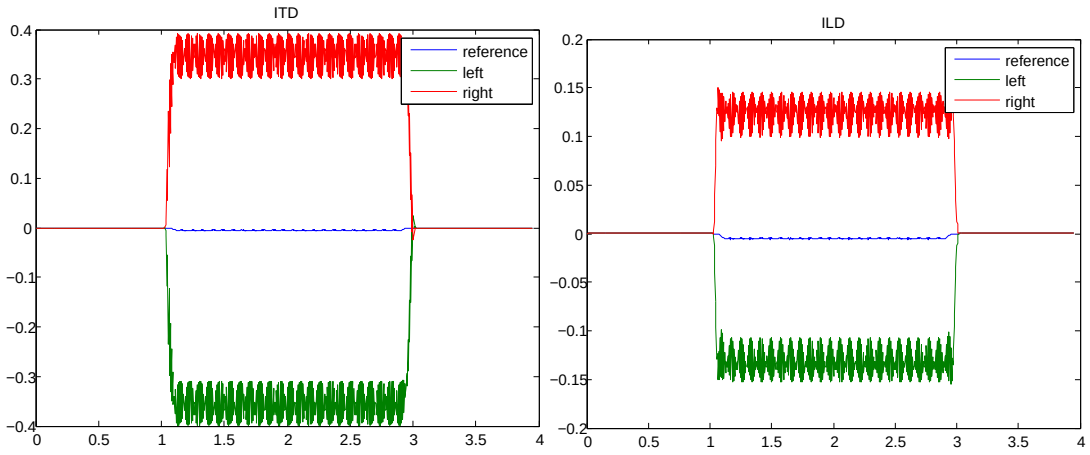


Figure 3.7: Centroid of the CASP-Lindemann correlation shown over time, for 500 Hz tone signals, with the source on the left (green) or right (red), and for a mono signal (blue). Left: for interaural time differences, right: for interaural level differences.

3.5.2 CASP-B

The Breebaart binaural model can be calculated for any specified characteristic time or intensity difference. Figure 3.8 on the next page illustrates the output of the model for three test signals. Only the output corresponding to time t = 2 s is shown.

To see more clearly, what impact ITDs and ILDs have on the output of the model, let us look at figures 3.9 on page 27 and 3.10 on page 27. The first only plots the characteristic interaural time difference for one chosen $\alpha = 0$, for a test signal with no ILD or ITD, and a test signal with the left channel 0.5 ms earlier. Figure 3.10, in turn, shows the output of the model as a function of α, for $\tau = 0$, for 4 test signals with different ILDs introduced.
It can be seen, that the minimum of the output function indicates the sound location. A change in ITD is clearly visible in the output of the model, as it produces such shift of the output, which corresponds exactly to the ITD change (see 3.9). Impact of introducing ILD on the output of the model is visible, however, it does not directly reflect the level difference between the channels.

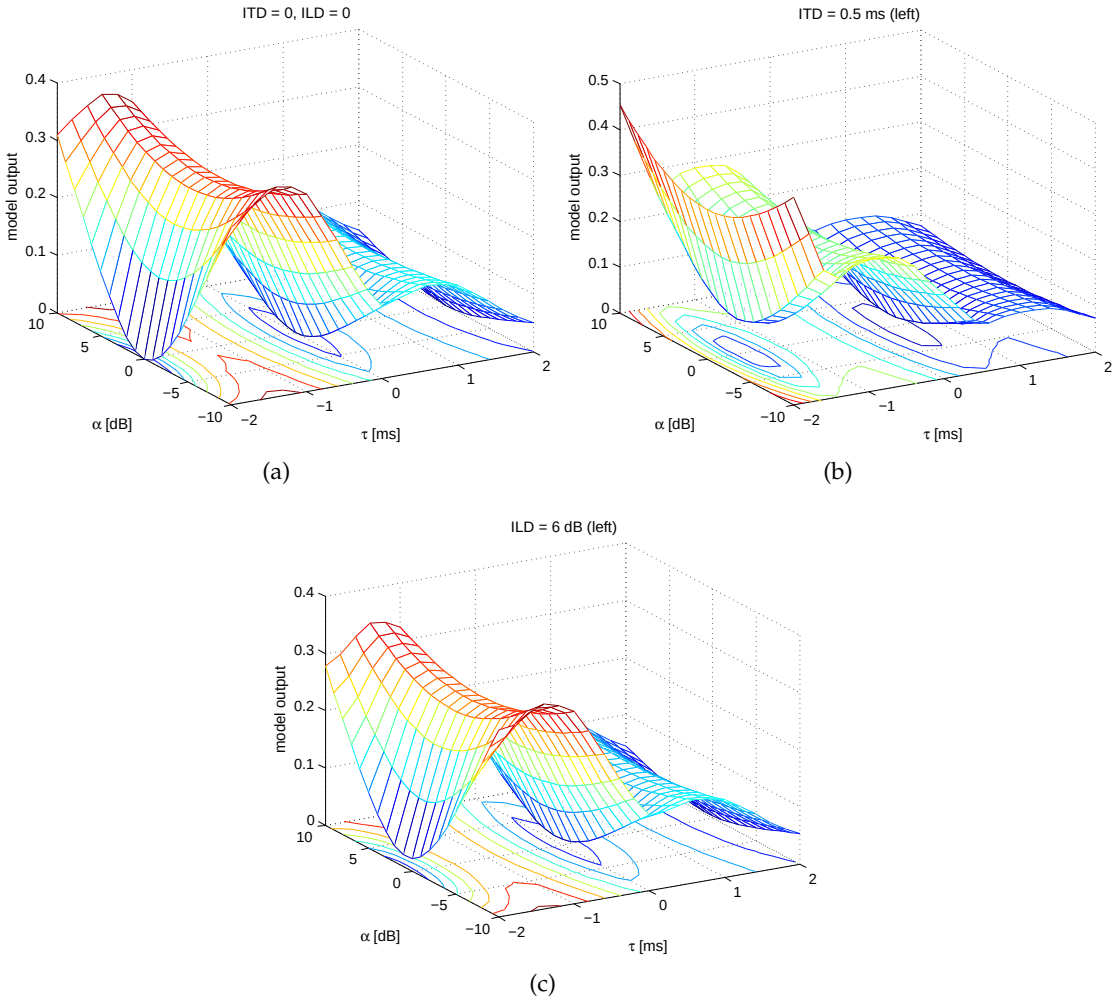


Figure 3.8: Output of the CASP-B model, for three different tone signals, (a) – a 2-channel signal with an 500 Hz tone, no ILD or ITD; (b) – a 2 channel signal with an 500 Hz tone, 0.5 ms earlier in the left channel; (c) – a 2 channel signal with an 500 Hz tone, the left channel 6 dB higher. τ corresponds to the characteristic interaural time delay, and α is the characteristic interaural attenuation. A clear shift in the output across τ, in response to introducing ITD, is visible (from (a) to (b)). Change in output as a result of introducing ILD (from (a) to (c)) is less obvious, but can also be noticed.

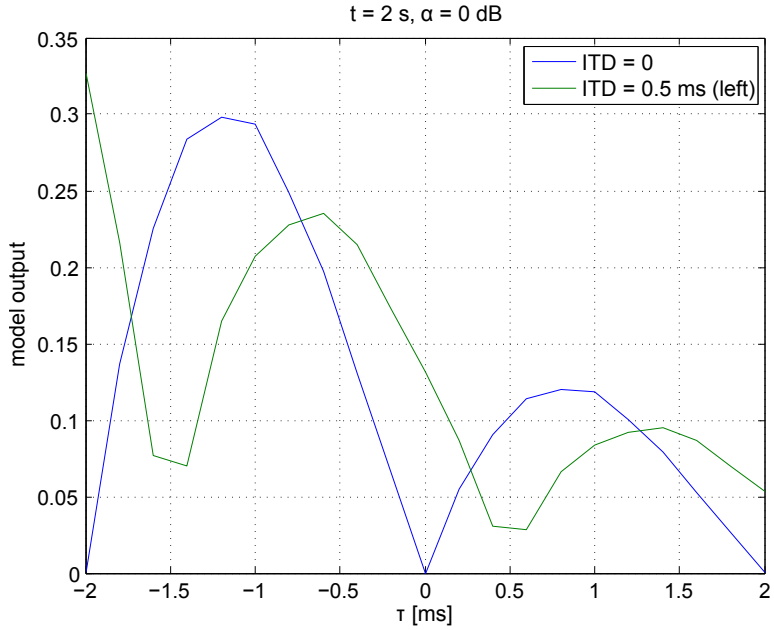


Figure 3.9: Output of the CASP-B model, for $\tau = 0$, and at a time $t = 2$ s. Input signals are a mono 2-channel 500 Hz tone (blue) and the same signal with ITD of +0.5 ms introduced to the left channel (green).

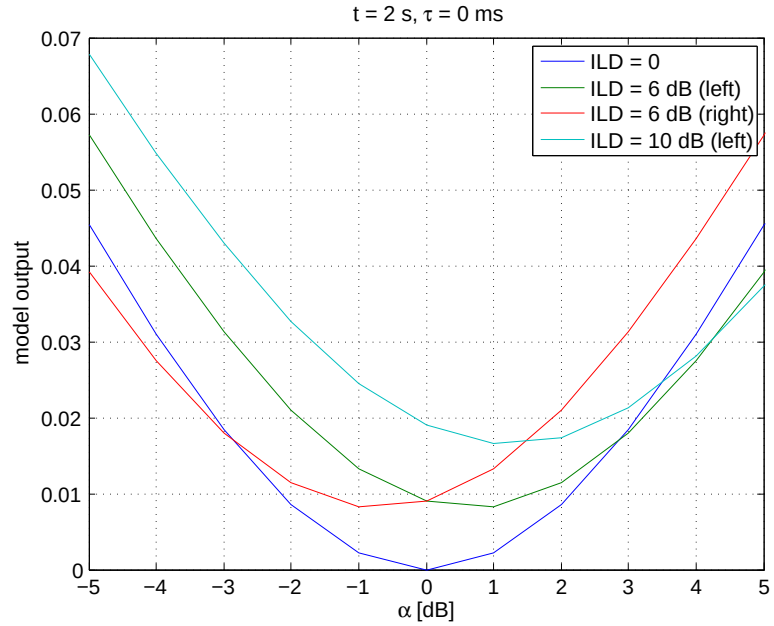


Figure 3.10: Output of the CASP-B model, for $\alpha = 0$, and at a time $t = 2$ s.

3.5.3 CASP-D

Unlike in the case of the two models described above, output of the Dietz interaural phase rate coding, for a specified frequency channel and in a given time t, is not a vector, but a single number. However, there are three potentially useful indicators calculated by the model: IPD from a fine structure filter, IPD from a modulation filter, and ILD. Figures 3.11 and 3.12 show outputs for two test signals.

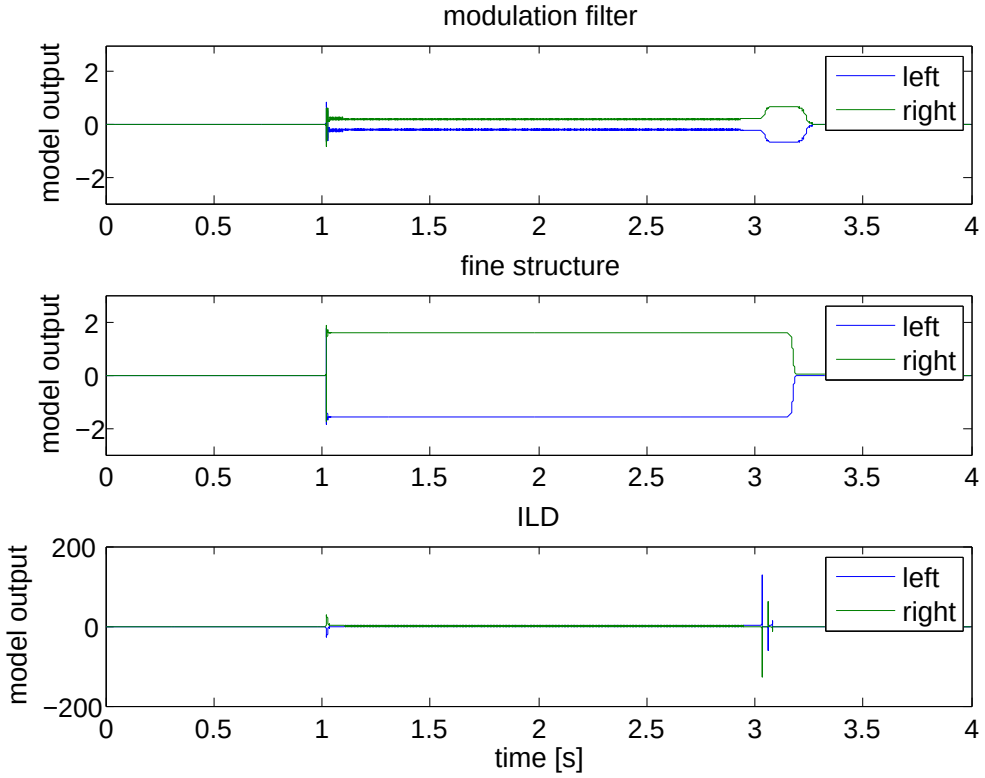


Figure 3.11: IPD output of the Dietz model over time, for ITD introduced for the left and right channel. The test signal is a 500 Hz tone which starts at 1 s at ends at 3 s.

As expected, IPD and ILD outputs of the model cope well with detecting interaural time and level differences, respectively. The fine-structure filter output and ILD output combination looks particularly promising.

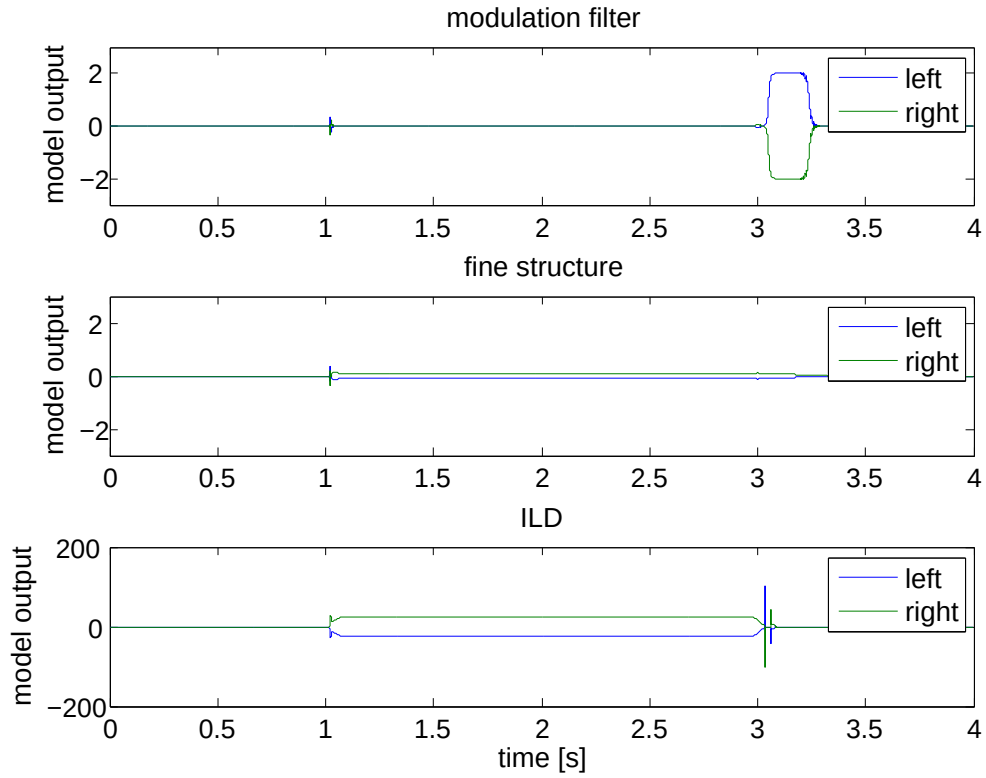


Figure 3.12: ILD output of the Dietz model over time, for ILD introduced for the left and right channel. The test signal is a 500 Hz tone which starts at 1 s at ends at 3 s.

3.6 Summary

In this chapter, binaural models used in the project were described in sections 3.2-3.4. Those models were combined with monaural CASP processing. In section 3.5, examples of the output of the combined models were presented and briefly discussed.

CHAPTER 4

Listening test

Contents

4.1 Introduction 31
4.2 Test method 31
4.3 Experimental set-up 33
4.4 Experimental procedure 35
4.5 Results 36
4.6 Comments from participants 39
4.7 Summary 39

4.1 Introduction

Despite a lot of investigative effort which has put into developing objective quality assessment tools (for various applications), a listening test is, and most likely will always remain the most reliable way of gaining information about human preference. Therefore, in this project, such an experiment was designed, in hope that its results will allow to adjust and test an objective audio quality assessment algorithm.

4.2 Test method

Test method used in the experiment was based on the double-blind multi-stimulus test method with hidden reference and hidden anchor, referred to as MUSHRA in recommendation ITU-R BS.1534-1 (2003). The participants' task was to rate the quality of short music excerpts compared to a reference, the quality of which was assumed to be the highest possible, *ideal* quality.

Keeping in mind that the results of the experiment should contribute to the development of a binaural audio quality model, degradations under test were divided into two groups, representing their two basic types: spatial impairments of the multichannel system, and other perceptual degradations, which do not have direct connection

to the spatial impression.

In order to make the task easier for the subjects, each of those two groups was tested on a separate user interface "page". A screenshot of the user interface (written in Matlab) can be seen in figure 4.1 on the facing page. On each page, there was a hidden reference and 7 different types of degradations, one of which was a hidden anchor, intended to represent the worst quality of all the samples. Recommendation ITU-R BS.1534-1 (2003) instructs to use a 3.5 kHz low-pass filtered sample as the low anchor, and such was chosen for the non-spatial perceptual degradations. For the spatial degradations, however, it did not seem to be appropriate. Instead, a mono signal from the left surround loudspeaker was chosen. This was used in Conetta et al. (2008), and the selection was "based upon the results of informal listening undertaken by the first author" of the paper.

Degradations used in the experiment were as follows (the numbers used here will be consistently used for those degradations throughout the report):

Non-spatial degradations:

1. hard limiter at -15 dB maximum signal level,
2. high pass filter at 500 Hz,
3. correlated noise in all channels (the same noise sample in all channels),
4. uncorrelated noise in all channels,
5. hard limiter at -20 dB maximum signal level,
6. mp3 codec, 64 kbps for each channel,
7. low pass filter at 3.5 kHz (*low anchor*),

Spatial degradations:

8. downmix to stereo (channels L and R)
9. downmix to mono – channel C
10. channel L moved 30° to the left
11. channel order moved to the right (L becomes C, C becomes R, R becomes SR etc.)
12. downmix to mono – all channels
13. channels L, R and C 6 dB lower than the rest
14. downmix to mono – channel LS

The scales used were visual analog scales, divided into five equal intervals, with the following labels: *Bad*, *Poor*, *Fair*, *Good* and *Excellent*. It was possible to choose any point on the scale (from 0 to 100, with 4th decimal precision).

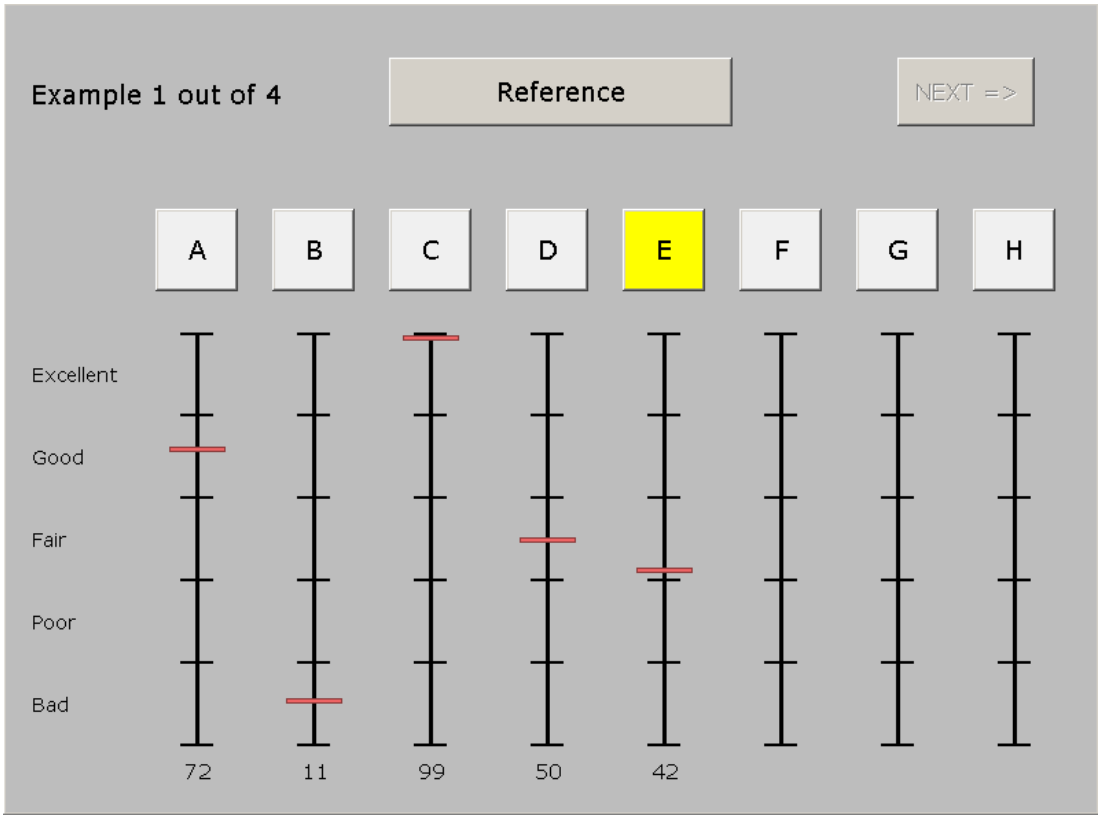


Figure 4.1: Graphical user interface

Two different excerpts of commercially available multichannel music recordings were chosen. Sound 1 was 8.4 s long, and sound 2 was 5.2 s long. Both were taken from the same DVD Audio disc (Steely Dan "Everything Must Go", a jazz-rock record), and were cut so that a full musical phrase would be included.

4.3 Experimental set-up

The experiment was conducted in a multi-channel listening room, which conforms to the recommendation ITU-R BS775-1 for multichannel/surround setups. The setup consisted of 6 loudspeakers, placed in positions indicated in figure 4.2. Five of the

loudspeakers (L, R, C, LS, RS) were positioned in accordance with ITU-R BS.775-2 (2006), and the additional loudspeaker (A) was placed 60 degrees to the left (30 degree misplacement for the left loudspeaker). They were all placed at 1.25 m hight, which was approximately the hight of the listener's ears, and at a distance of 2.5 m from the listener.

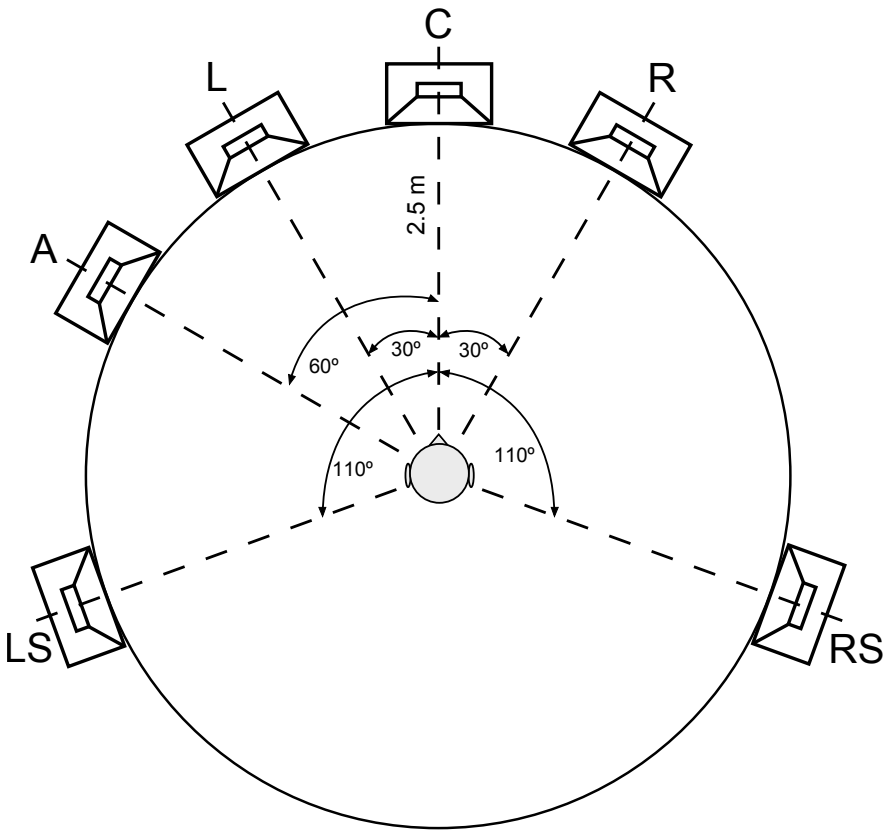


Figure 4.2: Loudspeaker setup for the listening experiment

The loudspeakers were active Genelec 1031A Bi-amplified Monitoring System. Their anechoic frequency characteristic is relatively flat from 50 to about 20 000 Hz (see: Gen), and a decision was made not to apply any equalisation to them.

For the listening experiment, the 6 loudspeakers were controlled from a PC connected through an ADAT connection to Behringer ADA8000 8-channel A/D and D/A converter.

4.4 Experimental procedure

Before the experiment, audiometry was performed for each participant, to rule out subjects with significant hearing loss. 18 people agreed to take part in the listening test, 10 male and 8 female, between 21 and 28. Hearing thresholds of 3 of them exceeded 20 dB hearing level at one or two audiometric frequencies (around 4 or 8 kHz), however, it can be argued, that such loss is not critical for the purpose of audio quality assessment. Therefore, after some considerations, a decision was made to include the responses from all of them in the analysis.

During the experiment, subjects were seated in the middle of the listening room, surrounded by a curtain, in order to prevent them from seeing the loudspeakers (see figure 4.3). This was done firstly, because of the unsymmetrical placement of the loudspeakers, and secondly, to avoid visual distractions during the listening session. Before entering the room they were also asked to close their eyes and were lead to the chair. They were asked to place their head on an headrest and not move it during the experiment. The ratings were submitted through a graphical user interface displayed on a touch-screen.

Figure 4.3: The experimental setup (two of the loudspeakers are not visible).

Due to time limitations, there was no full training session provided for the subjects. However, before starting the experiment, each of them was instructed on how to use the interface. Then, they were asked to rate perceived quality of the audio excerpts that they would hear, in comparison to an *ideal* reference, on the scales shown in figure 4.1 on page 33. They were told that they could play each sound as many times as they wanted, and in any order, although they were encouraged to always compare to the reference. Additionally, they were informed, that one of the sounds they were going to rate is a hidden reference.

None of the subjects reported any problems during the experiment and all of them finished it successfully.

4.5 Results

An example response from one subject is shown in figure 4.4.

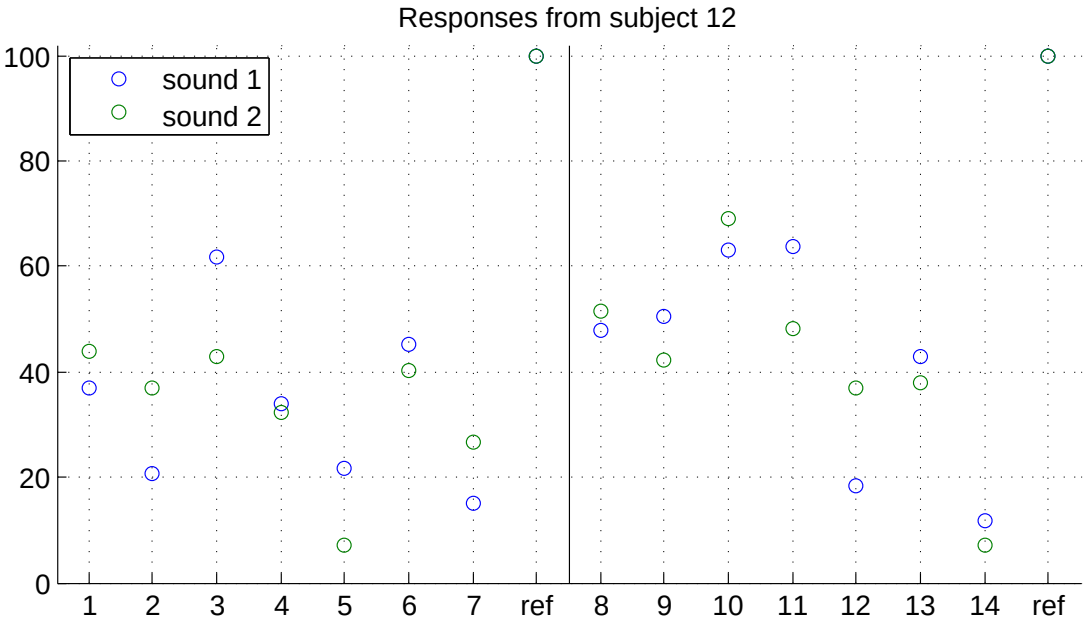


Figure 4.4: Responses obtained from one of the subjects. Numbers on the x-axis correspond to degradation types described in section 4.2 on page 31.

Summary of the obtained results can be seen in figures 4.6 on page 38 (combined results for spatial degradation types, for both sounds) and 4.5 on the next page (results for other types of degradations, for both sounds).

In majority of cases, subjects rated the hidden reference as the sample with the highest quality. 23 out of the total 32 ratings (18 subjects, 2 sounds) given to the hidden reference in the "spatial degradations" page indicated its best quality. For the other degradations, the fraction was even bigger: 26 out of 32 times the hidden reference was rated as best.

The spatial anchor, mono signal coming from the back left loudspeaker, was chosen as the worst of all samples 24 out of 32 times. The situation was, however, not so clear for the low-pass filtered anchor. It was marked worst only 14 times, with the degradation 5, -20dB hard limiter, being chosen 13 times. It seems that in the case of degradations other than the spatial ones, choice of the worst sample was more ambiguous.

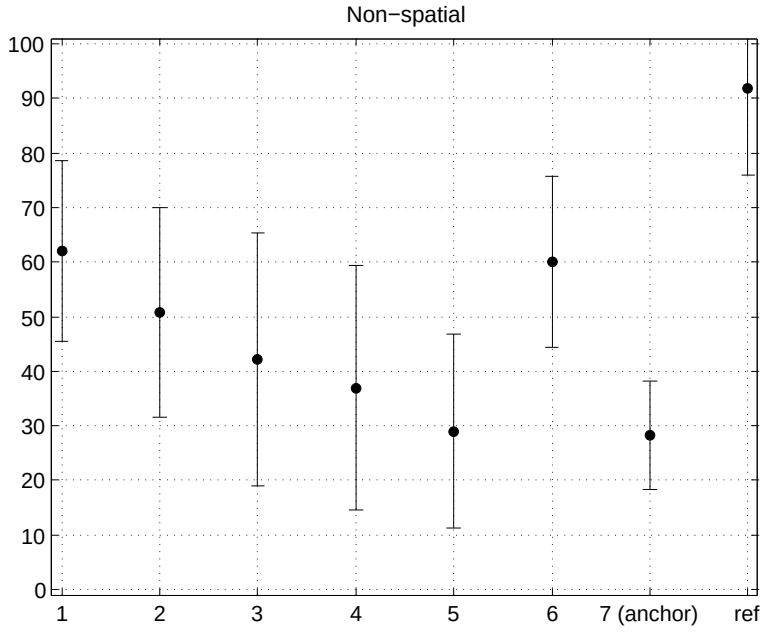


Figure 4.5: Means and standard deviations for non-spatial degradations, both sounds.

It can be noticed, that the spatial degradations were generally rated higher than other types of degradations. Figure 4.7 on the following page shows responses for all types of degradations within each group, averaged over all subjects, for sound 1 and sound 2. The mean for all responses, for both sounds, is 63.5 for the spatial degradations, and 50.1 for the other types. This trend was not equally prominent for all subjects. Figure 4.8 on page 39 shows results from two different subjects: subject 17

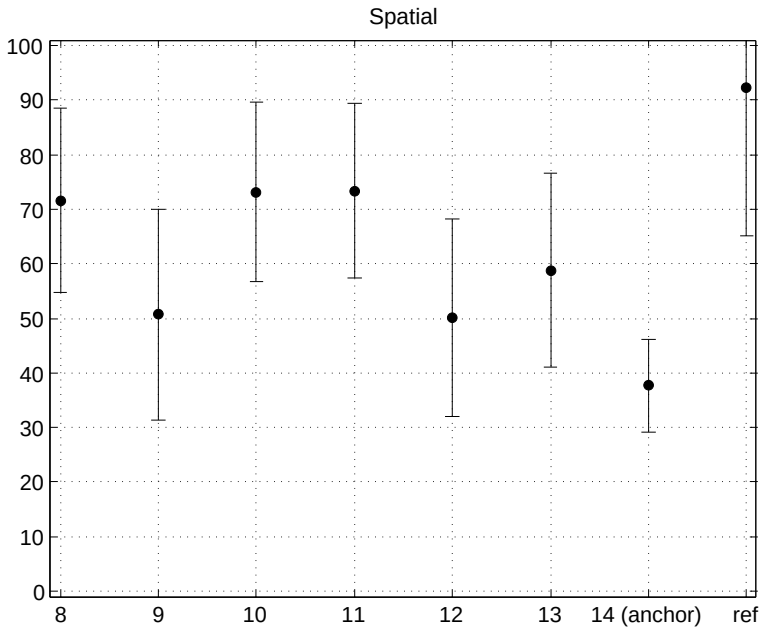


Figure 4.6: Means and standard deviations for spatial degradations, both sounds.

rated spatial degradations across the same range as the other ones, while subject 6 clearly favoured spatial degradations.

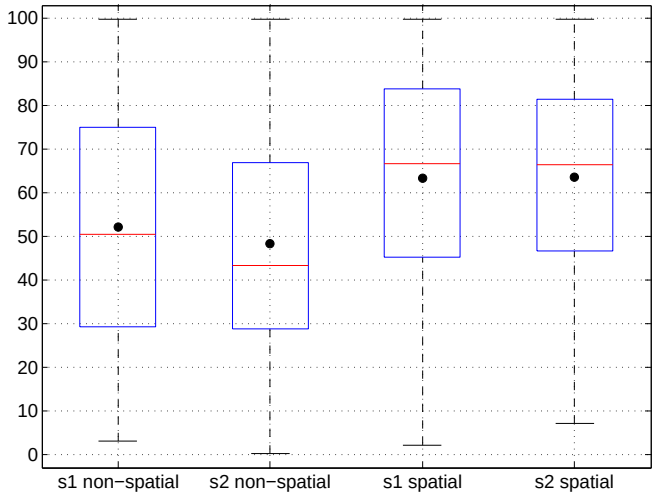


Figure 4.7: Boxplot showing results for sound 1 and sound 2 for all the subjects. Spatial degradations were generally rated higher than other types of degradations.

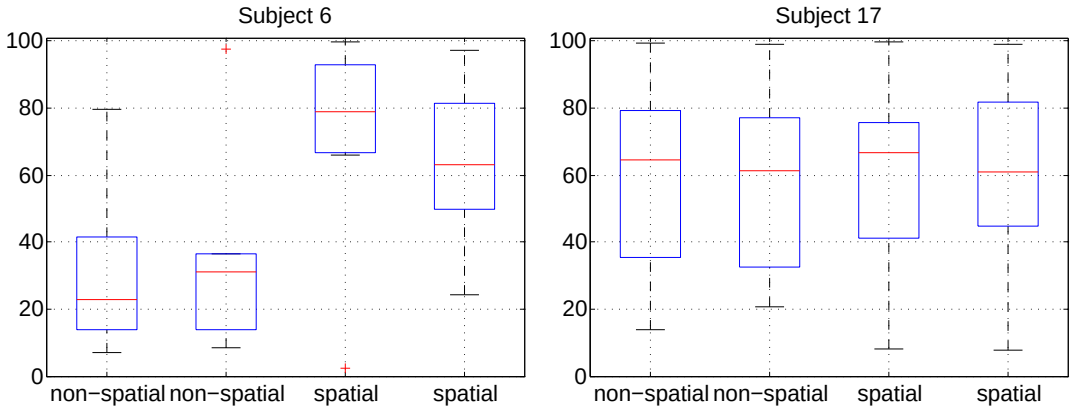


Figure 4.8: Some differences in how subjects rated sounds

4.6 Comments from participants

Most participants did not report having significant problems with the task. Many, however, said that rating spatial degradations was more challenging than rating the other group. Two people pointed specifically to the anchor (mono in left surround loudspeaker) as the most difficult to rate, because the sound was clear, only coming from the "wrong" direction. However, one of those subjects added, that they kept in mind that they should compare to the reference, so they rated the anchor as "bad".

Another interesting remark was from a person who thought that the second excerpt was in fact "too spatial", and that it sounded better coming from only one direction. This kind of surround sound would, according to them, be more appropriate for a film soundtrack, than for a musical piece.

Besides that, two or three other participants reported that they did in fact like one or more sound samples more than they liked the reference.

4.7 Summary

A listening test was conducted in order to find subjective quality ratings for 14 different quality degradations. MUSHRA method, slightly modified to better fit the purpose, was used. Results obtained from 14 subjects were presented in this chapter.

CHAPTER 5

Quality prediction with binaural models

Contents

5.1 Introduction 41
5.2 Sound samples pre-processing 41
5.2.1 Binaural room impulse response measurements 42
5.2.2 Gain adjustment 43
5.2.3 Sound pressure at the blocked ear canal 45
5.2.4 Ear canal transfer function 45
5.3 Decision device for audio quality assessment 46
5.4 Simulation results 48
5.5 Summary 52

5.1 Introduction

The following chapter illustrates the process of simulating audio quality assessment with auditory models described in chapter 3 on page 15. This assessment is based on comparing a test sound sample to a reference sound, thus obtaining a distance measure, which would indicate the perceived change in quality. The model here works as an artificial listener.

5.2 Sound samples pre-processing

In order to find the perceptual models' internal representations of the degraded signals under test, first the sound pressure entering the ears of the artificial listener needs to be found. In order to do that, binaural room impulse response (BRIR) of the room, where the listening test was carried out, was measured, and convolved with 5-channel sound samples.

Although the CASP model includes outer- and middle-ear filters, a decision was made not to include the first, since recorded BRIR already included the influence of the pinna (as well as torso). Instead, an ear canal filter was applied to the sound samples, giving the sound pressure at the eardrum, and the CASP model started from that point.

5.2.1 Binaural room impulse response measurements

BRIR measurements were made in the listening room with the exact same loudspeaker setup as used for the listening test. Impulse responses from each loudspeaker were recorded with Valdemar, an artificial head and torso simulator made at AAU (Christensen et al., 2000). The impulse response measurements were made with the maximum length sequence method (MLS), using MLSSA analyser ver. 7.0 (Rife, 1991).

The measurement resulted in 12 different transfer functions, from 6 loudspeakers and 2 microphones (ears) of Valdemar. Recorded impulse responses were 340 ms long. After being extracted from MLSSA, they were imported to Matlab and scaled with the stimulus amplitude. Figure 5.1 shows an example result - the measurement made for channel 1, left ear.

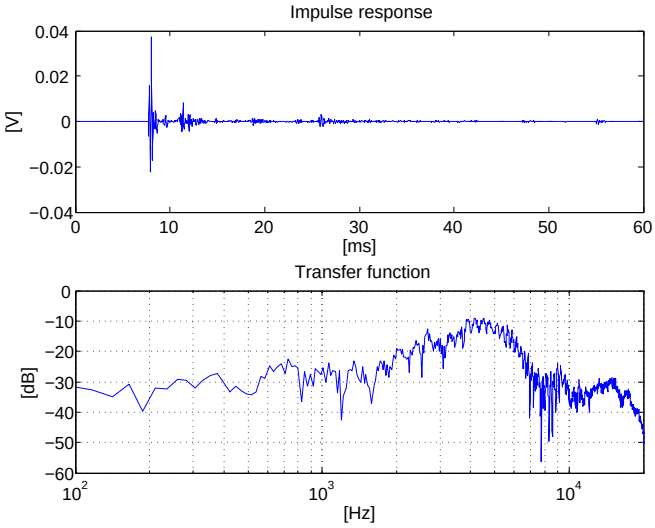


Figure 5.1: Impulse response and transfer function recorded from channel 1 (loudspeaker L) to the left ear microphone of Valdemar.

5.2.2 Gain adjustment

The setup used for the experiment was not exactly the same as the one used for BRIR measurements. Specifically, the latter were made using only the MLSSA system (for playback and recoding), while during the experiment a PC with an external sound interface was used to play the sound samples. Therefore, in order to accurately represent the sound pressure level at the ear of the listener, gain adjustments needed to be made.

Figure 5.2 illustrates the difference in the two setups. Setup in the top of the figure is the one used for measuring BRIR, middle - in the listening test, and bottom - in the gain adjustment measurement.

To find the soundcard gain, additional measurement was carried out. A 500 ms long white noise sample was played through the soundcard and each loudspeaker, and recorded with MLSSA. A white noise sample of the same length and with the same RMS value was then convolved with BRIR for each channel and ear, and its transfer function was compared to the recorded one.

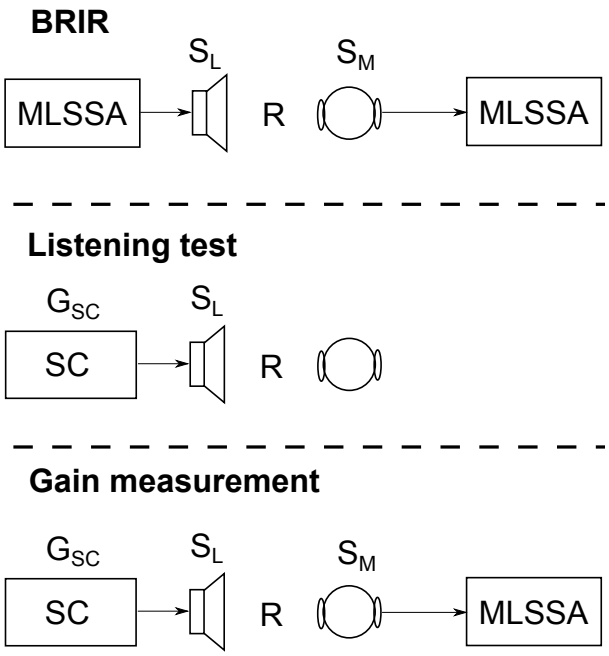


Figure 5.2: Comparison of three set-ups, top: during the BRIR measurements, middle: during the listening test (also for the artificial listener), bottom: during the gain adjustment measurement.

Looking back at figure 5.2, if x is the noise sample played through the sound card and the loudspeakers, $\mathrm{S_L}$ – the loudspeaker transfer function, $\mathrm{S_M}$ – the microphone sensitivity, and R – the response of the room, the sound card gain $\mathrm{S_L}$ can be found from:

$$\mathrm{G_{SC}} = \frac{x_{measured}}{x * \mathrm{S_L R S_M}} = \frac{x_{measured}}{x * \mathrm{BRIR_n}} \tag{5.1}$$

where $\mathrm{BRIR_n}$ is the actual measured impulse response for a given channel and ear.

The difference was found to be about 17 dB and is illustrated in figure 5.3. The reason why the two transfer functions are not exactly the same is that due to practical issues the gain adjustment measurement was not fully optimal. At the point when this measurement was made, only the MLSSA system was calibrated with Valdemar microphones, and the input of the sound card was not. Therefore, the recordings were made with MLSSA, and because of the limitations of the system's available acquisition length, and differences in delays between sound card output and MLSSA input, it was practically impossible to synchronize the two so that it is known exactly which part of the noise signal was recorded. However, for this purpose, it should be enough to use a different sequence of the same kind of noise.

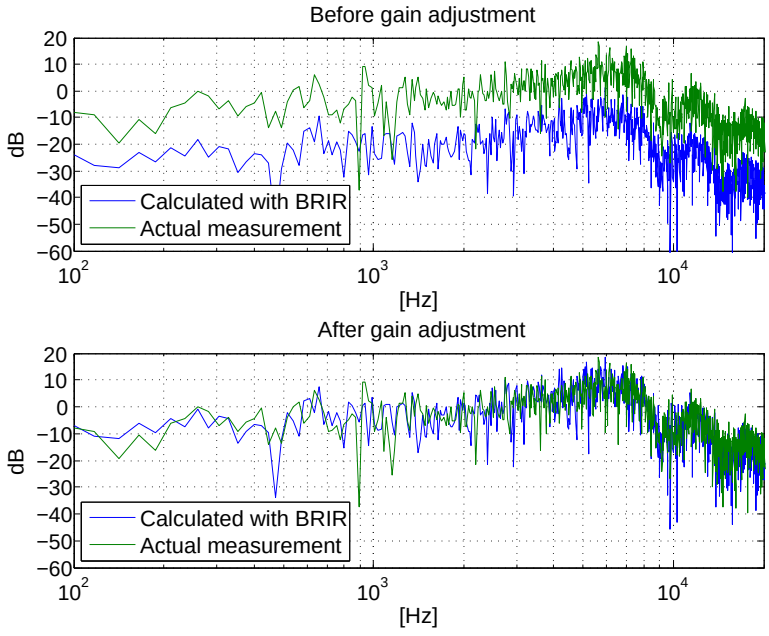


Figure 5.3: Gain adjustment of the white noise sample (channel 4, left ear)

5.2.3 Sound pressure at the blocked ear canal

By convolving the sound samples with the gain adjusted binaural room impulse response, the sound pressure at the entrance to the ear canal of the artificial listener can be found. Each audio channel of the sound samples was convolved with the impulse response of a corresponding system channel (corresponding loudspeaker), and this was done for both ears. For example, for one sound sample S, first, convolution with the room impulse response B was applied for each channel n:

$$\begin{aligned} S_n * B_L^n &= P_L^n, \qquad \text{for} \quad n = 1, 2, \ldots, 5 \\ S_n * B_R^n &= P_R^n, \qquad \text{for} \quad n = 1, 2, \ldots, 5 \end{aligned} \tag{5.2}$$

and then the sound pressure at each ear (L and R) was summed together:

$$\begin{aligned} P_L &= P_L^1 + P_L^2 + P_L^3 + P_L^4 + P_L^5 \\ P_R &= P_R^1 + P_R^2 + P_R^3 + P_R^4 + P_R^5. \end{aligned} \tag{5.3}$$

5.2.4 Ear canal transfer function

Processing applied to the sound samples so far gives the sound pressure as measured at the blocked entrance to the ear canal. In order to find sound pressure at the eardrum, transfer function from blocked ear canal entrance to the eardrum was used. Results of measurements made by Hammershoi and Moller (1996) in the ear canals of 12 subjects were used to construct a minimum-phase FIR filter, the magnitude response of which is shown in figure 5.4. This filter was applied to each binaural sound sample.

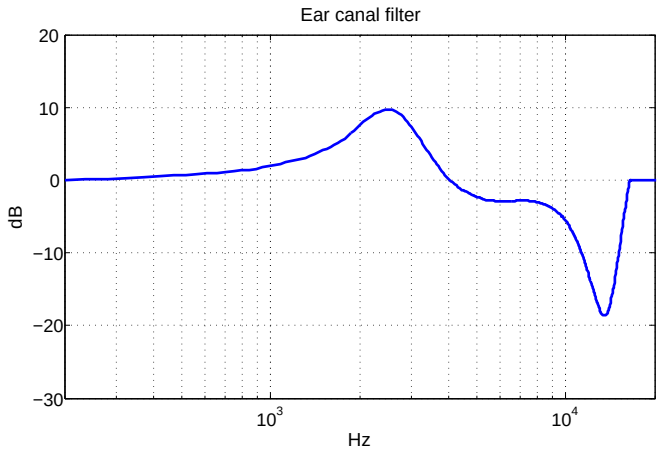


Figure 5.4: Transfer function of the FIR filter used for ear canal filtering

5.3 Decision device for audio quality assessment

In the next step, signals representing the sound pressure at the eardrum of the left and the right ear were fed into the combined binaural models, and each channel separately was processed by monaural CASP model. Due to technical difficulties (not enough RAM for the computation), the modulation filterbank was not included in CASP.

Internal representations of a sound sample under test, obtained both from the binaural models and CASP, were compared with the corresponding internal representations of a reference sound. The comparison was made in a detector, which is described below. This basic idea of objective quality assessment is illustrated in figure 5.5.

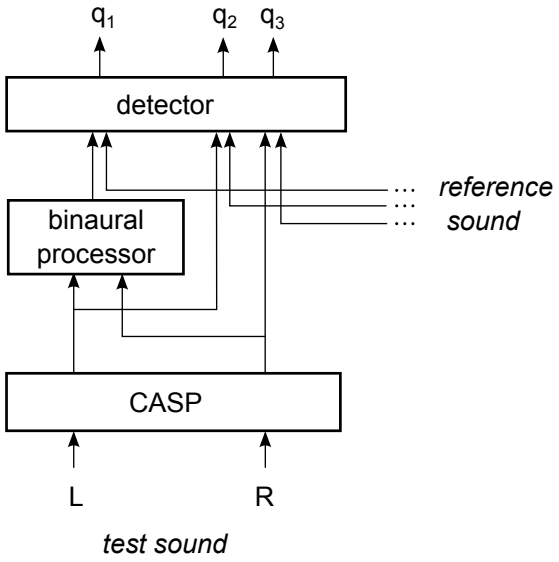


Figure 5.5: Outline of the full objective quality assessment model. q_1 is the binaural prediction, q_2 and q_3 are the monophonic predictions for the left and right channel.

Detector used in this project is a correlation measure based on the work by Oldenburg and Aps (2000). First, time integration is applied to the internal representations in 20 ms windows with no overlap. Then, a frequency-weighted correlation measure q is calculated according to the equation:

$$q = \frac{\sum_i \sum_j (w_j X_{i,j} - \overline{X})(w_j Y_{i,j} - \overline{Y})}{\sqrt{\sum_i \sum_j (w_j X_{i,j} - \overline{X})^2} \sqrt{\sum_i \sum_j (w_j Y_{i,j} - \overline{Y})^2}} \tag{5.4}$$

where X and Y are the test and reference sounds, and indices i and j correspond to time and frequency channel.

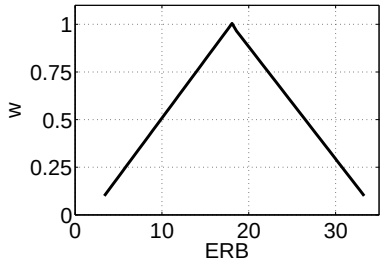
Input to this detector is a 2-dimensional matrix of time vs. frequency channel, thus a single number had to be chosen for each time frame for CASP-L and CASP-B models. This problem has previously been addressed in section 3.5 on page 21. For the analysis shown below, the following was chosen:

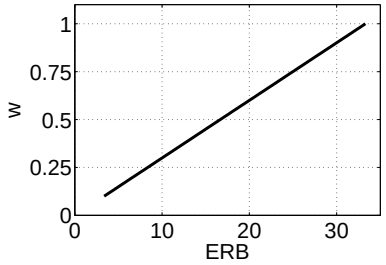
- Lindemann: value at $\tau = 0$,
- Breebaart: value at $\tau = 0, \quad \alpha = 0$.

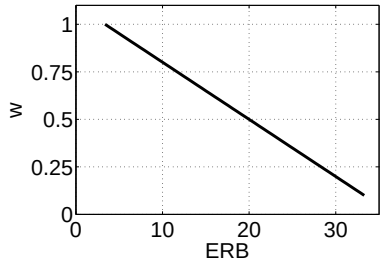
In the case of the Lindemann model, a centroid along the delay line, and the location of the maximum were also briefly considered, but they seemed to provide worse results.

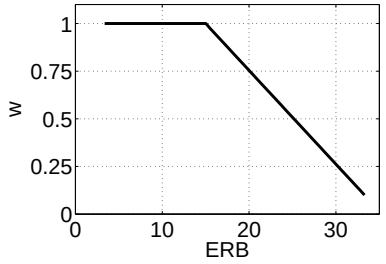
To examine which frequency weighting would be most appropriate for the purpose of audio quality assessment, or if there is any benefit of frequency weighting at all, 10 different weightings were chosen, as shown in table 5.1.

Table 5.1: Frequency weightings

1. 

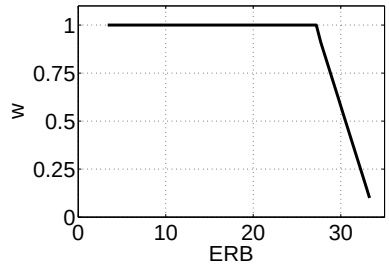
2. 

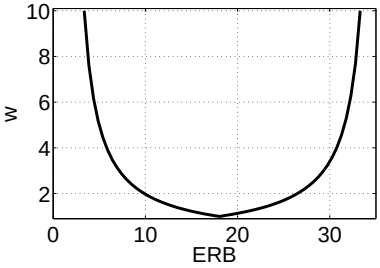
3. 

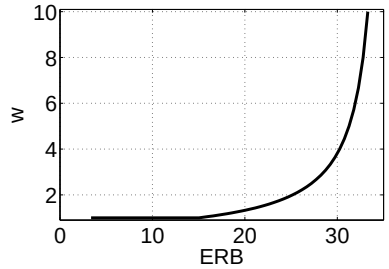
4. 

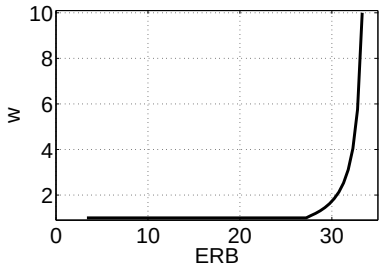
Continued on next page

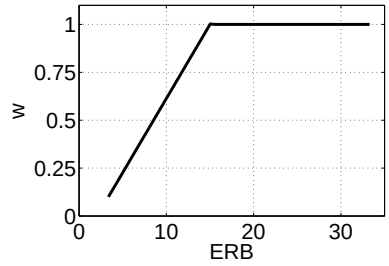
Table 5.1 – *Continued from previous page*

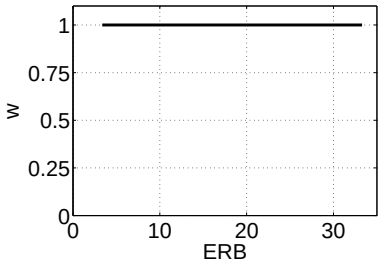
5.


6.


7.


8.


9.


10.


5.4 Simulation results

Output values of the detector were compared to responses obtained from the listening experiment. The results are presented in this section. Table 5.2 on the facing page summarises all the possible model combinations, by presenting the correlation coefficients between the particular model's prediction and the actual responses from subjects participating in the listening test. This is shown for each binaural model, and for each frequency weighting in the detector. The "mono" results are the mean of the detector outputs from the left and right channel.

It should be noted, that results for spatial and non-spatial degradations are presented, and will be analysed separately. This is partly because those two tasks were separated in the listening experiment, and partly to see how the models' predictions differ when processing a sound degraded in space, and a sound degraded in a different manner. Additionally, since it is not yet clear how the three different outputs of the CASP-D model could be optimally combined, all of them are presented individually.

Certain observations can be made when analysing results in table 5.2. Firstly, clearly not in all the cases there is a significant correlation between the prediction and the subjective response. Some of the predictors, such as the CASP-D modulation or CASP-D fine structure, do not offer significant correlation with the responses under any of the tested weightings. Some of the other, however, show relatively high correlation – such as the CASP-L for spatially degraded sounds.

Table 5.2: Correlation coefficients between predictions and subjective responses, averaged sound sample 1 and sound sample 2; in bold - correlation is significant (p-value < 0.05).

weighting	**mono**		**L**		**B**	
	o	s	o	s	o	s
1.	0.559	**0.750**	0.093	**0.801**	**0.729**	0.678
2.	**0.805**	**0.798**	0.068	**0.859**	0.587	0.629
3.	0.283	**0.758**	0.013	**0.783**	0.550	**0.840**
4.	0.344	**0.764**	0.021	**0.786**	0.573	**0.826**
5.	0.443	**0.771**	0.021	**0.781**	**0.792**	**0.834**
6.	0.175	**0.749**	0.006	**0.785**	0.579	**0.831**
7.	0.696	**0.826**	0.031	**0.884**	0.545	0.665
8.	0.560	**0.816**	0.028	**0.835**	0.581	0.681
9.	0.686	**0.772**	0.094	**0.803**	0.683	0.673
10.	0.462	**0.776**	0.021	**0.778**	**0.718**	**0.828**

weighting	**D** fine		**D** mod		**D** ILD	
	o	s	o	s	o	s
1.	0.267	0.645	0.398	0.661	0.660	**0.769**
2.	0.266	0.648	0.527	0.673	**0.812**	**0.777**
3.	0.244	0.689	0.379	0.677	0.537	**0.747**
4.	0.251	0.680	0.378	0.674	0.563	**0.758**
5.	0.254	0.673	0.442	0.671	**0.715**	**0.748**
6.	0.206	0.704	0.502	0.673	0.687	**0.776**
7.	0.257	0.666	0.524	0.672	**0.719**	**0.805**
8.	0.254	0.673	0.473	0.664	**0.744**	**0.805**
9.	0.263	0.655	0.471	0.669	**0.792**	**0.768**
10.	0.254	0.673	0.464	0.672	**0.742**	**0.758**

Secondly, for some of the predictors, introducing frequency weighting substantially improves their performance. This is perhaps most clearly seen in the ability of

the monaural CASP output to predict non-spatial degradations, where correlation increases from 0.462 to 0.805 as a result of introducing an optimal frequency weighting. However, it can also be noticed, that the optimal weighting varies across the predictors.

Figures 5.6-5.9 show scatter plots of the objective prediction versus subjective response, both the spatial and non-spatial degradations, for the frequency weighting which was found most optimal for the detector. Sometimes, in the case of binaural predictors, priority was given to detecting spatial degradations - such as when choosing weighting 7 for CASP-D ILD, rather than weighting 2, even though the latter shows higher correlation in general. Altogether, it seems that the best binaural predictor is CASP-L with weighting 7, and the best predictor for non-spatial degradations is CASP-D ILD with weighting 2, although CASP monaural detector with weighting 2 also performs well in this task.

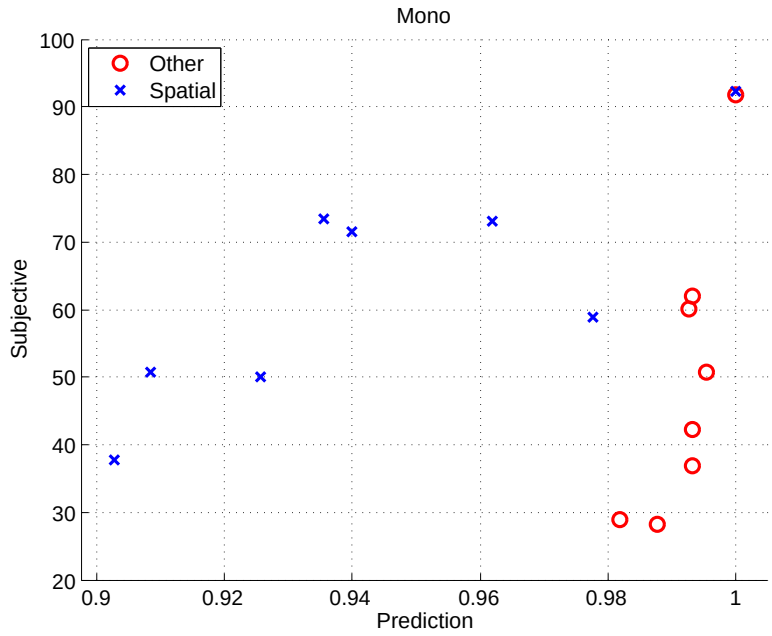


Figure 5.6: Mono (mean between 2 channels), weighting 2

It is interesting to see, that for most optimally weighted predictors, there seems to be a clear distinction between how the model rates degradations that are of spatial and non-spatial nature. For the same perceived subjective quality, all the predictors will rate the spatial degradation lower than the other kinds.

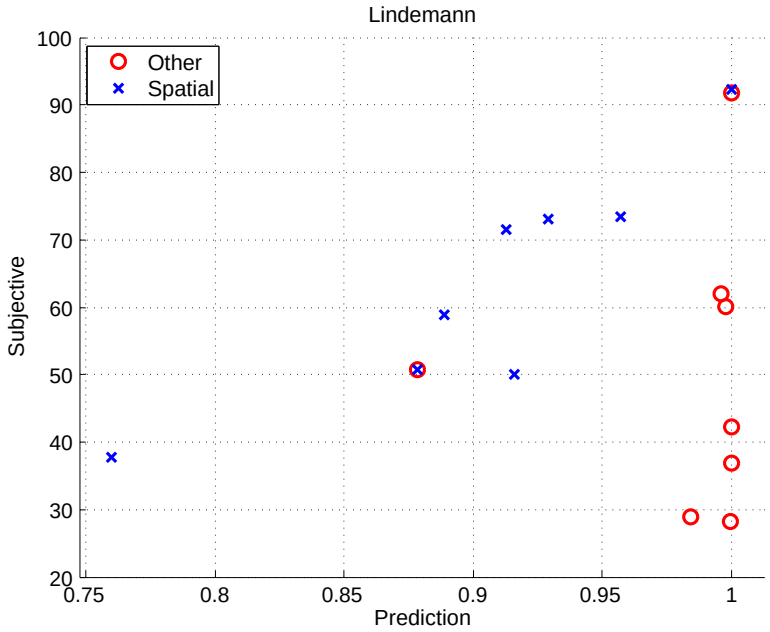


Figure 5.7: Lindemann ($\tau = 0$), weighting 7

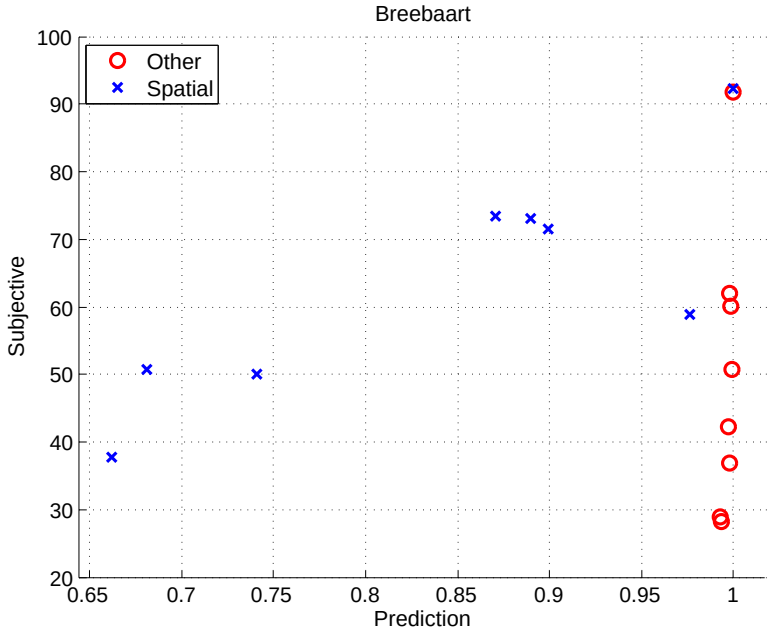


Figure 5.8: Breebaart ($\tau = 0, \alpha = 0$), weighting 5

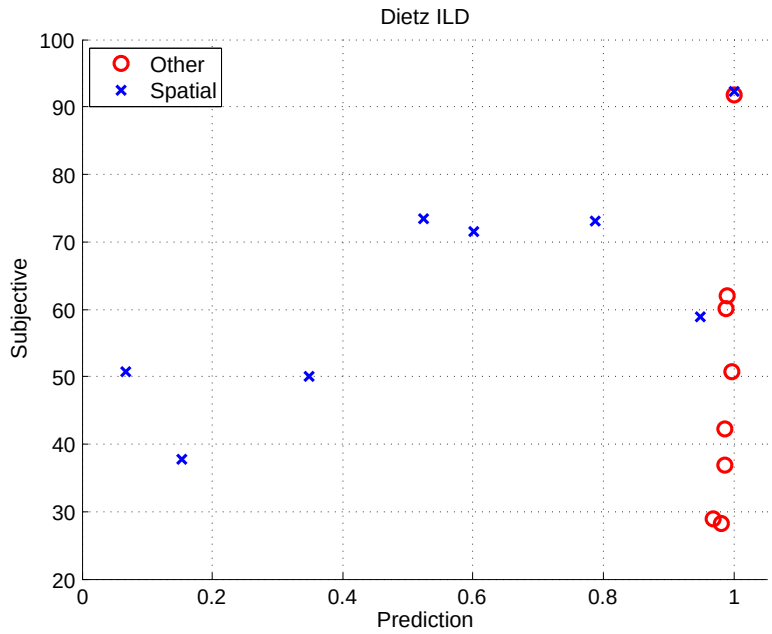


Figure 5.9: Dietz ILD, weighting 8

5.5 Summary

In this chapter, a process of obtaining objective quality predictors from an artificial listener was described. Three models presented in chapter 3, and a frequency-weighted correlation-based detector were used for the purpose. Then, prediction results were illustrated by means of correlation with subjective responses from the listening test. CASP-L model proved to be superior in predicting spatial degradations.

CHAPTER 6

Discussion and Conclusions

Contents

6.1 Discussion **53**
6.1.1 Listening test 53
6.1.2 Objective quality prediction 54
6.1.3 Areas of potential future work 55
6.2 Conclusions **56**

6.1 Discussion

6.1.1 Listening test

Although the listening test was successful and provided useful information for the project, certainly there were some aspects of it that could be improved.

Firstly, other types of sound recordings than used in this project should be tested as well. Those could include other musical genres, but also film soundtracks, as they are very often designed to be played in a multichannel setup. It is likely, that the change in quality a certain degradation to a sound produces, will be somehow related to the nature of the sound itself.

Secondly, although costly, it could be beneficial to get a more controlled group of trained subjects. From the comments obtained from the subjects after the experiment, as well as some of the responses they provided, it is presumed that not only did some of the subjects have problems with the task, but some of them were most likely not even able to perform the basic task of distinguishing between the reference and degraded signal. There were a few cases, in which sounds different than the hidden reference were rated as the best – even a sound sample with added noise. Another subject rated the hidden reference as worse than all the spatial degradations. It is difficult to say, if the reason for this was – as reported by some of the participants – the fact that they liked other sounds more than the reference, or that they simply were

not able to recognize it. On the contrary, two subjects who reported that the task was "very easy" and they had "no problems at all", provided some of the most consistent (sound 1 compared to sound 2) reponses. They also had no problems recognizing the reference and rated it the highest on the scales.

A solution to those problems could be giving the subjects an intensive training session prior the experiment, and setting a rejection criterion, which would be the ability to correctly identify the hidden reference. The main reasons why this was not done in this project, was that firstly, it is much more time consuming, and secondly, the number of available subjects was limited.

6.1.2 Objective quality prediction

As mentioned in section 5.4, some perceptual models tested in this project do correlate with subjective quality ratings, however, they seem to behave differently when processing sounds with spatial degradations, compared to sounds with non-spatial degradations. Clearly, the model overestimates spatial degradations, which in reality do not cause such a big change in quality (or, the opposite – it underestimates non-spatial degradations).

It is interesting to see, that for no weighting applied, as well as for so many other weightings, the monophonic CASP model predicts the non-spatial degradations so poorly, compared to the spatial ones, while some of the binaural models cope with this task better. Further investigation of the results would be needed to try to find a reason for this situation.

A possible combined model, which should maximise the prediction accuracy, could consist of the CASP monaural model (which predicts the non-spatial degradations well), and the CASP-L binaural part, which should take care of the spatial degradations. Figure 6.1 shows those two parts, together with 2-nd order polynomials fitted to the data.

In order for this model to work, a way would have to be found to determine, what kind of degradation the test sound contains, and then an appropriate prediction could be obtained from one of the fitted polynomials. In the case of degradations which were tested in this project, all the non-spatial ones are rated above, say, 0.97, while most spatial fall much lower than 0.97 (and this is largely true also for other predictors). It is, therefore, tempting to say, that an easy way of determining which kind of degradations are present, is to check if the monaural detector's prediction is greater

than 0.97 – if it isn't, the monaural prediction (red polynomial) is not accurate, and the binaural model (blue polynomial) should be applied instead. Of course, this is a simple case, where there are no combinations of spatial and non-spatial degradation types, or no degradations in-between. Furthermore, the value 0.97 was chosen quite arbitrarily, and may not be the most optimal one.

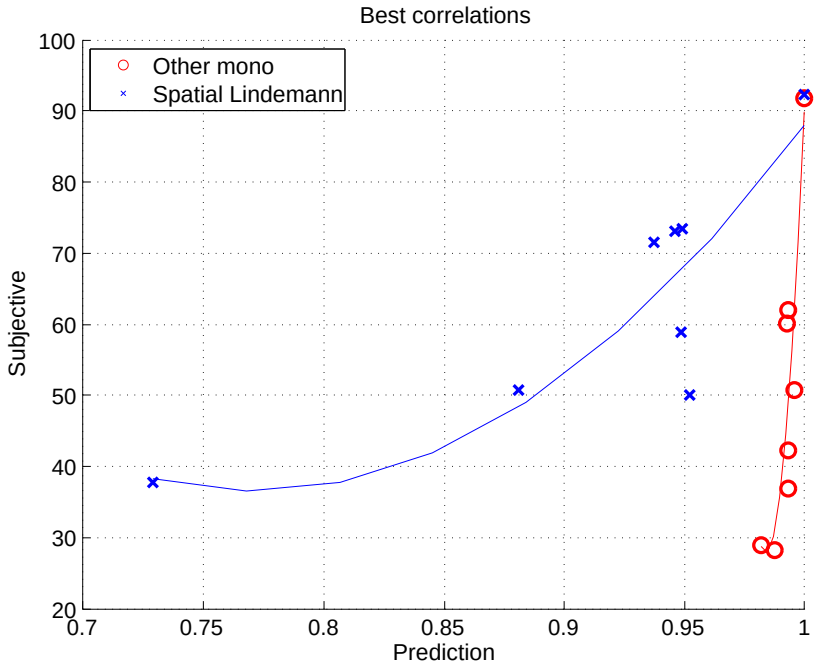


Figure 6.1: A combination of the CASP monophonic model (only non-spatial degradations shown) and the Lindemann model (only spatial degradations shown). The red and blue lines are 2nd order polynomials fitted to the data.

6.1.3 Areas of potential future work

There are several ways, in which the investigation started in this project could be continued. Due to time limitations of the project, not all the possible variations of the models were considered. There are still some parameters, which could be changed and tested to see if the change can improve the model's predictions. An example of such a parameter is the time integration constant in the decision device. The currently used $\tau = 20\text{ms}$ is taken from a speech quality algorithm (Oldenburg and Aps, 2000), and is "oriented at the typical frame rate used in the analysis and synthesis algorithms in speech coders". Perhaps a better constant for audio could be found.

Another improvement could be considering a different detector, for example one based on a difference rather than correlation, as used by Breebaart et al. (2001). Their model is, however, fitted to the purpose of signal detection rather than quality prediction, and would have to be modified in some way.

Moreover, the different outputs of CASP-D model – IPD from the modulation filter, from the fine-structure filter, and ILD – should be analysed in more detail to find out if they could be combined in a way, that would improve audio quality predictions. Also, in the case of CASP-B model, different approaches than simply taking into account the value at $\tau = 0$, $\alpha = 0$ could be investigated.

Finally, the modulation filterbank should be included in the monaural CASP model. This gives the possibility to obtain more information from the monophonic predictions. This is particularly interesting, because the modulation filterbank is, next to the DRNL filterbank, the main difference between CASP and other similar models. It is also rather straightforward to do, as long as the technical difficulties have been overcome.

6.2 Conclusions

In the project, three models were tested to see if they could be used for audio quality assessment. Each of those models consisted of a part of the CASP computational auditory model of perception, as well as one of the three binaural processors, considered for the project. Those were based on the work of Lindemann (1986), Breebaart et al. (2001) and Dietz et al. (2008). The last stage of the assessment model was a decision device, based on a frequency weighted correlation measure.

A listening test was conducted, in order to validate the combined objective models, as well as to adjust some of their parameters in order to optimise their predictions. The experiment was carried out in a multi-channel set-up, in order to be able to include various types of spatial degradations it in. The subjects' task was divided into either rating spatial, or non-spatial degradations to the given sound samples.

Optimal frequency weightings for each binaural processor (as well as the monophonic one) were selected. It was found, that the best prediction of spatial degradations can be obtained with the CASP-L model, based on the Lindemann binaural processor. In the case of non-spatial degradations, the best performance was given by CASP-D (Dietz) ILD, however monophonic CASP output was also good.

Still, a lot is left to investigate if the model should work automatically, with the user being able to obtain a single number predicting sound quality degradation, only by feeding the model with reference and test sounds – which is the ideal situation. Some of the areas, which could be improved, were discussed in this chapter.

APPENDIX A

Responses from the listening test

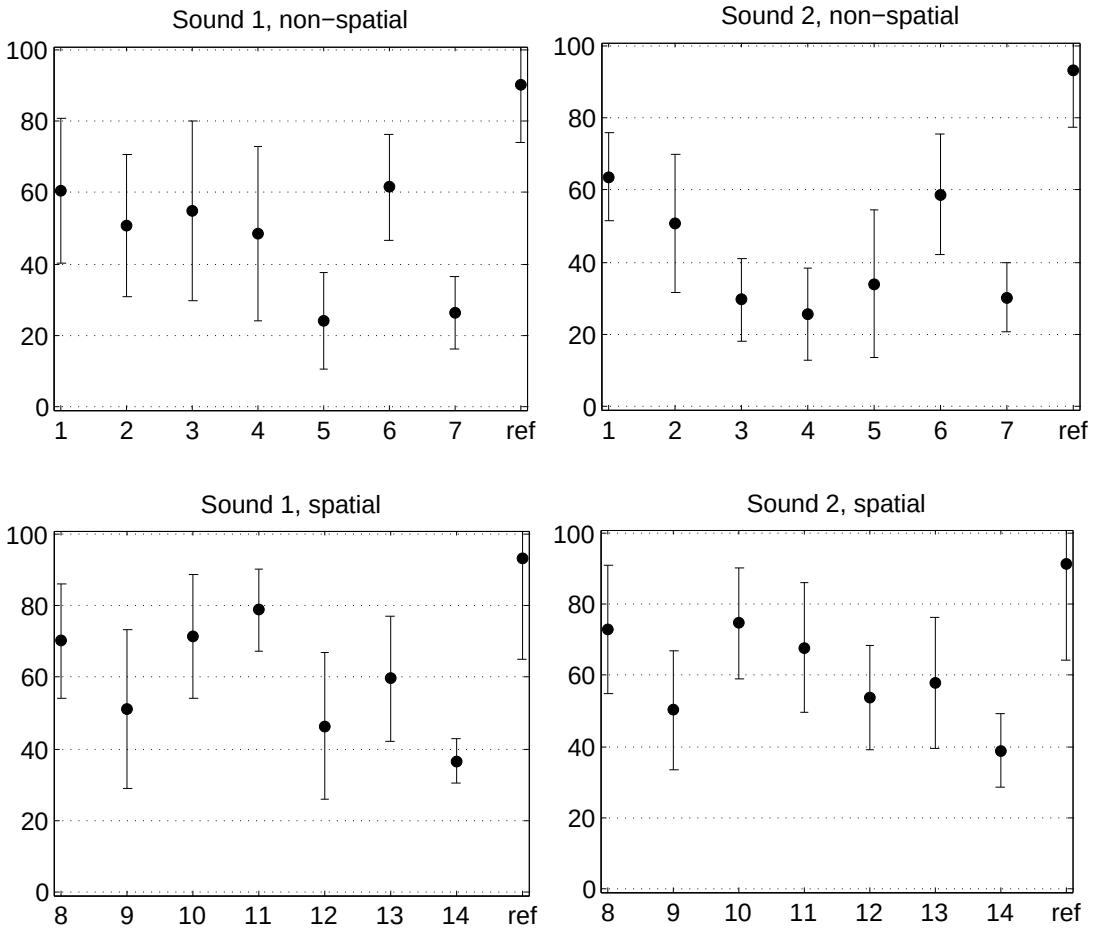


Figure A.1: Mean values and standard deviations of responses averaged across subjects.

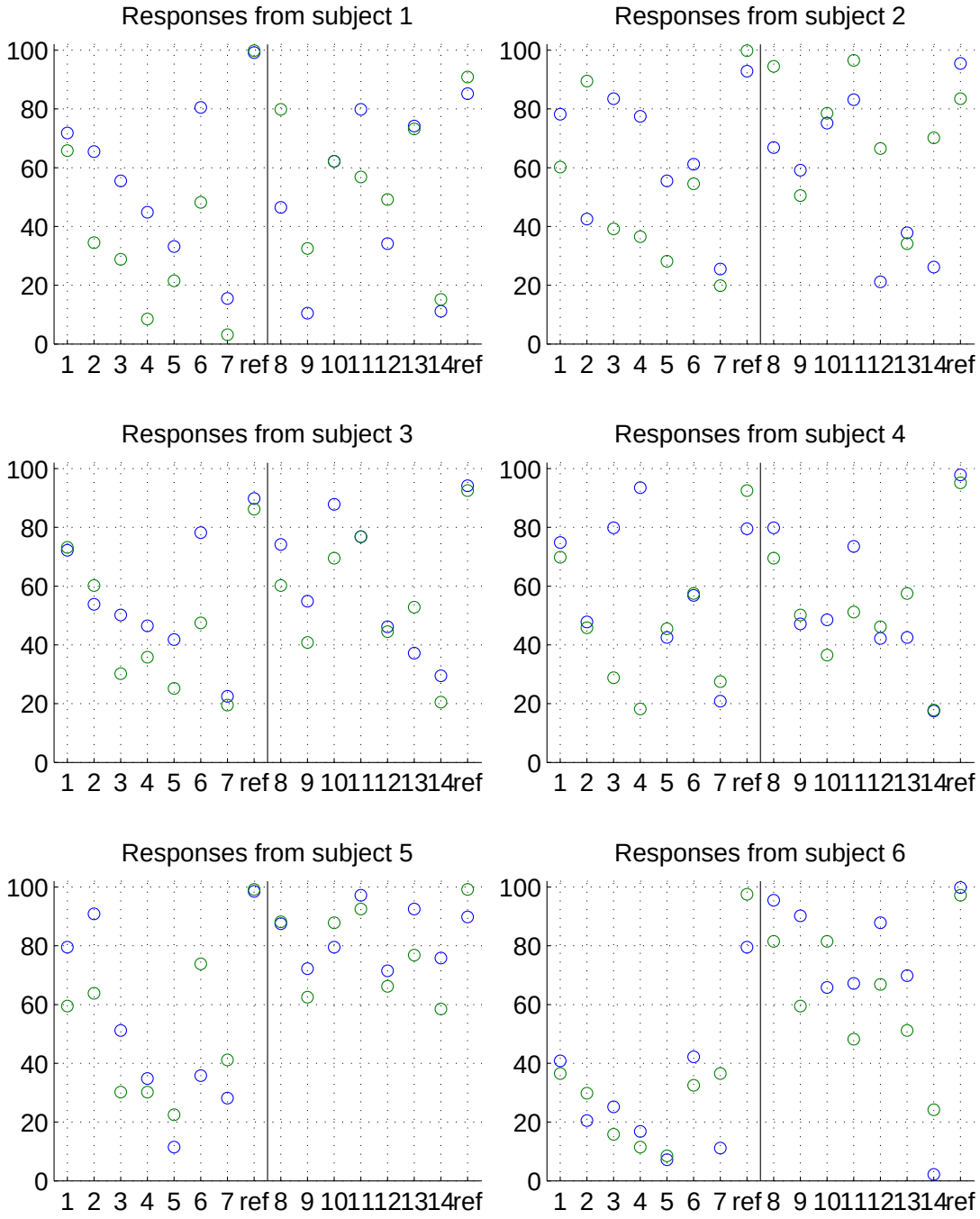


Figure A.2: All subjective responses pt.1, blue - sound 1, green - sound 2

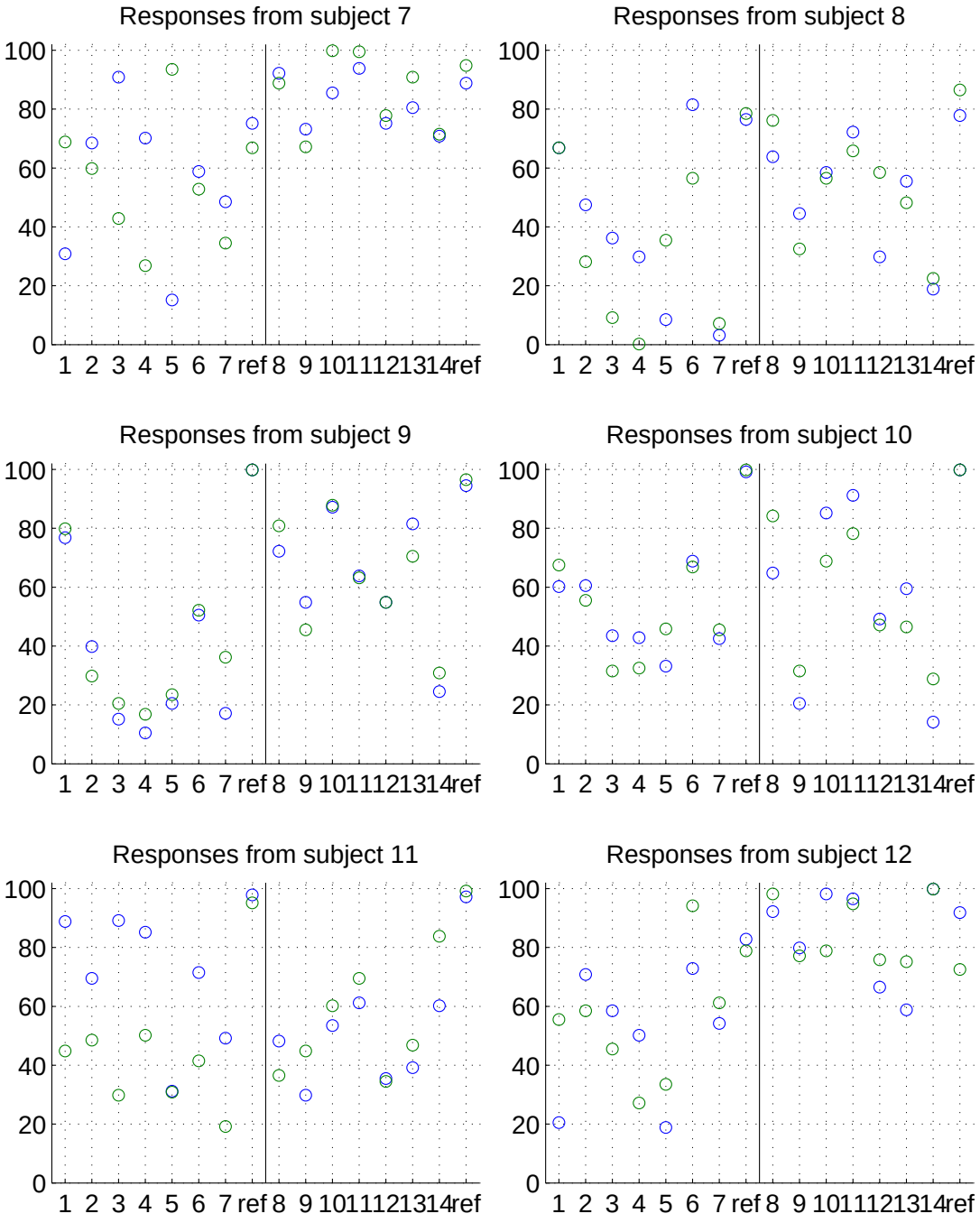


Figure A.3: All subjective responses pt.2, blue - sound 1, green - sound 2

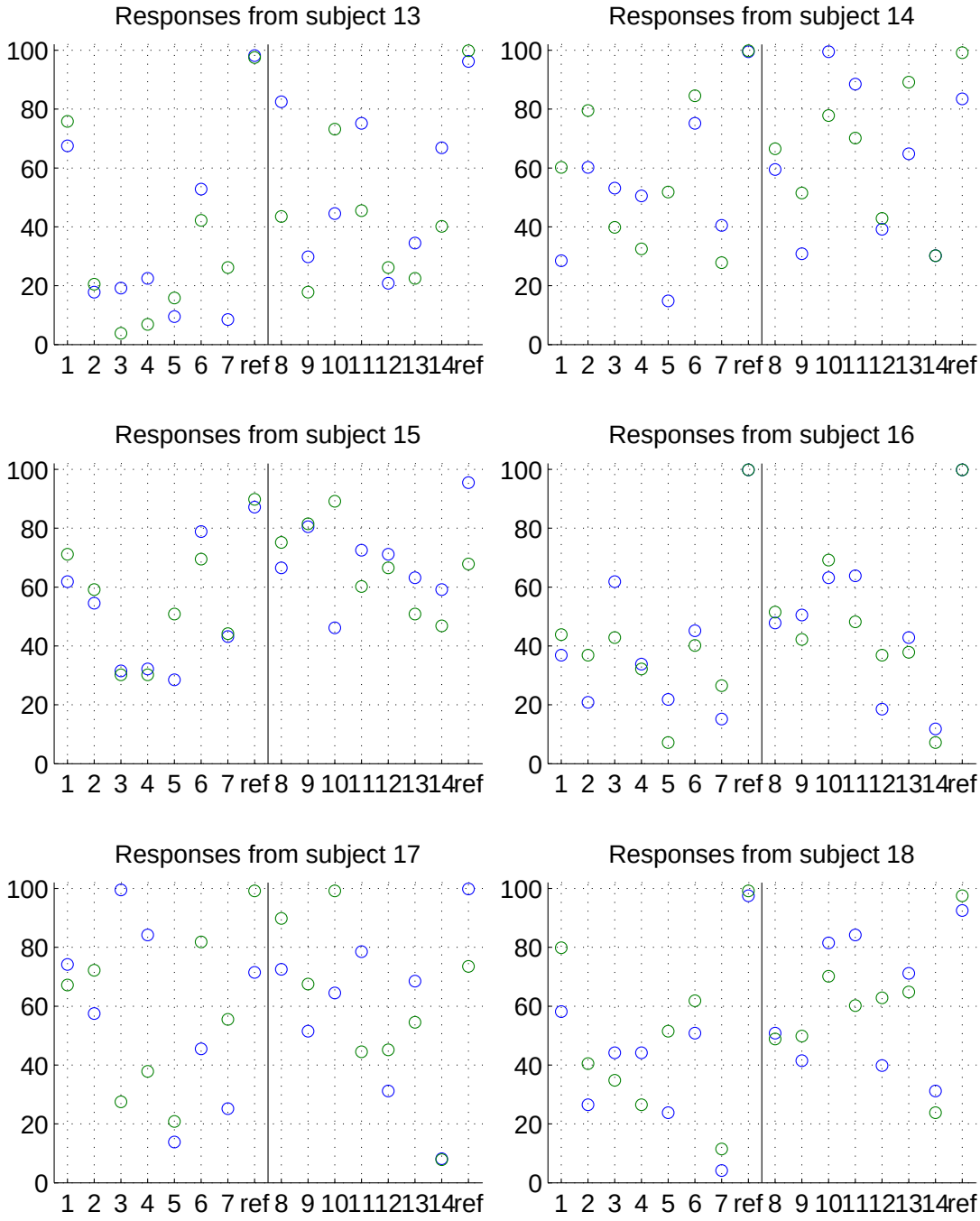


Figure A.4: All subjective responses pt.3, blue - sound 1, green - sound 2

APPENDIX B

Enclosed DVD contents

`Report.pdf`: Full report.

- **`BRIR/`**
  - 12 .TIM files with BRIR measurements obtained from MLSSA, for 6 channels and 2 microphones (ears)
  - `valdemar.mat`:
    a Matlab file including 12 impulse response files measured with Valdemar
- **`Detector/`**
  - `corrDetector.m`:
    function used to calculate the correlation predictor
  - `freqWeighting.m`:
    function calculating one of 10 tested frequency weightings
- **`GUI + sound samples/`**
  - **`samples/`**
    4 folders (2 sound samples, 2 types of degradations), each including 7 degraded samples, a reference and a hidden reference, all .wav files
  - `gui.m`:
    graphical user interface for the listening experiment
  - `get_samples.m`:
    function used by `gui.m` to read sample paths from the folder 'samples'
- **`Models/`**
  folder including all the functions needed to process samples with the auditory models used in this project. Before it can be used, the AMToolbox and LTFAT packages need to be initialized, which can be done by running the file `combined_init.m`. After initialisation, the combined binaural models can be run with the function `combined_models.m`.

Bibliography

Data sheet: Genelec 1031A Bi-amplified Monitoring System. (Cited on page 34.)

Søren Bech and Nick Zacharov. *Perceptual audio evaluation : theory, method and application*. John Wiley and Sons, 2006. (Cited on page 2.)

Jeroen Breebaart, Steven Van De Par, and Armin Kohlrausch. Binaural processing model based on contralateral inhibition. i. model structure. *Journal of the Acoustical Society of America*, 110(2):1074–1088, 2001. (Cited on pages 15, 18, 19 and 56.)

Flemming Christensen, Clemen Boje Jensen, and Henrik Møller. The Design of VALDEMAR - An Artificial Head for Binaural Recording Purposes. *109th Convention of the Audio Engineering Society*, 2000. (Cited on page 42.)

R. Conetta, F. Rumsey, S. Zielinski, P.J.B. Jackson, M. Dewhirst, S. Bech, D. Meares, and S. George. QESTRAL (part 2): Calibrating the QESTRAL model using listening test data. In *Proc. 125th AES Conv., San Francisco CA*, number 7596, Oct. 2008. (Cited on page 32.)

Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *Journal of Acoustical Society of America*, 1997a. (Cited on page 5.)

Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *Journal of Acoustical Society of America*, 1997b. (Cited on page 5.)

Mathias Dietz, Stephan Ewert, Volker Hohmann, and Birger Kollmeier. Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences. *Brain Research*, 1220:234–245, 2008. (Cited on pages 15, 20 and 56.)

Mathias Dietz, Stephan D Ewert, and Volker Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605, 2011. (Cited on pages 20 and 21.)

N I Durlach. Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, 35(8):1206, 1963. (Cited on page 18.)

Dorte Hammershoi and Henrik Moller. Sound transmission to and within the human ear canal. *The Journal of the Acoustical Society of America*, 1996. (Cited on page 45.)

ITU-R BS.1387-1. Method for objective measurements of perceived audio quality, 2001. (Cited on page 3.)

ITU-R BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems, 2003. (Cited on pages 31 and 32.)

ITU-R BS.775-2. Multichannel stereophonic sound system with and without accompanying picture, 2006. (Cited on page 34.)

L A Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35–39, 1948. (Cited on pages 16 and 17.)

Morten L Jepsen, Stephan D Ewert, and Torsten Dau. A computational model of human auditory signal processing and perception. *Journal of Acoustical Society of America*, 2008. (Cited on pages 5 and 6.)

W Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals. *Journal of the Acoustical Society of America*, 80(6):1608–1622, 1986. (Cited on pages 15, 16, 17 and 56.)

Enrique A. Lopez-Poveda and Ray Meddis. A human nonlinear cochlear filterbank. *Journal of Acoustical Society of America*, 2001. (Cited on pages 7 and 8.)

Brian C. J. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition*. Academic Press, April 2003. ISBN 0125056281. (Cited on page 1.)

Universitiit Oldenburg and Westermann Aps. Objective modeling of speech quality with a psychoacoustically validated auditory model. *October*, 48(5):395–409, 2000. (Cited on pages 46 and 55.)

D. Rife. *MLSSA Reference Manual vr. 7.0*. DRA Laboratories, Sterling, VA, 1991. (Cited on page 42.)

David J M Robinson. *Perceptual model for assessment of coded audio*. PhD thesis, University of Essex, 2002. (Cited on page 3.)

Francis Rumsey, Slawomir Zielinski, Philip Jackson, Martin Dewhirst, Robert Conetta, Sunish George, Søren Bech, and David Meares. Qestral (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener. In *October*, pages 1–8, 2008. URL `http://www.aes.org/e-lib/browse.cfm?elib=14746`. (Cited on page 3.)

Peter L. Søndergaard, John F. Culling, Torsten Dau, Nicolas Le Goff, Morten L. Jepsen, Piotr Majdak, and Hagen Wierstorf. Towards a binaural modelling toolbox. In *Proceedings of the Forum Acousticum 2011*, 2011. (Cited on page 15.)