Audio Engineering Society

Convention Paper 7367

Presented at the 124th Convention
2008 May 17–20 Amsterdam, The Netherlands

Virtualized Listening Tests for Loudspeakers

Timo Hiekkanen[1], Aki Mäkivirta[2], and Matti Karjalainen[1]

[1] *Department of Signal Processing and Acoustics, Helsinki University of Technology, Espoo, Finland*

[2] *Genelec Oy, Iisalmi, Finland*

Correspondence should be addressed to Timo Hiekkanen (thiekkan@cc.hut.fi)

ABSTRACT

The precise location of a loudspeaker in a listening room is known to affect loudspeaker preference ratings. When multiple loudspeakers are compared the evaluation is limited by the poor human auditory memory. To overcome these problems, a method to evaluate and compare loudspeakers using headphones is proposed. The method utilizes personal head-related transfer functions in rendering the sound field recorded in a standard listening room with an artificial head. Equalization of circumaural headphones and the artificial head are investigated. Formal listening tests are conducted to examine differences between the proposed binaural method and real loudspeakers in a standard listening room. Listening tests show that the virtualized loudspeakers can be nearly imperceptible from reality in many but not in all cases.

1. INTRODUCTION

Traditionally, subjective evaluation of loudspeakers is done in room acoustics, usually in standardized listening rooms. Listening tests are conducted to assess loudspeaker performance or to establish the preference ranking of several loudspeakers. Initially, the performance of the loudspeakers is evaluated by the designer and the final evaluation is made by the consumer when making a purchase decision. However, there are several aspects that can prevent reliable direct comparisons between loudspeakers.

Human auditory memory is short. We cannot accurately remember complex sound images for longer than few seconds. Our long-term auditory memory does not yield solid references and our mood-of-the-day can severely affect the preference ratings if we try to compare a current loudspeaker to a loudspeaker which is not presently at hand.

It is well established that the position of a loudspeaker in a room can strongly affect the perceived sound quality. Also, the room itself affects preference ratings even if the loudspeakers to be evaluated

are placed at the same position. Bech [1] showed that the listening room will influence the perceived differences between loudspeakers in different positions as well as the perceived differences between loudspeakers in similar positions in a different room. Olive *et al.* [2] came to similar conclusions. They found that loudspeaker location was the most significant factor in listener preference ratings.

Unfortunately, what we see is often what we hear. Visual cues can seriously affect the results of listening tests, and should be prevented by using an acoustically transparent curtain.

To achieve reliable and consistent results that can be compared across tests when evaluating loudspeakers in a listening test, all loudspeakers should be evaluated in the same room placed at exactly the same physical position. The time taken to switch between loudspeakers should be small due to the short time span of the human auditory memory. The listener should remain in exactly the same position all the time. Any visual cue should be eliminated. It is difficult to fulfill all these requirements in the real life.

Spatial radiation properties of a loudspeaker are an important part of its fidelity. In anechoic conditions, the direct sound radiating from the loudspeaker to the receiving point determines the properties of a loudspeaker. In room conditions, sound radiated to directions other than the listening direction can make a significant difference. Depending on the loudspeaker and its position, different room modes are excited and the early reflection pattern received at the listening position changes. To evaluate such spatial properties, loudspeakers must be listened to or measured in room conditions.

Binaural techniques have been used to ease the listening test methods and to ensure that the listening conditions are equal for every test subject [3, 2]. Recently, Olive *et al.* [4] have showed that similar loudspeaker preference ratings are achieved with binaural room scanning method and real loudspeakers. Blauert [5] points out the benefits of binaural technology in measurement and evaluation of audio signals. The performance of binaural recordings and the binaural synthesis has been evaluated in numerous studies [6, 7, 8, 9, 10] (for more see [11]). Research has mainly focused on the localization performance of measurements and recordings done at the entrance to a closed ear canal, ignoring issues related to sound coloration.

In the present paper, measurements are done at the entrance to an open ear canal. A method for binaural measurement and synthesis using head-related loudspeaker–room responses is proposed and its use in loudspeaker evaluation task is discussed. Spatial and spectral attributes of the method are compared to real loudspeakers in formal listening tests.

1.1. Binaural Recording, Synthesis, and Reproduction

According to Møller [12], the motivation for binaural techniques is that the input to our hearing system consists of only two signals: the sound pressures at the eardrums. If these signals are recreated precisely, all auditory aspects of an auditory event are repeated perfectly. Headphones are the most practical reproduction device for binaural recordings since they offer almost complete channel separation.

Binaural signals can be recorded either with a head and torso simulator or with a true-head using miniature microphones. Different types of head and torso simulators have been built, starting from spheres with two microphones to full scale replicas of average human upper body. Møller *et al.* have shown that in terms of localization, the best results are always achieved with individual recordings [6]. An artificial head is only an approximation and can not provide good localization and timbre for everyone, and the individual variations in the quality of reproduction are large.

Without compromising the reproduction of spatial information, individualized recordings can be made at any point between the ear drum and few millimeters outside the ear canal entrance. However, three recording positions are of special interest: at the ear drum, at the entrance to an open ear canal and at the entrance to a closed ear canal [12]. Position at the entrance to an open ear canal is chosen here for the following reasons.

- Only the measured headphone response needs to be compensated if the microphones used to measure the binaural responses are small enough not to disturb significantly the sound field at the entrance to an open ear canal.

- Measurement of binaural responses as well as headphone responses is straightforward.
- Measurements at the entrance to an open ear canal give the maximum comfort to test subjects.

The auditory event produced by the loudspeakers can be simulated with headphones if transfer functions from each loudspeaker to each ear and from each headphone terminal to each ear are known. In a stereophonic listening setup, as in Figure 1, proper signals for headphone reproduction are

$$Y_\mathrm{l} = (X_\mathrm{l} H_\mathrm{ll} + X_\mathrm{r} H_\mathrm{rl})/P_\mathrm{l} \quad (1)$$

$$Y_\mathrm{r} = (X_\mathrm{l} H_\mathrm{lr} + X_\mathrm{r} H_\mathrm{rr})/P_\mathrm{r} \quad (2)$$

where Y_l and Y_r are the signals for headphone reproduction, X_l and X_r are input signals of stereophonic reproduction, H represents the transfer functions from loudspeakers to ears as in Figure 1, and P_l and P_r are transfer functions from headphone terminals to ears.

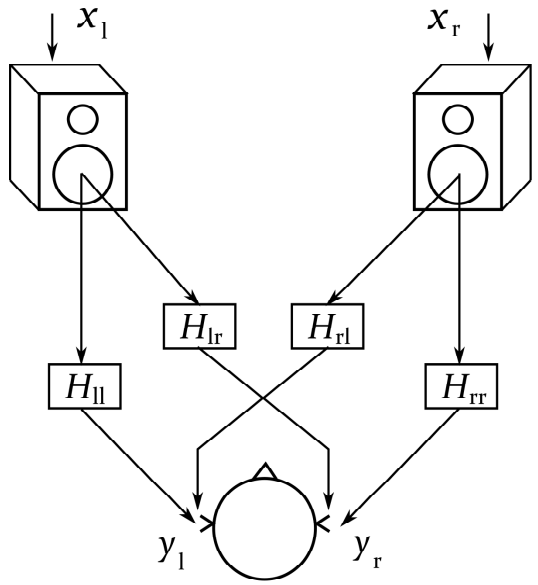


Fig. 1: Transmission paths in the stereophonic listening setup.

Directional hearing of humans is based on inter-aural level differences (ILD), inter-aural time differences (ITD), and spectral cues [13]. If an artificial head provides approximately correct ITD cues, a question arises if the localization properties of an artificial head could be improved by correcting the frequency-dependent ILD cues and spectral cues with an equalizer.

2. MEASUREMENTS

Transfer functions from loudspeakers to ears are needed for binaural synthesis. A series of measurements were conducted in a standardized listening room to understand how repeatable binaural measurements are in room conditions, and to compare true-head measurements with artificial head measurements.

All measurements and processing is performed at 44.1 kHz sampling rate, or if it is not possible, responses are resampled to 44.1 kHz before processing. The head-related spherical coordinate system, where φ denotes azimuthal angle and δ denotes elevation angle, is used. Also, a stereophonic listening setup is used if not mentioned otherwise.

2.1. Equipment

Binaural true-head measurements can be made by attaching small microphones to test subject's ears. Alternatively, a head and torso simulator representing an average human upper body can be used. The artificial head has properties that make it superior to true-head measurements. It can be placed accurately and repeatably. Due to the sensitivity of the room responses to placement differences, the exact placement is essential for comparable results. The microphones of the artificial head are mounted permanently, which removes the variance caused by microphone locations.

The artificial head[1] used in the measurements is made of polyurethane with Nextel coating. The ear shape complies with the IEC 959 and DIN V 45608 standards. Microphones are 1/2 inch condenser microphones positioned at the end of the 20 mm long ear canal. The microphone signal is transferred through an AES/EBU connection.

[1] Manikin MK1 by 01dB-Metravib

Small electret microphone capsules[2] were used in the true-head measurements. The diameter of the capsules is 4.75 mm and height is 4.2 mm and the manufacturer promises flat frequency response from 40 Hz to 20 kHz. The capsules were soldered to cables and a thin and solid wire was wrapped around the cable to give support and shape.

A two-channel preamplifier[3] provided polarization voltage for the microphones. The microphones were attached to test subject's head as shown in Figure 2. The wire was twisted to fit behind the ear and tape was applied to relief strain and keep the microphones in place.

Fig. 2: Microphone attached to subject's head.

Circumaural dynamic headphones[4] were used in the measurements as well as in reproduction of binaural synthesis. Measurements were all done in an ITU-R BS.1116 [14] compliant listening room using the logarithmic sine sweep technique [15].

2.2. Artificial Head Measurements

To test the repeatability of artificial head measurements and to find the positioning accuracy needed, the following measurements were done.

The manikin was placed on a chair, head raised to the level of a true-head listener. Precise and repeatable positioning was confirmed with a plumb line hanging from the ceiling and markings on the floor. Distances from each loudspeaker to the plumb line were measured to be 240 cm.

First, the artificial head was moved towards the line between the loudspeakers and binaural responses were measured for every two centimeters from each loudspeaker. Beyond 10 cm, one measurement was made at the 15 cm displacement. Secondly, the artificial head was moved to the left parallel to the line between the loudspeakers one centimeter at a time. Thirdly, the artificial head was rotated horizontally 2.5° at a time from 0° to 10° and measurements were made as earlier.

The measured impulse responses were convolved with stereophonic commercial rock music (Porcupine Tree: 'Trains' from the record 'In Absentia') and monophonic pink noise, and summed as in Eqs. (1) and (2). The results were listened to with the headphones. Fast and seamless switching between different convolutions was enabled using the Pure Data programming environment [16].

As expected, moving the artificial head forward was found to cause less perceivable differences than moving sideways. With music, a 15 cm movement to forward direction provides a difference that is just noticeable. With pink noise, a 10 cm movement is noticeable. Displacement to the side direction causes perceivable differences much faster. A one centimeter sideways displacement is noticeable when listening to pink noise, while a displacement of three to four centimeters is perceivable with music.

Sensitivity to rotation depends highly on audio material. With pink noise, a 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° [13]. However, even a direction change of 10° was found difficult to notice with certain music signals.

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

[2]Sennheiser KE 4-211-2
[3]Unides Design UD-MPA10e
[4]Sennheiser HD590

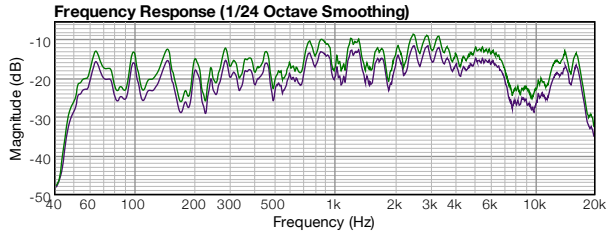


Fig. 3: Responses in the listening room from a loudspeaker at $\varphi = -30°$ to the left ear of an artificial head. The artificial head and the loudspeaker were relocated between measurements. Curves are separated by 3 dB on purpose.

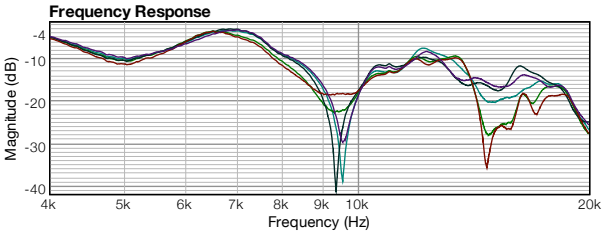


Fig. 4: Five consecutive measurements of headphone transfer functions with the artificial head. Zoomed to frequency range 4 kHz – 20 kHz.

accurately enough to achieve repeatable results. No difference was heard with music or pink noise.

It must be stressed that although these results are based on informal listening by the author they give an idea of how accurate the placement of the the artificial head must be in order to avoid audible errors due to placement differences. The lateral accuracy should be ±1 cm at least, and the forward direction should be well specified. Inaccuracy of placement in frontal direction is not as critical as rotation and lateral displacement but it should not be overlooked. It seems possible to repeat artificial head measurements without perceivable differences between the measurements. Figure 3 shows a typical magnitude response of an artificial head measurement and illustrates the difference between two measurements.

Repeatability of headphone responses of the artificial head was also investigated. Figure 4 demonstrates the repeatability using circumaural dynamic headphones. Responses were measured five times consecutively. Headphones were taken off and put back on between the measurements. Albeit effort was made to place the headphones in the same way, more than 10 dB differences can be seen at frequencies above 7 kHz.

Møller *et al.* have studied headphone responses with human subjects and came to the conclusion that the responses are reliable only up to 7 kHz [17]. Riederer investigated the repeatability of dummy head responses and noted that below 7 kHz responses agree very well [18]. He achieved ±3 dB repeatability up to 13 kHz with circumaural headphones[5].

2.3. True-Head Measurements

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

The location and orientation of subject's head was controlled with a plumb line hanging above the head. The subject was asked to look at a black dot in the front wall and to keep his head still.

Measurements agree very well up to 1 kHz, but above that curves differ. Figure 5 illustrates the differences above 500 Hz. As could be expected, these differences in measured responses are audible if the responses are compared to each other with headphones in the similar way as for the artificial head measurements earlier.

Reasons for the differences are not known, but a few guesses can be made. The microphone locations are probably not exact causing variance in the measurements. The test subject's head cannot be located as accurately as the artificial head and it may move during a measurement. Finally, the human body is a time-varying noise source: blood circulation, breathing and swallowing cause interferences.

To explore the repeatability of true-head headphone responses, five consecutive measurements

[5]Sennheiser HD580

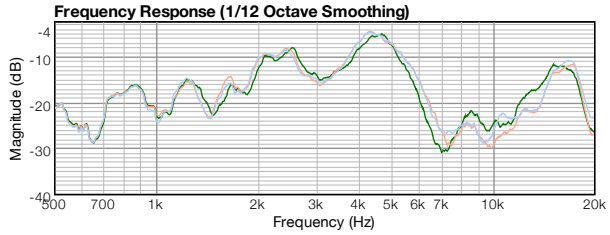


Fig. 5: Comparison of three true-head measurements. Microphones were removed between measurements and the test subject was allowed to move.

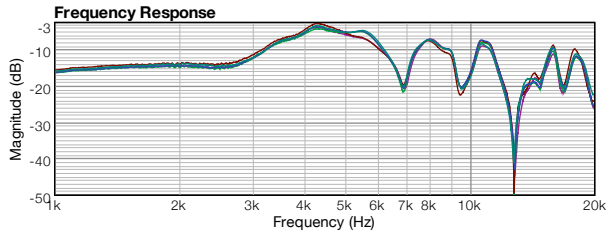


Fig. 6: Repeated true-head headphone response measurements.

were made. Microphones were attached to subject's head by experimenter. The headphones were placed by the test subject, since Møller *et al.* have noted that this produces good repeatability [19]. Figure 6 shows that repeatability appears to be better than with the artificial head. The headphones were taken off and put on a table between the measurements. According to Figure 6, frequency responses are within 3 dB up to the frequency of 13 kHz. Variation is almost constant with respect to frequency in contrast to the artificial head measurements, where much less variation was present at low frequencies.

The true-head headphone measurements seem to be rather well repeatable. However, measurements were all made in one session and an effort was made to place the headphones similarly. Much greater variations are seen if longer pauses are taken, microphones are replaced or the headphones are put on carelessly.

Measurements of several test subjects show that binaural loudspeaker-room responses as well as headphone responses are highly individual. Also, the responses are asymmetrical, meaning that the left and right ear responses differ.

2.4. Conclusion from Measurements

The measurements indicate that true-head measurements alone cannot be used to compare loudspeakers in a stereophonic listening setup. Differences between measurements can be greater than differences between loudspeaker responses.

Loudspeaker–room measurements using an artificial head are repeatable but cannot be used since the artificial head cannot offer correct localization and timbre for everyone due to the averaged nature of its responses.

To be able to evaluate and compare loudspeakers in the stereophonic listening setup through headphones, both artificial and true-head measurements were used. The artificial head gives good measurement accuracy and repeatability, and binaural synthesis using true-head responses gives good timbre and correct localization.

3. METHOD

To use responses from an artificial head instead of individual true-head responses, the artificial head responses must be equalized to match with the individual true-head responses. Figure 7 shows the difference between a true-head response and an artificial head response from a loudspeaker to the right ear in anechoic conditions. As can be seen, responses agree only below 1 kHz. This could be expected since the artificial head has microphones at the ear drum position.

In theory, artificial head responses can be used together with individual true-head responses as in Eqs. (3) and (4)

$$Y_\mathrm{l} = \left(X_\mathrm{l} \frac{H_\mathrm{ll}^\mathrm{ref} G_\mathrm{ll}}{G_\mathrm{ll}^\mathrm{ref}} + X_\mathrm{r} \frac{H_\mathrm{rl}^\mathrm{ref} G_\mathrm{rl}}{G_\mathrm{rl}^\mathrm{ref}} \right) \cdot \frac{1}{P_\mathrm{l}} \qquad (3)$$

$$Y_\mathrm{r} = \left(X_\mathrm{l} \frac{H_\mathrm{lr}^\mathrm{ref} G_\mathrm{lr}}{G_\mathrm{lr}^\mathrm{ref}} + X_\mathrm{r} \frac{H_\mathrm{rr}^\mathrm{ref} G_\mathrm{rr}}{G_\mathrm{rr}^\mathrm{ref}} \right) \cdot \frac{1}{P_\mathrm{r}} \qquad (4)$$

where H^ref and G^ref refer to true-head and artificial head measurements of a reference loudspeaker, G refers to an artificial head measurement of a loudspeaker to be evaluated, P refers to headphone responses, and Y, X, and indices are as in Figure 1.

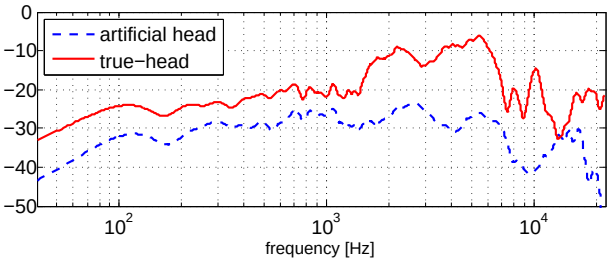


Fig. 7: The magnitude responses of the artificial head and a test subject in anechoic conditions. The measurements are from right ear, loudspeaker being at $\varphi = +30°$ angle. Different resonances can be seen at high frequencies.

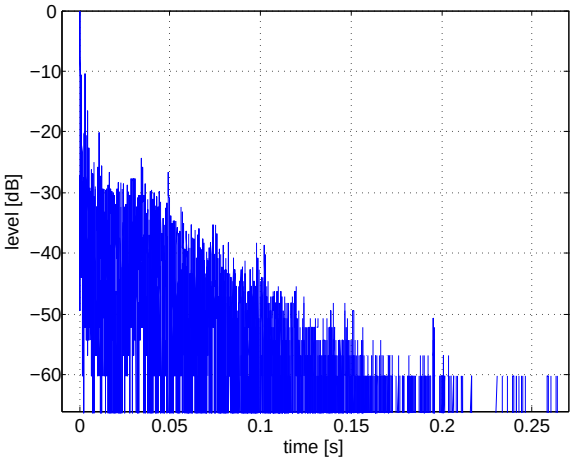


Fig. 8: 12000 first samples of an ipsilateral true-head response squared and plotted on the logarithmic scale.

The problem is to design filters $H^{\text{ref}}/G^{\text{ref}}$, which equalize the artificial head responses to match with individual responses, and individual headphone equalizers $1/P$.

3.1. Equalization of Artificial Head Responses

There is always some noise in the measured impulse responses. The noise is not part of the loudspeaker–room–head transfer function and makes it difficult to determine where the magnitude of the transfer function becomes insignificantly small from auralization point of view. Because of this binaural responses should be truncated to use them in binaural synthesis. Here, all responses are truncated before other processing. Figure 8 demonstrates the SNR achieved in true-head measurements. The starting point of a response was decided based on a fixed threshold. The responses were faded linearly to zero at the point where the signal fell under the estimated noise floor. Truncation of the responses was not considered to be critical since the SNR was good (around 60 dB).

To design the correction filters $H^{\text{ref}}/G^{\text{ref}}$, the magnitude information is used, and a minimum phase impulse responses are created. This enables smoothing of responses in the frequency domain. It is advantageous since the target was to equalize the general shape of artificial head responses but not the individual room resonances. Use of Kautz filters [20][21] was investigated shortly, but the minimum phase design method was selected for its flexibility and easiness. Complex smoothing technique [22] could give one starting point, but it was not investigated here.

In the minimum phase method, 32768 point magnitude responses of the true-head and the artificial head responses are smoothed in the frequency domain using a moving hanning window. The smoothed true-head magnitude response is divided by the smoothed magnitude response of the artificial head. Different window lengths from 1/24 octave to 1/2 octave were tested and in preliminary listening, 1/4 octave smoothing was found to perform well. A minimum phase time domain response is created and the resulting impulse response is truncated after decay of 60 dB. Figure 9 shows the magnitude response of a typical correction filter achieved by the minimum phase method.

The minimum phase method does not result in imperceptible difference between equalized artificial head responses and true-head responses but on the other hand, there is no risk of annoying resonances since the magnitude response of the minimum phase filter is smooth as in Figure 9.

3.2. Headphone Equalization

Equalization of headphone transfer functions is critical in relation to colorations in binaural reproduction. Headphones are equalized to produce a flat frequency response at the physical location of the

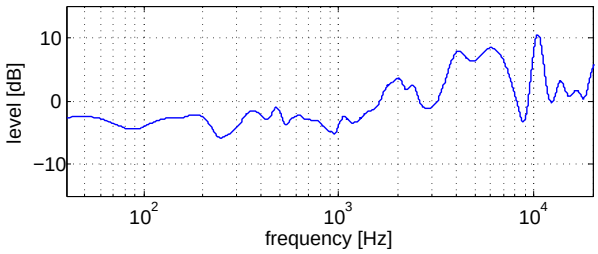


Fig. 9: The magnitude response of a filter achieved by the minimum phase method.

binaural measurement, in our case at the entrance to an open ear canal.

In theory, it is sufficient to design an inverse filter, $1/P$. However, direct inversion of the magnitude response does not provide an optimal solution because of the variance in frequency response produced by headphone placement inaccuracy. At high frequencies, magnitudes and frequencies of the resonances vary from one measurement to another depending on the position of headphones.

According to Bücklein, peaks in the magnitude response should be avoided. A peak is more audible in the reproduction than a corresponding dip [23]. In addition, Toole and Olive state in [24] that wide resonances are detected more readily than narrow peaks. Two guidelines can be now formulated for headphone equalization.

1. Avoid high peaks, especially the wide ones.
2. Do not widen the existing peaks and dips if possible.

The first requirement implies that peaks in the inverted magnitude response of the headphone transfer function should be compressed to ensure that there are no peaks above the average level in the equalized response. The second requirement implies that the inverted magnitude response should not be smoothed excessively since the smoothing widens resonances, and on the other hand it flattens notches, which are needed to compensate for the peaks of the headphone transfer function in the reproduction phase.

The proposed method for the headphone equalization is as follows. The measured headphone response is truncated to 512 samples. The magnitude of a one-sided, 4096 point spectrum of the headphone response is smoothed to remove small variations caused by noise. 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 10, a typical result of the inversion of the magnitude response of the headphones is shown. The dashed curve in Figure 10 represents the smoothed and inverted response. After smoothing, a reference level for peak reduction is decided based on the average level of the inverted response from 40 Hz to the frequency of the first minimum magnitude value below 4000 Hz.

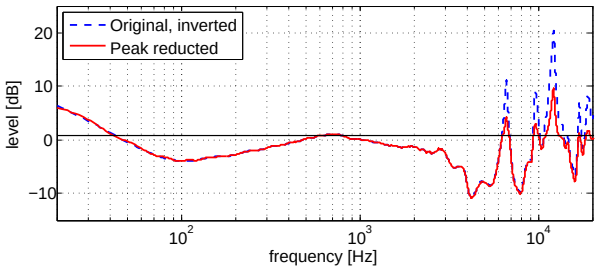


Fig. 10: A typical inverted headphone response shown by dashed line. Horizontal line indicates the peak reduction level. Solid line is the inverted headphone response after peak reduction.

Magnitude values exceeding the peak reduction level are compressed. Compression is active only above the frequency of the minimum magnitude value below 4000 Hz. Amplitude values exceeding the peak reduction level are multiplied by a compression ratio. For instance, if peak reduction level is l, amplitude value of a specific frequency is a and the compression ratio is r, the result for a specific frequency would be $(a - l) \cdot r + l$. In informal listening, 1/4 compression ratio was found to give good results. The effect of the peak reduction is shown in Figure 10. Peaks at high frequencies are taken down from five to ten decibels. A slight frequency roll-off starting

from 4000 Hz was designed to the inverse filter. The roll-off was used to compensate the sharpness caused by unsuccessfully equalized resonances.

Finally, a minimum phase impulse response is created and truncated after decay of 60 dB. An individual filter is designed for each ear. It is strongly recommended that the headphone response is measured in the same session where the binaural loudspeaker-room responses are measured. Remounting the measurement microphones can shift the resonance frequencies significantly, resulting in improper headphone equalization for a specific set of binaural loudspeaker-room responses.

3.3. Equalization Method

In this section, a new method for subjective loudspeaker evaluation was proposed. The method trades problems of loudspeaker placement and loudspeaker swapping for problems related to measurement and equalization accuracy. To sum up, the method consists of the following steps:

1. Using a reference loudspeaker pair, true-head loudspeaker-room responses are measured at the entrance to an open ear canal.
2. Headphone responses are measured with the same microphone placements. The individual inverse headphone filters are calculated.
3. Using the reference loudspeaker pair the artificial head is measured in the same position where the true-head measurements were made.
4. All the loudspeakers to be listening-tested are measured in the similar manner using the artificial head. More loudspeakers can be measured later, if the measurement position as well as loudspeaker positions are well documented.
5. Four individual filters are calculated to equalize the artificial head responses to the true-head responses.
6. Preliminary loudness normalization of the equalized responses is done.
7. Convolutions between the equalized artificial head responses and test signals are calculated and headphone equalization is done using the pre-calculated filters.

To allow changing of responses and to improve flexibility, the impulse responses of the equalized artificial head responses and the impulse responses of the headphone equalization filters are stored separately. A graphical programming environment [16] was used to create a program which performs realtime convolutions needed for binaural reproduction of any audio material. Parallel convolutions enable seamless switching between different responses. Four pairs of virtual loudspeakers can be compared without a delay of switching loudspeakers or any physical arrangements. In the demonstration system, program was run on a laptop computer[6] with 2 GHz dual core processor[7] and firewire audio interface[8]. The four parallel binaural convolutions (16 channels of convolution altogether) for 32768 sample impulse responses took about 65 % of processor time.

4. LISTENING TEST: COMPARISON TO REALITY

To explore the differences between the virtual and real loudspeakers, formal listening tests were organized. The aim was to understand, how the binaural reproduction differs from real loudspeakers in a real room and to evaluate differences between auralization using true-head responses (true-head method) and auralization using individually equalized artificial head responses (artificial head method).

The task given to the test subjects was to evaluate differences in reproduction in terms of five attributes: *apparent source width, direction of events, distance to events, spaciousness, and tone color*. The three first attributes 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 auditory event appears to originate, and distance to events is the distance from the listening position to the point where the auditory event appears to happen. Spaciousness describes the amount of space present in the listening. Tone color describes the spectral content of the sample.

The test subjects were asked to rate the difference between real and virtual loudspeakers using the verbally anchored ITU small impairment scale from 1

[6] MacBook
[7] Intel Core 2 Duo
[8] MOTU Traveler

to 5 [14]. The anchor points and the verbal descriptions are shown in Table 1. The test subjects were able to set the difference rating in increments of 0.1 point.

Table 1: ITU small impairment scale.

Grade	Impairment
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The listening tests were conducted in an ITU-R BS.1116 [14] compliant listening room. Two pairs of studio monitoring loudspeakers[9] were used in the test. Loudspeaker placement was controlled with plumb lines hanging from the ceiling.

4.1. Test Subjects

Seven males and one female subjects participated in the test. Seven of the test subjects reported no hearing damages; one subject reported continuous tinnitus. Although all test subjects cannot be considered experts in loudspeaker evaluation, all had at least some experience in participating in listening tests.

4.2. Samples and Processing

Three different audio test signals were used in the test. *Anechoic male speech*, moving slowly from left to right and back to left, gave an easy way to evaluate the directions and discolorations since the human hearing is specialized to analyze speech signals. A forty second excerpt of *a jazz song* (Screen Play on 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, was used as the most critical test signal for evaluating the sound discolorations.

The test signals were auralized using the truncated true-head responses and the individually equalized artificial head responses. The responses of Loudspeaker A were used to design the equalizers for Loudspeaker B and vice versa, resulting six different test cases for one loudspeaker pair. Table 2 shows the different cases. Each case was repeated once and all attributes were rated.

[9] Genelec 1030A and Genelec 8030A

Table 2: The different samples in the test.

	Loudspeaker A		Loudspeaker B	
	method		method	
speech	true	artificial	true	artificial
music	true	artificial	true	artificial
noise	true	artificial	true	artificial

4.3. Test Procedure

The test was divided into four sections. First, the experimenter attached the microphones on test subject's head and measured the binaural true-head impulse responses in the 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 were given. The test subject was advised not to pay attention to possible loudness differences or background noise, to keep his/her head still and to look forward when listening to the virtual loudspeakers through the headphones. The listening order of headphones first and real loudspeakers then was recommended but not forced. The test subject was allowed to familiarize with the material and to experiment switching between the virtual and real loudspeakers. Processing of the measured responses 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 taken between the two parts and the loudspeakers were switched.

In the evaluation phase the test subject rated the difference of one virtual loudspeaker pair and one real loudspeaker pair using a computer mouse and the

user interface shown in Appendix A. The test subject could switch between the headphones and the loudspeakers at any time. Pressing the play button started the audio clip from the beginning but switching between the virtual and real loudspeakers was instantaneous. The test subjects were able to adjust the volumes of the virtual and real loudspeakers to equalize the loudnesses and was advised to do so if perceived loudnesses were not the same. 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 was an extra case for test subject training only, and it was excluded in the analysis.

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 obtained.

4.4. Results

The received data was analyzed using the multi-way analysis of variances (ANOVA) and multiple comparison tests. 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 [25].

In Figure 11, the means and 95% confidence intervals for each attribute, test signal, and processing method are shown.

True-head responses and equalized artificial head responses worked well for the speech signal. Apparent source width, direction, and distance of events are all rated above 4.5, which corresponds to imperceptible on the ITU small impairment scale. Spaciousness and coloration lie between 4 and 4.5 (perceptible but not annoying). Although the means of the artificial head method are worse, the differences are small (< 0.1) and the confidence intervals of the true-head and artificial head results overlap.

All attributes get lower ratings with music and noise signals. With the music signal, the difference to reality is rated as perceptible but not annoying. The difference between the auralization methods is greatest in terms of distance to events but the confidence intervals overlap. With the noise signal all attributes except tone color are above 3.5 corresponding to perceptible but not annoying. In terms of coloration, the difference to reality was rated as slightly annoying.

The six main effects in the ANOVA 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. A 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 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.

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 of Figure 11 it was concluded that there is no perceptual difference between the true-head method and the artificial head 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.

The significant ($p < 0.01$) second-order interactions in the ANOVA table were sample*attrib, sample*subj, attrib*subj and repet*subj. Figure 11 confirms the sample*attrib interaction. The three other

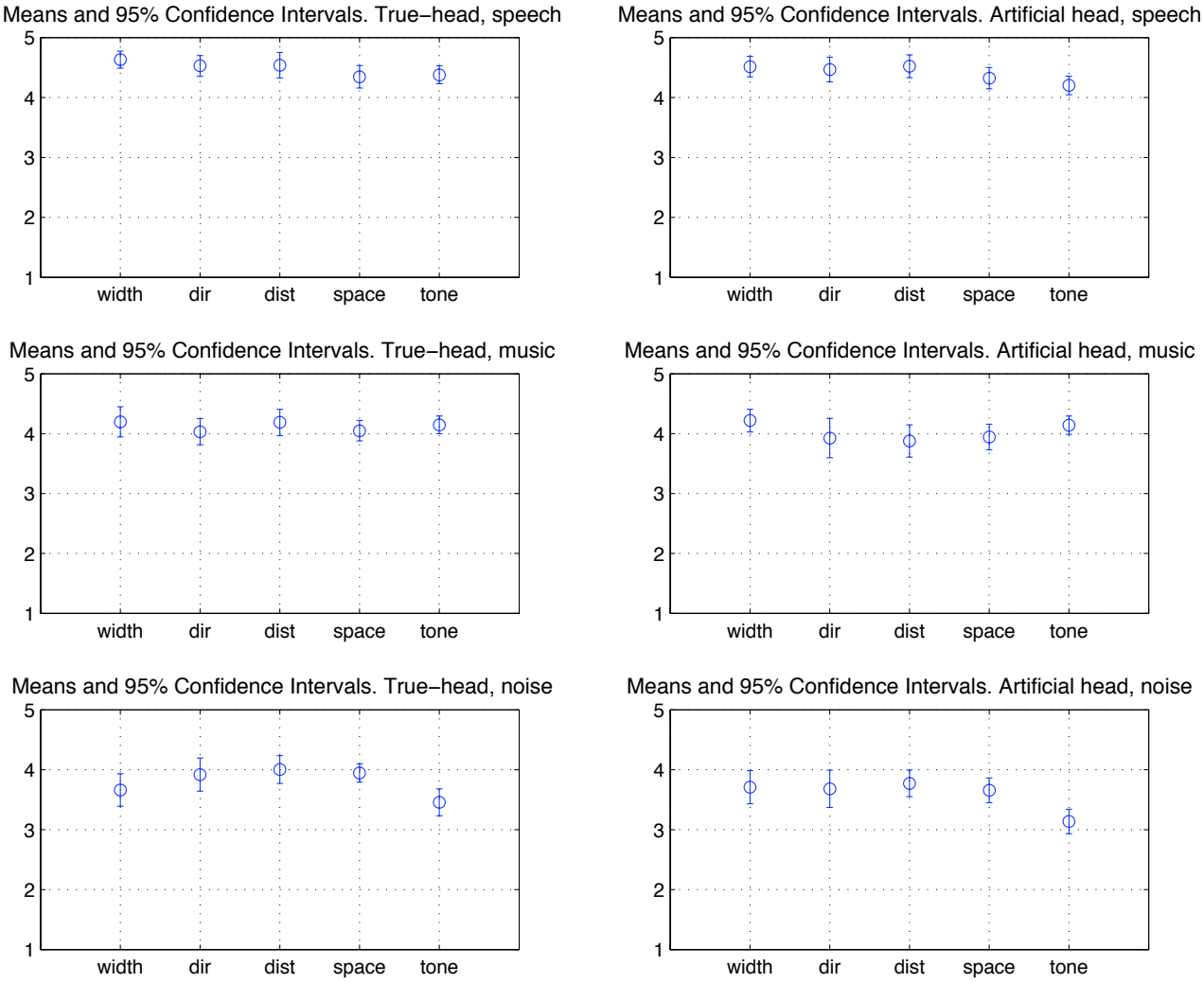


Fig. 11: The means and 95% confidence intervals. The data from both loudspeaker pairs is combined. Y axis scale refers to the ITU small impairment scale.

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.

5. SUMMARY, DISCUSSION AND CONCLUSIONS

In this paper, repeatability of true-head and artificial head measurements was investigated in room conditions. Repeatability of headphone transfer functions for an artificial head and a true-head was studied. All true-head measurements were done at

the entrance to an open ear canal. A method for individual equalization of artificial head responses and headphone equalization of binaural reproduction were proposed. The performance of binaural synthesis using the proposed methods was compared to loudspeakers in a real room in formal listening tests. It was found that the performance depends highly on the test signal used.

The strong dependence of the ratings on the test signal is probably connected with the measurement and equalization inaccuracies at high frequencies. Most of the speech signal energy is below 4 kHz. Above this frequency the headphone equalization is not exact. There is more energy at high frequencies in the music and noise signals, which may lead to the audible differences observed in direct comparison with real loudspeakers. One explanation for the differences can be found from the well-known problems of binaural techniques [3]. 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.

Many of the test subjects were astonished by the quality of externalization of the male speech. All critical comments from the test subjects were related to differences in high frequencies. Either the sound color was too bright or sharp or the high frequencies were not located correctly. This is in agreement with the authors' subjetive findings that the limiting factor in the comparison to reality is the coloration. Unnaturalness of the sound color decreases the usability of the binaural method since it draws attention away from other attributes of sound.

Sound coloration could possibly be reduced if it was possible to place headphones exactly similarly every time. With circumaural headphones it is difficult, but with intra-aural headphones it might be possible. The use of intra-aural headphones would make the measurement procedure much more complicated since the measurement microphones should be inserted inside the ear canal.

The listening test indicates that it is possible to use equalized artificial head responses instead of true-head responses in binaural synthesis. Individual equalization of the artificial head responses improves spatial properties and reduces sound coloration of the binaural synthesis close to those of binaural synthesis using true-head responses. To assess this difference in detail, an experiment should be organized to compare the performances of these methods.

The proposed listening test method trades the problems of traditional loudspeaker listening tests for the problems related to the measurement accuracy and equalization of the virtualized listening setup. Nonlinear properties of loudspeakers cannot be represented by the virtualized listening test method. Since there are no dynamic localization cues in the method used, and the binaural responses represent only a single position in the listening room, the virtualized method is not recommended as the sole method to evaluate loudspeakers. Also, virtual loudspeakers should not be compared to real ones. Instead, all loudspeakers to be compared should be virtualized for the comparison task. Virtual loudspeakers are comparable in the sense that the same processing is done to all virtual loudspeaker pairs.

6. REFERENCES

[1] S. Bech. The influence of the room and of loudspeaker position on the timbre of reproduced sound in domestic rooms. *12th International Conference of the Audio Engineering Society*, May 1993.

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

[3] F. 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] S. E. Olive, T. Welti, and W. L. Martens. Listener loudspeaker preference ratings obtained in situ match those obtained via a binaural room scanning measurement and playback system. *the 122nd Convention of the Audio Engineering Society, preprint no. 7034*, May 2007.

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

[6] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings. *Journal of the Audio Engineering Society*, 44(6):451–469, June 1996.

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

[8] H. Møller, D. 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.

[9] 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.

[10] 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.

[11] J. Blauert, editor. *Communication Acoustics*. Springer, 2005.

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

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

[14] 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.

[15] A. Farina. Simultaneous measurement of impulse response and distortion with a swep-sine technique. *the 108th Convention of the Audio Engineering Society (AES), preprint 5093*, 2000.

[16] 2007. URL `http://puredata.info/`.

[17] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen. Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society*, 43(4):203–217, April 1995.

[18] K. 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.

[19] H. Møller, C. B. Jensen, D. Hammershøi, and M. Friis Sørensen. Design criteria for headphones. *Journal of the Audio Engineering Society*, 43(4):218–232, April 1995.

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

[21] M. 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.

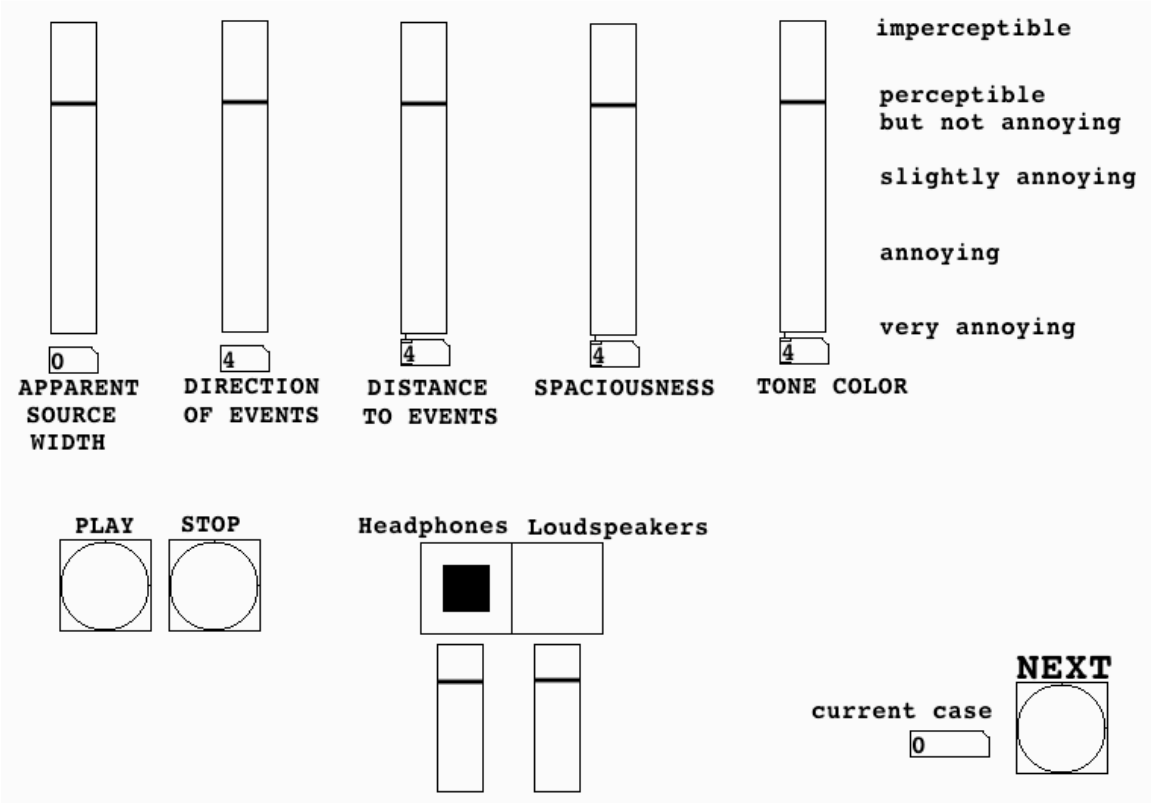
[22] P. D. Hatziantoniou and J. H. Mourjopoulos. Generalized fractional-octave smoothing of audio and acoustic responses. *Journal of the Audio Engineering Society*, 48(4):259–280, April 2000.

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

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

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

APPENDIX A. USER INTERFACE OF THE LISTENING TEST



APPENDIX B. 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