A motion-tracked binaural microphone array for recording spatial sound

Technische Universität Berlin
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation

Masterarbeit

vorgelegt von Felicitas Fiedler

[redacted]

Berlin, 9. Juni 2018

Erstgutachter: Prof. Dr. Stefan Weinzierl
Zweitgutachter: David Ackermann

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, 9. Juni 2018

Felicitas Fiedler

Abstract

The motion-tracked binaural (MTB) technique allows the pseudo binaural recording and reproduction of spatial sound scenes. The recording device consists of a spherical microphone array with a certain number of microphone capsules on its horizontal circumference. Using a head tracking system and a rendering software, a dynamic sound scene can be reproduced via headphones. The system can also transmit signals in real time. A 16-channel MTB array with omnidirectional diffuse field corrected electret condenser microphone capsules was built at the Technical University Berlin. It is also to be perceptively evaluated in two listening tests. First the plausibility of an 8- and 16-channel MTB recording with real signals as reference was tested. The analysis of the obtained values from a yes/no pardigm was done by the signal detection theory. Furthermore, the MTB sound quality was investigated by means of various items from the Spatial Audio Quality Inventory (SAQI) compared to a dynamic binaural synthesis as reference. In terms of plausibility, the results show that the detection rate of a 16-channel MTB rendering with optimal interpolation is 62 %, compared to a guessing rate of 50 %. When evaluating the perceptual qualities with the SAQI, there were significant differences between the MTB signal and the binaural reference in a change in tone color. It should be emphasized, however, that the degree of externalization and the elevation of the sources were approximately at the same level as the dynamic binaural synthesis.

Zusammenfassung

Das Motion-Tracked Binaural (MTB) Verfahren ermöglicht die pseudo-binaurale Aufnahme und Wiedergabe räumlicher Schallfelder. Das Aufnahmegerät besteht aus einem kugelförmigen Mikrofonarray, auf dessen horizontalen Umfang eine bestimmte Anzahl von Mikrofonkapseln eingelassen ist. Mit Hilfe eines Head-Tracking-Systems und einer Rendering-Software kann eine dynamische akustische Szene über Kopfhörer wiedergegeben werden. Mit diesem System ist es außerdem möglich, Signale in Echtzeit zu übertragen. Am Fachgebiet Audiokommunikation der Technischen Universität Berlin wurde ein 16-kanaliges MTB-Mikrofonarray mit omnidirektionalen Elektret-Kondensator-Mikrofonkapseln gebaut. Anschließend wurde das Array diffusfeldentzerrt und die Wiedergabequalität in zwei verschiedenen Hörversuchen perzeptiv evaluiert. Zuerst wurde die Plausibilität einer 8- und einer 16-kanaligen MTB-Aufnahme mit realen Signalen als Referenz getestet. Die Auswertung der erhaltenen Werte aus einem Ja/Nein-Pardigma erfolgte durch die Signalentdeckungstheorie. Darüber hinaus wurde die MTB-Klangqualität anhand verschiedener Items aus dem Spatial Audio Quality Inventory (SAQI) im Vergleich zu einer dynamischen Binauralsynthese als Referenz untersucht. Im Hinblick auf die Plausibilität zeigen die Ergebnisse, dass die Erkennungsrate gegenüber einer Ratewahrscheinlichkeit von 50 % bei einer 16-kanaligen MTB Wiedergabe bei 62 % liegt, sofern mit optimaler Interpolation gearbeitet wird. Bei der Auswertung der perzeptiven Qualitäten mit dem SAQI zeigen sich signifikante Unterschiede zwischen den MTB-Signalen und der binauralen Referenz in einer Veränderung der

Klangfarbe. Hervorzuheben ist jedoch, dass der Grad an Externalisierung und die Elevation der Quellen in etwa auf gleichem Niveau wie die dynamische Binauralsynthese liegen.

Contents

PART I - Paper

1. Introduction

One ability of human hearing is to give a three-dimensional character to sound sources [1]. Both monaural and interaural cues are important for the directional localization of sound sources [2, 3, 4]. For the localization of sound events on the median plane, monaural spectral cues, introduced by pinna and torso, are evaluated [5]. The spectral influence of the pinna above 3 kHz provides the most important elevation cues [6]. On the horizontal plane the interaural difference cues are exploited for the localization of sound sources [4]. The interaural time difference (ITD) is analyzed in the low frequency range up to about 1.5 kHz, the interaural level difference (ILD) in the corresponding high frequency range, as the duplex theory of Lord Rayleigh [2] states. These properties can be exploited for the simulation of virtual acoustic environments (VAEs). The acoustic signals are played back via headphones in a way they exist at the listeners' ears in reality and thus create an immersive listening experience. Important for stable sound localization and to reduce front/back confusion, which can occur with static signals, are the motion cues. They result from tracking the listeners' head movements [7].

A method for recording and reproducing spatial sound fields is the motion-tracked binaural technique, published by Algazi et al. [2] in 2004. The recording device consists of a rigid sphere with a diameter, which is approximated to a human's head. On the horizontal circumference a certain number of microphones is embedded evenly distributed in the surface of the sphere. Head tracking allows to determine the ear positions of the listener on the sphere relative to the microphone positions. Using interpolation algorithms, a stereo signal can be reconstructed from the multi-channel recording. If one ear position coincides exactly with a microphone position, the signal is transmitted directly to the headphones. However, if the ear position is between two adjacent microphones, the signal is first interpolated. The MTB method enables a dynamic virtual reproduction of a real sound environment and can also transmit signals in real time. It can also be exploited using synthetically generated signals by convolving MTB recorded room impulse responses with anechoic audio. Particularly, the proposed method does not rely on head-related transfer functions (HRTFs).

It thus offers an alternative to the dynamic binaural synthesis. For this, binaural room impulse responses (BRIRs) are measured with an artificial head for different head orientations. In the frequency domain they describe the HRTFs of a signal between the sound source and the listener's ear. When convolving the BRIRs with an anechoic audio signal, they give all necessary informations about room and source locations it [8]. Using a head tracking system, the head position of the listener can be identified. A real time convolution software can be used to generate a high-resolution, dynamic auralization that creates an immersive sound image. The binaural signals are played back via headphones. With this method, it is possible to simulate a variety of different source positions [9]. In contrast to the MTB method, this system does not allow a dynamic reconstruction of time variant acoustic scenes and a dynamic real time transmission.

To investigate the quality of a virtual acoustic environment, the authenticity and plausibility criteria can be evaluated. The authenticity criterion examines the perceived identity between a real (as an external reference) and a simulated sound field [3], whereas the plausibility is determined by means of an internal reference, i.e. the expectation of a real sound field [10]. The perceived plausibility of an MTB microphone array was already tested in [11], but the study was performed without a reference as suggested in [10].
The aim of our study is now to test the plausibility of a motion-tracked binaural microphone array with real stimuli as reference in a yes/no paradigm. The evaluation of the test is made by the signal detection theory. Since in [10] a dynamic binaural synthesis has already been investigated with regard to its plausibility, for a further comparison the MTB sound quality is also to be examined in a second listening test using various items from the Spatial Audio Quality Inventory (SAQI) [12] with a dynamic binaural synthesis as reference. For evaluating the results, the median values of all subjects' ratings with 95 % bootstrap confidence intervals are presented. Previously, for this purpose, a MTB recording device was developed according to [2] at the Technical University Berlin, involving the implementation of a diffuse field equalization filter.

2. Method

2.1. Array design and manufacturing

The MTB microphone array has the shape of a sphere that approximates the size of a human's head. The calculation of the diameter $d = 17.6$ cm was based on a study by Algazi et al. [13] and the anthropometric dimensions from DIN 33402-2 [14]. Too large deviations from the individual head size can, according to [15], lead to erroneous effects in the sound localization. The spherical body was produced by the additive manufacturing technique selective laser sintering (SLS). Here a powdery starting material is fused layer by layer with a laser to a spatial structure, which was described before with a virtual CAD model [16]. Due to the used thermoplastic material and the subsequent painting, the body has reverberant properties. The sphere consists of an upper and a bottom shell as well as a horizontal center ring, which was placed between the shells. 16 microphone capsules are evenly spaced on this ring, so that they complete with the surface of the sphere. For a better grip, the capsules were previously covered with an elastic ring. In the interior of the spherical body there is also a mount for the converter circuit boards of the microphones, which is also manufactured by using the SLS method and shielded by a steel sheet case lying around it to prevent influences of electromagnetic waves. The incoming signals are balanced at this point. The integrated voltage converter, downshifting the voltage level to the operating voltage required for the capsules, allows using the array with a phantom voltage of 12 to 48 V. The microphone array is connected to a 16-channel multicore cable at the lower shell of the sphere with as many XLR connectors. To fix the sphercial array, a 3/8-inch thread is attached at the bottom, which allows to be screwed onto a standard microphone stand. Omnidirectional electret

condenser capsules from Sennheiser (KE14) operating at 9 V were used for the array. They have a diameter of 14 mm and were made available in cooperation with Sennheiser electronic GmbH & Co. KG for the project. The selection of the microphone capsules was made by a matching method in which a set of capsules, suitable for the MTB microphone, was selected. For this purpose, the amplitude responses of individual capsules measured by Sennheiser were compared. The aim was to find 16 capsules that differ in amplitude less than 1 dB over the entire frequency range.

After assembling the recording device, the frequency responses of the installed microphones were measured in an anechoic chamber at Georg Neumann GmbH in Berlin. For calibration, a reference measurement was conducted on a Microtech Gefell MK202 microphone to determine the influence of the measurement system. The sound source was a Neumann coaxial loudspeaker system. A motorized device enabled an accurate measurement of the array. The excitation signal was a continuous sine sweep. During the measuring process, the sphere was turned on its own axis. The frequency response of each microphone was measured at a resolution of 2.8125° cw. Afterwards the frequency responses were adjusted with the correction data of the calibration microphone. For audio recordings in the diffuse field, a diffuse field equalization filter for the MTB array has been implemented on this basis. Due to the symmetrical properties of the sphere, only the measuring data between 0° and 180° on the horizontal plane was considered for the calculation of the filter. In a first step, a mean amplitude response was calculated for each position in a resolution of 0°:2.8125°:180° cw over all 16 microphone capsules. Next, the amplitude responses were multiplied by various weighting factors prior to further data processing. The sphere is scanned by discrete points. Each point is represented by an surface area that describes the mean amplitude response of this point. Since the area at 90° has the largest size, the amplitude response has the strongest influence at this measuring point and is therefore heavily weighted. To 0° and 180° the area sizes as well as the weighting factors decrease. If the factors for all measuring points between 0° and 180° are summed, the result is 1. Subsequently, an averaging of the data in all directions was executed. In order to obtain the equalization filter, the inverse of the amplitude response has been generated. The black curve in Figure 1 shows the raw data of this filter. It has a large boost at 21 kHz by about 23 dB. In addition, since the noise in the signal would be amplified too much at this point, the filter was compressed in a ratio of 4:1 and with a threshold of -23 dB. The compressed filter, represented by the red curve, is the best compromise between tone coloration and signal-to-noise ratio.

2.2. Recording system und player software

For recording sound with the MTB microphone, 8 channels each are connected to the XLR inputs of two Presonus Digimax DP88 preamplifiers. The devices supply the microphone array with phantom power and the incoming signals are getting amplified and a/d converted. The preamplifiers are each connected via ADAT with an RME Fireface UFX, synchronizing the devices. With a USB port, the sound card can be connected to a computer running the recording

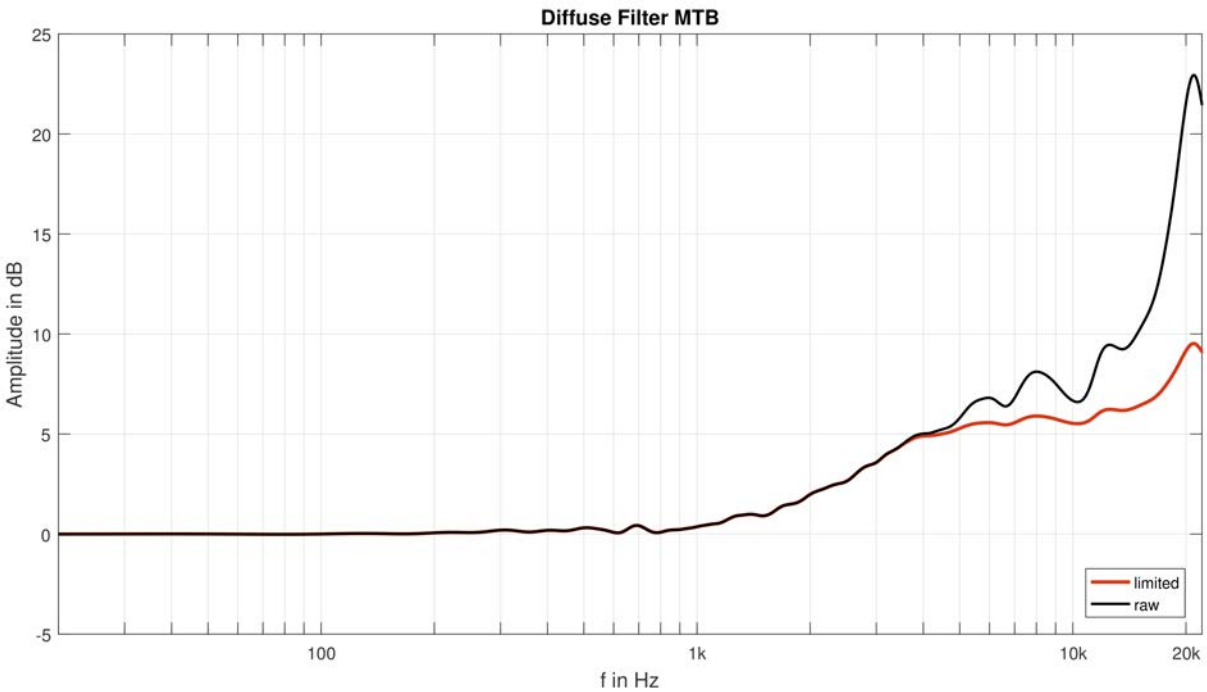


Figure 1: Raw data of the diffuse field equalization filter (black) and compressed filter (red) with a 4:1 ratio and a threshold of -23 dB.

software. The real time MTB Renderer [11] is used as playback software. The software was primarily developed for the Linux system in C++ and has an OSC interface for the tracking data. The MTB Renderer was also designed as a JACK client for real time audio streaming. The graphical user interface allows to set various parameters. On the one hand, the desired number of input channels can be entered. Here it is possible to choose between 8, 16, 24 and 32 input channels (in our case, we can choose between 8 and 16 channels). Furthermore, one of five implemented interpolation algorithms, already introduced in [2], can be selected. The short-time Fourier transform (STFT) size can be used to set the block size of the STFT. It is applied during the interploation in the frequency domain. With the STFT window option, the window shape of the STFT can be selected. At this point you can choose between Hanning, Hamming, Blackmann, Bartlett and Rectangular window. Moreover, the cutoff frequency option can be used to set the crossover frequency for the frequency-band-separated interpolation. In addition to real time streaming, it is also possible to play with the MTB microphone recorded audio via a dedicated player software called MTB Player [11]. The software can be connected to the MTB Renderer with the JACK Audio Connection Kit and can also be controlled by OSC commands. For the transformation back into the time domain, necessary for the interpolation in separate frequency ranges, the weighted-overlap-and-add (WOLA) method is used here [17]. For tracking the head movements of a person a Polhemus Patriot Headtracker is used. The tracking data output as OSC command can be received by the MTB Renderer. For the following listening tests we use the Spectral-Interpolation Restoration, which – as mentioned in section 2.3

– achieved the best results of the five algorithms in terms of perceived plausibility ([11]). Since our plausibility test is performed with real signals for internal reference, we will again investigate an 8- and a 16-channel MTB recording. Furthermore, a Hanning window with 75 % overlap and a block size of 128 samples was defined for both listening tests. The cutoff frequency for the different number of input channels N is calculated according to equation 3. Thus, for $N = 8$ and a radius of $r = 87.5$ mm, we obtain a cutoff frequency of $f_{max} \approx 1250$ Hz. For $N = 16$ we get a cutoff frequency of $f_{max} \approx 2500$ Hz.

2.3. Interpolation methods

Due to the limited number of microphones integrated in the MTB array, the sound field is sampled discretely. For the continuous reproduction of the two ear signals while rotating the head, it is necessary to interpolate signals at every point between two adjacent microphones. [2] presents five different interpolation methods for this purpose, which are briefly presented below. The first two methods interpolate in the time domain, whereas the three subsequent algorithms implement a frequency-band-separated interpolation. In 2010, the quality of MTB recordings was investigated in [11] by evaluating the number of microphones (8, 16, 24, 32), the signal content (noise, music, speech) and the interpolation method, concerning the perceived plausibility without reference. The Spectral-Interpolation Restoration received the best assessments of all interpolation algorithms. An increase in the number of microphones and a distinction in content show minor differences, when using this interpolation method. As this algorithm is used for the listening tests in this study, it will be described in more detail below, as well as the linear interpolation that the algorithm uses in the low-frequency range.
The *Nearest Microphone Selection* is the easiest way to reconstruct the ear signals. The circle is divided into different sectors N, corresponding to the number of microphones in the array. Depending on the ear position, the nearest microphone is selected. This approach leads to signal jumps at the boundaries of the sectors.
The *Full-Bandwidth Linear Interpolation* is a linear interpolation between two adjacent microphones, where $x_n(t)$ is the nearest microphone signal and $x_{nn}(t)$ the second nearest microphone signal. With the interpolation coefficient

$$w = \frac{b}{b_N}, \tag{1}$$

where b describes the angle between the ear and the nearest microphone and b_N the angle between the adjacent microphones, the signal $x(t)$ at the ear position can be calculated with

$$x(t) = (1 - w)x_n(t) + wx_{nn}(t). \tag{2}$$

However, the linear interpolation of two microphone signals can cause comb filter effects and distortions. For this case in [2] was calculated a cutoff frequency f_{max}. Below this frequency,

the comb filter effects can be avoided within a range of ±3 dB. It is given by

$$f_{max} = \frac{Nc}{8\pi r}, \tag{3}$$

where N is the number of microphones and r the radius of the sphere. To cover the entire frequency range, a high number of microphones would be required.
Further methods pursue the approach of applying different interpolation methods in separate frequency ranges. Low-pass filtering can eliminate high frequency artifacts that rely on phase interference. The high frequency components can then be reconstructed separately. The ITD represents the main cue of directional localization on the horizontal plane in the low frequency range up to about 1500 Hz [2]. For this range, the linear interpolation is suitable. The crossover frequency depends on the number of microphones. The methods described below perform a linear interpolation in the low frequency range and follow different approaches for restoring the high frequency signal components.
For the *Fixed-Microphone Restoration*, a high-pass filtered signal of a fixed omnidirectional microphone, also called complementary microphone, is used to reconstruct the high frequencies.
The *Nearest Microphone Restoration*, however, uses a high frequency sectoral interpolation, which transmits the signal from the nearest microphone.
The last in [2] introduced method is the *Spectral-Interpolation Restoration*. Here, a spectral interpolation is performed in the high frequency range. For the transformation into the frequency domain the STFT is applied. The amplitude of the Fourier transform $M_c(\omega)$ of the time signal $x(t)$ from equation 2 can now be defined by

$$M_c(\omega) = (1 - w)M_n(\omega) + wM_{nn}(\omega), \tag{4}$$

where $M_n(\omega)$ is the amplitude of the STFT of x_n and $M_{nn}(\omega)$ the amplitude of the STFT of x_{nn}. The last step is the restoration of the time signal. For this purpose, real time methods can be used, as described in [17].

2.4. Perceptual analysis

The following section focuses on the perceptual evaluation of the MTB playback quality in two listening tests. It describes the methodology, the experimental setup and the performance of both investigations. First, the plausibility of the system is tested with real signals as an internal reference. In the second listening test, the quality is evaluated by means of items from the SAQI [12]. In comparison, a dynamic binaural synthesis is used as a reference. The BRIRs were recorded with the artificial head FABIAN [18], developed by the Audio Communication Group at TU Berlin.

2.4.1. Assessment of plausibility

In the first listening test, the plausibility of the MTB system will be investigated. The plausibility of a virtual acoustic environment is described in [10] as a simulation that corresponds to the expectation of the listener to a real sound field. Within the experiment, the subjects are listening to real and simulated stimuli in an identical experimental environment. The test setup, procedure and evaluation of the listening test were performed as in [10]. Using a yes/no paradigm, the subjects' decisions are queried and the results are analyzed using the signal detection theory (SDT). With the SDT it is possible to measure the sensitivity of a perceived stimulus. The model is used to investigate detection rates of weak signals in the presence of noise. The response behavior is divided into two components: the sensitivity d and the response bias β. The sensitivity describes the person's ability to differentiate between the conditions 'signal' and 'no signal', where the response bias describes the tendency of a person to one of the conditions. In order to find the true detection power, the response bias is excluded.
In our study, the real stimuli represent the pure noise signal and the simulated stimuli the existing stimulus with noise. The SDT observer model assumes a normal distribution with equal variance for the conditions reality and simulation. The sensory difference between the distributions is given by the sensitivity. In a 2AFC paradigm, sensitivity and detection rate P_c depend on each other. The simulation would be perfectly plausible if the sensitivity $d' = 0$ or $P_c = 0.50$ (pure guessing rate). In [10] the plausibility criterion was set to $P_c = 0.55$, whereby $P_c = 0.75$ is the usually targeted detection threshold of a psychometric function as stated there. For testing the hypothesis, the minimum effect hypothesis ($d' = d_{min}$) is considered. This case assumes that there is an effect, but it is perceptually insignificant. The required sample size (the number of individual decisions) N_{opt} depends on the alpha and beta error (type I and II error) of the test as well as the d'_{min} value to be evaluated. Since the simulation in [10] is still considered as plausible for $d' = 0.1777$, for this sensitivity value as well as an alpha error of 25 % and a beta error of 5 % a sample size of $N_{opt} = 1071$ is calculated. Taking into account the representativity of the sample and a minimum of 100 decisions per person (necessary to obtain stable parameters for STD), 11 subjects were used for the listening test, resulting in a sample size of 1100 decisions.
The listening test took place in the Electronic Studio of the Technical University Berlin. The room has a dry acoustic with a reverberation time of $RT \approx 0.26$ s and a volume of $V = 140$ m^3. The room proved to be a suitable experimental environment as it allows the presentation of real and simulated stimuli within one experiment. As a listening position, we chose a central location in the room where the subjects were placed later. Five Genelec studio monitors with different positions (frontal, side, elevated, back) were positioned in the room, directly aligned to the listening position. Figure 2 shows the positions of the loudspeakers and the receiver position. The coordinates of the speaker positions can be seen in table 1. The coordinates are aligned relative to the listener's frontal view (azimuth = 0°, elevation = 0°).
In the first step, the impulse responses for the simulated stimuli were measured for each loudspeaker position with the MTB microphone array at the listening position. To generate the

Table 1: Loudspeaker positions in the plausibility listening test.

Loudspeaker	Azimuth	Elevation	Distance
1	0°	0°	1.56 m
2	−28°	6°	3.69 m
3	−60°	8°	2.65 m
4	45°	48°	1.65 m
5	126°	7°	3.16 m

simulated stimuli, the received impulse responses were convolved with anechoic audio material. The content ranged from male and female speech in different languages to recordings of solo instruments and extracts of pop songs. The signal length varied between 3 and 6 seconds. In total, 20 different stimuli were generated for each speaker position. During the listening test, the participants sat on a chair, which had an individually adjustable neck rest to stabilize the head position and a table that could be used to take off the arms or to guide the computer mouse. On a desk directly in front of the subject, there was a laptop with the graphical user interface of the listening test, which was independently operated by the subjects. An extraaural headphone (BK2/11) [19] was used to present the simulated stimuli. It could be worn by the participants throughout the entire experiment without affecting the presentation of the real signals. For tracking the head movements, a Polhemus Patriot head tracking system was used. The simulated signals were also high-pass filtered with a cutoff frequency of 50 Hz and equalized with the diffuse field filter of the MTB microphone array. Furthermore, they were provided with the head-

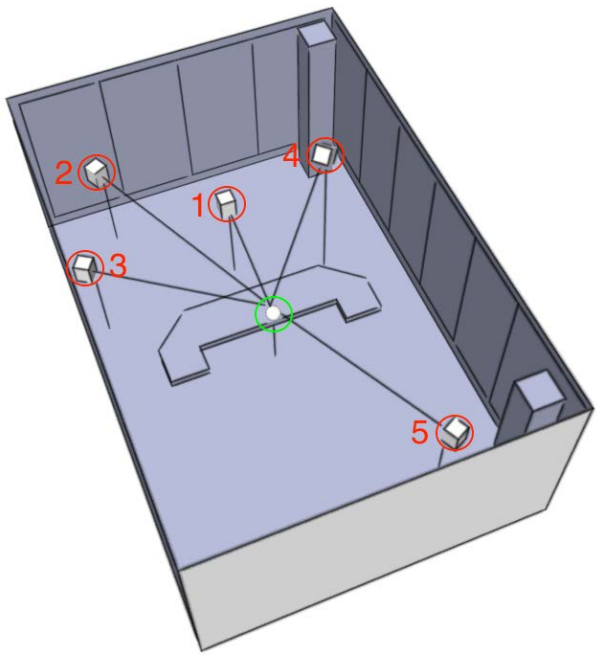


Figure 2: Loudspeaker positions (red) and receiver position (green) in the plausibility listening test.

phone equalization filter and the diffuse field equalization filter of the artificial head FABIAN. The volume level of the real stimuli as well as the volume level between real and simulated stimuli at the listening position for each loudspeaker has been adjusted perceptually in advance, so that the perceived loudness at the listening position was at the same level for all stimuli. During the entire experiment, the participants had to be alone in the Studio, so the control center was set up in a neighboring room. Talkback enabled the communication with the person in the test room. The signal processing was done on an external computer, communicating via OSC with the laptop operated by the subject.
Two different groups were investigated for an 8- and a 16-channel MTB recording. A total of 22 subjects completed the test, with 11 subjects per test group. Each person took about 15 minutes for the test. The participants in both groups had an average age of 29 years (16 male, 6 female) and no hearing defects except for two people with tinnitus. 16 subjects averaged several years of musical education (school, study, music school, job) and can therefore be assessed as experienced listeners. The other 6 subjects did not receive any musical education. In addition, 8 participants have already gained experience with listening tests on dynamic binaural synthesis.
At the beginning, the participants were introduced to the listening test and familiarized with the operation of the user interface. During the experiment, a total of 100 stimuli was presented to the subjects. In this process, the real stimuli were played via loudspeakers and the simulated stimuli via headphones in approximately equal parts. In order to prevent memory effects of perceived differences between real and simulated stimuli, that could influence the response bias of the subjects, content and source position were randomized and combined only once per participant as a real or simulated stimulus. After starting the test through the MATLAB based user interface, the command for the next stimulus to be played was sent via OSC to the external computer. A Pure Data patch controlled the corresponding combination of speaker position, presentation type (real or simulated) and audio content. With JACK Audio connections the real signals were transmitted via MADI to the speakers in the test room. The MTB signals had to pass the MTB Renderer previously, where the tracking data of the head movements is scanned and the signals are interpolated. The output signals of the MTB Renderer are convolved with the equalization filters and the headphone compensation filter and then sent via MADI to the headphones. After playing each stimulus, the participants should decide in a yes/no paradigm whether the played stimulus was a real loudspeaker or not. The next stimulus was played after entering the answer. So the subjects could proceed at their own tempo. Thus, the subjects were previously instructed to focus their full concentration on the current stimulus before playing the next one. The head movements of the listeners could be checked by the test supervisor during the experiment while observing the tracking data.

2.4.2. Evaluation with the SAQI

The Spatial Audio Quality Inventory describes a vocabulary for the perceptual evaluation of virtual acoustic environments. It was developed by a focus group of the Audio Communication

Group at the Technical University Berlin [12] and contains 48 items in 8 categories: Timbre, Tonalness, Geometry, Room, Time behavior, Dynamics, Artifacts and General. The vocabulary exists in english and german. The test interface used for the experiment is a Matlab based software named WhisPER [20]. It includes a special test design with a SAQI questionnaire. In this listening test, the perceptual sound quality of an 8-channel and a 16-channel MTB recording is tested. Since the plausibility of a dynamic binaural synthesis has already been assessed in [10], here the test signals are investigated compared to a dynamic binaural synthesis with the BRIRs of the artificial head FABIAN as reference in means of six items from this inventory for the variables *room*, *source position* and *signal content*, to complete the comparison. The evaluation of the obtained values is done by calculating the median values with 95 % bootstrap confidence intervals for each condition. During the listening test, three different variables should be investigated:

- **Room:** short, long reverberation time
- **Source position:** frontal, top right
- **Content:** speech, noise

It gives a total of eight conditional variations for each test signal per SAQI item. First, the stimuli for the listening test were generated. Therefore, the impulse responses with the MTB array and the BRIRs with FABIAN (with a range of $\pm 40°$ and a resolution of $1°$ on the azimuthal plane) were recorded for two loudspeaker positions in two rooms at the listening position. We opted for a reverberant room (Audimax TU Berlin, $V = 8500$ m^3, $RT \approx 2.0$ s) and an acoustically dry room (Electronic Studio TU Berlin, $V = 140$ m^3, $RT \approx 0.26$ s). In the positioning of the sound sources we decided in both rooms for a frontal and an elevated loudspeaker at the right relative to the frontal view of the listener (azimuth = $0°$, elevation = $0°$). The loudspeaker positions and the receiver position are shown in Figure 3. Table 2 shows the coordinates of the speaker positions.

Table 2: Loudspeaker positions in the SAQI listening test

Speaker	Azimuth	Elevation	Distance	
1	0°	0°	12.7 m	Audimax
2	70°	17°	16.6 m	
1	0°	0°	1.50 m	Studio
2	45°	48°	1.65 m	

According to the mixing time, the dynamic and the static parts of the binaural room impulse responses were separated [22]. For the reproduction of the stimuli, the impulse responses (MTB) and the BRIRs (FABIAN) were convolved with anechoic audio material. Here we chose a male speaker and a white noise signal. From the SAQI vocabulary, we selected the following six in [12] described items with opposite scale poles for the test, which proved to be critical for the evaluation of the MTB system in advance:

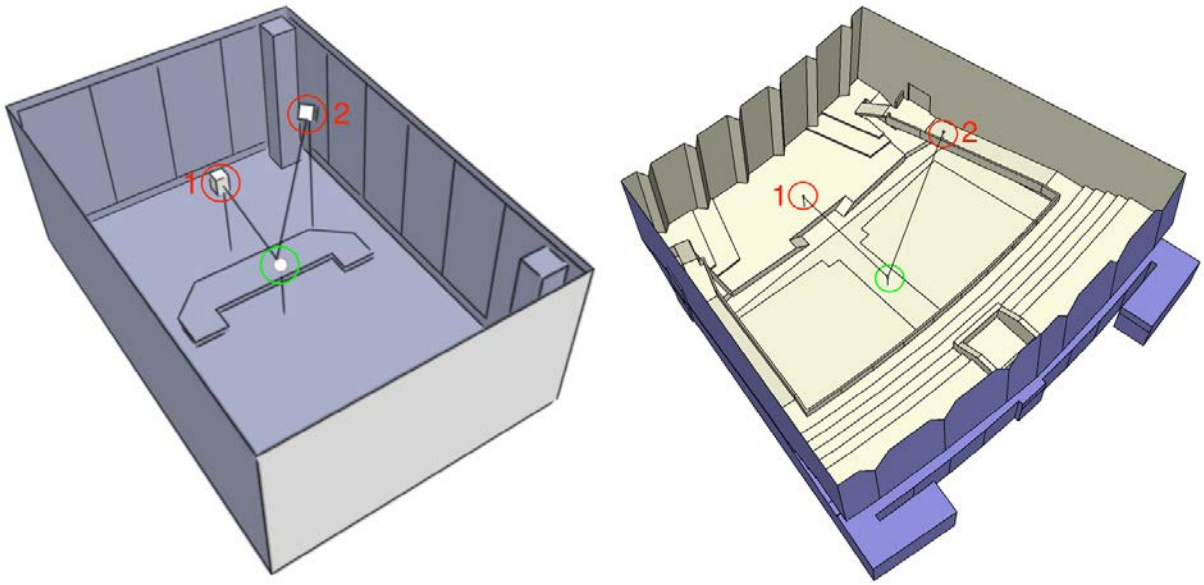


Figure 3: Loudspeaker positions (red) and the receiver position (green) in the SAQI listening test.

- **Tone color:** darker - brighter
- **Vertical direction:** shiftet down - shifted up (in °)
- **Distance:** closer - more distant
- **Externalization:** more internalized - more externalized
- **Localizability:** more difficult - easier
- **Naturalness:** lower - higher

The experiment took place in the Electronic Studio. The subjects were able to enter the inputs in a WhisPER user interface on a laptop independently. For playback, Sennheiser HD 800 headphones with head tracking system (Polhemus Patriot) were used. In order to achieve the best possible auralization, the interaural time difference in the binaural signals was customized by ITD extraction and manipulation [21] for each subject. Thus, we measured the participants' intertragus distance previously. In addition, the 8- and 16-channel MTB recordings were diffuse field equalized and, as well as the dynamic binaural synthesis, high-pass filtered with a cutoff frequency of 50 Hz. In this investigation, the entire audio processing and rendering was also done on an external computer.
A total of 30 people (21 male, 8 female, 1 non-binary) participated in the experiment. They had an average age of 29 years. 23 subjects can be considered as experienced listeners, because of their musical education in school, work, study and/or music school for several years. The remaining participants had no musical education. 17 subjects, just over half, have already participated in

listening tests on dynamic binaural synthesis. Again, almost all participants have no hearing defects except for three with a tinnitus.
At first the participants were made familiar with the experimental procedure and the user interface. Apart from that, the meaning of the individual SAQI items was discussed in order to avoid comprehension errors. Furthermore, a training process was completed with the user interface. The participants were also instructed to move their head a maximum of $\pm 40°$ on the horizontal plane, since the BRIRs for the reference signal were recorded in this range. With respect to each SAQI quality, the subjects were now able to indicate the perceived difference between the signals A and B on a scale with the aforementioned opposite scale labels from -1 to 1 in the user interface. A and B were randomly assigned by the test signal (8 or 16 MTB channels) and the reference (dynamic binaural synthesis). Initially a general difference between test signals and reference stimuli was retrieved. If a subject detected a difference in at least one stimulus, it could be indicated on a unipolar scale and after that the request of all other SAQI qualities was following. However, if no difference was specified for all stimuli, the test was automatically stopped. During the experiment, the participants were allowed to take their own tempo and to switch between the signals any number of times or to stop them. The test lasted about an hour. All audio processing and rendering was done on an external computer with a Linux system. The input commands containing all informations about the individual conditions, were sent from the laptop to the computer via OSC. A Pure Data patch controlled the commands and with the corresponding JACK Audio connections the signals with the valid audio content were sent via MADI to the headphones of the participants. The rendering of the BRIRs for the reference signal was done with the SoundScape Renderer (SSR) [23]. For rendering the test signals, the MTB Renderer was executed twice with a different number of input channels (8 and 16).

3. Results

A Plausibility

According to the plausibility, the individual sensitivity d'_i for each subject was calculated like in [10]. It depends on the individual correct detections and false alarms of each participant. The average sensitivity d'_{avg} in figure 4 shows the mean values of the individual sensitivities for both test groups. The values in both groups with $d'_{avg} > 0$ signify that there is a sensory difference between reality and simulation, with the simulation being recognized by the subjects more often than reality. The mean values of the response bias β are also shown in Figure 4. They were calculated from the individual response bias of the participants β_i. The values for both test groups with the average response bias $\beta_{avg} > 0$ show that the subjects tended to sense the stimuli (regardless of the sensitivity) as 'real'. On the right of figure 4 the equal-variance Gaussian signal detection model for reality and simulation with response behavior and sensitivity of both test groups is displayed. The data was tested in advance in a Kolmogorov-Smirnov test as normally distributed. A one-sided t-test showed that there is no significant difference between

an 8- and a 16-channel MTB playback with $p = .333$ for a significance level of $\alpha = .05$. Adjusting the values for the mean sensitivity with respect to the response bias, we received a sensitivity of $d' = 0.57$ for an MTB rendering with 8 channels and a sensitivity of $d' = 0.43$ for a 16 channel rendering with a type I error of 25 % and an type II error of 5 %. Converting these values into corresponding detection rates for a 2AFC paradigm, this results in a detection rate of $P_c = 0.66$ for 8-channel playback and a detection rate of $P_c = 0.62$ for 16 channels.

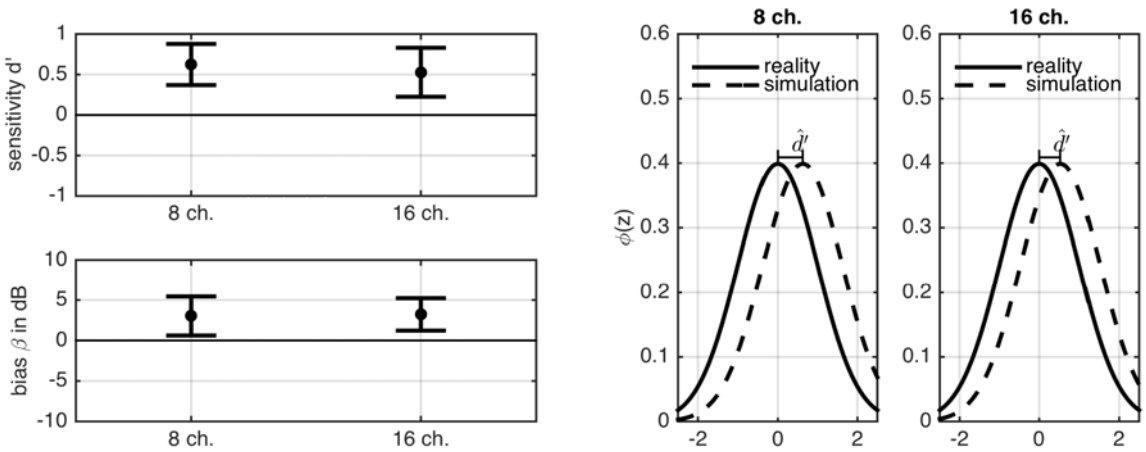


Figure 4: Left: Average sensitivity d' and bias β with 90 % confidence interval for 8-channel and 16-channel MTB playback. Right: Density distributions of the response behavior with equal variance for both test groups.

B SAQI

Before evaluating the SAQI listening test, the residuals for each SAQI item were examined for normal distribution using the Shapiro-Wilk test (significance level at 5 %). According to this criterion, the residual distributions of all qualities showed moderate outliers. For descriptive statistics we calculated thus the median as the location parameter and the bootstrap confidence intervals for interval estimation of this parameter with a number of samples $n_{boot} = 2000$ for performing the bootstrapping. The results of the SAQI listening test are shown in figure 5, illustrating the median values of all participants' ratings with 95 % bootstrap confidence intervals. They describe the perceived difference between the respective test signal and the reference with regard to the SAQI items. The value 0 indicates that no difference between test signal and reference could be detected. The gray sectors display the combinations of *room* and *source position*, whereby the speech signal is evaluated on the left, the noise signal on the right side. The two MTB test signals are represented in red (8 channels) and blue (16 channels). The y-scale describes the entire range from -1 to 1, within the subjects were able to rate. An exception is the item elevation. Here, the scale specifies the perceived source position displacement from $-90°$ to $90°$. The median values of the two test signals are at about the same level for all items.

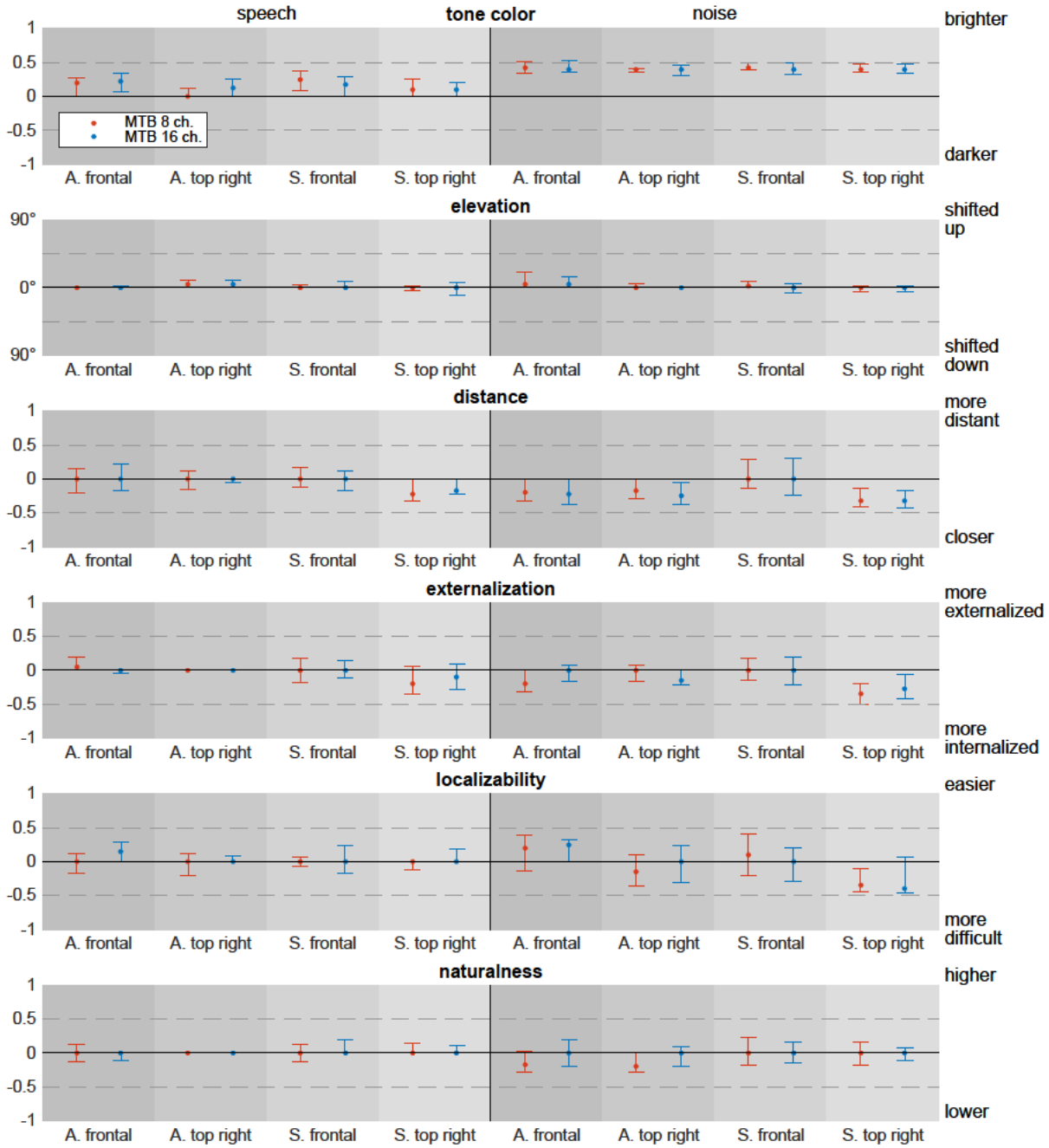


Figure 5: Median values with 95 % bootstrap confidence intervals. A = Audimax TU Berlin, S = Electronic Studio TU Berlin.

Speech signal: The differences between the MTB signals and the dynamic binaural synthesis show a small change in tone color for the speech stimulus. The median values of the test signals compared to the reference are in the range [0; 0.24]. There were perceived just minor differences in the source position on the vertical plane (elevation). A maximum position difference of 5° (shifted up) was noticed among the median values of all conditions for the speech signal. This refers to

the elevated source in the Audimax. All other median values are at 0. Concerning the distance, the median values for almost all conditions are at the same level as the binaural reference. Only the top right source in the Studio stands out slightly. In a 16-channel playback, it was felt a little closer with a value of -0.19. While listening to the 8-channel MTB playback, subjects also identified the same source as closer to the listening position than the reference source with a median value of -0.2. The same effect can be observed in the degree of externalization, where the elevated source in the Studio causes small deviations to the reference. During 16-channel MTB playback, the sound source was perceived as being more internalized with a median value of -0.1. When the speech signal was reproduced with 8 MTB channels, the source was felt to be more internalized with a rating of -0.19. The median values of the remaining conditions show no difference to the binaural reference with the exception of the frontal source in the Audimax. With a median of 0.04 it was perceived as more externalized. The localizability of the MTB test signals was more precise or at least as precise as the reference for each condition. Only the frontal source in the Audimax was rated with a median value of 0.14 as easier to locate. In terms of the naturalness of the test signals, the participants could perceive no difference to the binaural synthesis on average. The median values are 0 for all *room*position* conditions.

Noise signal: Larger differences were heard in the noise signal. When looking at the tone color, it can be seen that the MTB signals were perceived as brighter than the reference. The median values of both MTB signals range between 0.4 and 0.43. Only minor effects have been noticed in the elevation of the sources with a median position difference of 5° (frontal source in the Audimax, 8- and 16-channel MTB playback) and 1.5° (frontal source in the Studio, 8-channel playback) shifted up. Regarding to the distance, the test signals are at the same level, but compared to the dynamic binaural synthesis, the source positions in the noise signal were perceived closer to the listening position than in the speech signal with median values around -0.2. Especially the elevated source in the Studio was rated with a median value of -0.33, when playing with 8 and 16 MTB channels. Furthermore there was a slight decrease in externalization concerning the top right source position in the Studio. With a value of -0.27 for the 16-channel playback and -0.35 for the 8-channel playback, the source is rated more internalized as the reference source. In the Audimax, the frontal source during playback with 8 MTB channels (-0.19) and the elevated source during playback with 16 MTB channels (-0.14) were perceived as being more internalized. The localizability of the sound sources in the test signals varies a little bit more comparing to the speech signal. The frontal sound source in the Audimax could be located more easily with median values up to 0.24 for the 16-channel playback. However, subjects found it harder to locate the top right source in the Studio (maximum deviation at 0.4 for 16 MTB channels). The other values fluctuate around 0. The assessments of naturalness hardly deviate from the reference. Just the two source positions in the Audimax were perceived as less natural with median values down to -0.2.

To investigate the interaction effects between the variables, we used a MANOVA with repeated measures. Beforehand, for the conditions *content*room*position*test signal* we tested the par-

ticipants' ratings for normal distribution using the Shapiro-Wilk test with a significant level of 5 %. We were able to observe that less than half of the values for each condition could achieve a normal distribution according to this criterion. Even though there are outliers present in the remaining conditions, due to the data situation, the ratings can be regarded as nearly normally distributed. Furthermore, the linear relation between the SAQI items was investigated by calculating the Pearson correlation coefficient r as a measure of the effect size. For an average coefficient r_{avg} over all itmes, we got $r_{avg} = 0.08$ with a maximum correlation coefficient for externalization*distance of $r = 0.21$ ($p = .00$, $n = 30$). These are weak correlation effects between the items. Based on this assumptions, a MANOVA can be used.
The Mauchly test for sphericity is irrelevant in this case, since the inner subject factors each have only two forms and thus sphericity is given. When testing the main effects of the model with the software SPSS, it can be observed that at a significance level of $\alpha = .05$ there are significant interactions within the variables *content* ($F(6,24) = 18.045$, $p = .000$, partial $\eta^2 = .819$), *room* ($F(6,24) = 3.126$, $p = .021$, partial $\eta^2 = .439$) and *position* ($F(6,24) = 4.933$, $p = .002$, partial $\eta^2 = .552$). Furthermore, the interactions between *content* and *position* ($F(6,24) = 2.641$, $p = .041$, partial $\eta^2 = .398$) as well as *room* and *position* ($F(6,24) = 3.282$, $p = .017$, partial $\eta^2 = .451$) are significant. Within the two test signals (8 and 16 MTB channels) there are no significant interactions ($F(6,24) = .607$, $p = .722$, partial $\eta^2 = .132$). Bonferroni-corrected post-hoc tests show that the content has a significant effect on tone color, distance and externalization. Thus the tone color in the noise signal ($M = .377$, $SD = .026$) is significantly brighter than in the speech signal ($M = .168$, $SD = .025$), the sound sources at the noise signal ($M = -.182$, $SD = .027$) felt significantly closer to the listening position than at the speech signal ($M = -.037$, $SD = .02$) and the sources are perceived more internalized in noise ($M = -.166$, $SD = .03$) than in speech ($M = -.01$, $SD = .034$). In addition, the Bonferroni-corrected pairwise comparisons testify that the test sources in the Audimax ($M = 6.66°$, $SD = 2.52°$) were felt to be significantly shifted upwards compared to the Studio ($M = -.54°$, $SD = 1.98°$). Significant interactions also show up in the Bonferroni-corrected pairwise comparisons at the source positions. Firstly the tone color of the sources at frontal position ($M = 0.309$, $SD = .024$) was perceived as brighter than at the elevated position ($M = .236$, $SD = .024$). Likewise, the position affects the elevation perception of the sound sources. The frontal source were perceived to be significantly upshifted ($M = 5.95°$, $SD = 2.7°$; elevated source: $M = -.18$, $SD = 1.8$). Moreover, the top right source ($M = -.0166$, $SD = .024$) was felt to be significantly closer to the listener than the frontal source ($M = -.053$, $SD = .027$). The degree of externalization also showed significant differences. Thus, the top right source ($M = -.123$, $SD = .029$) appears more internalized than the frontal source ($M = -.002$, $SD = .033$). Finally, significant localizability interactions were also found. Here, the elevated source ($M = -.082$, $SD = .044$) could be located more difficult than the frontal source ($M = .074$, $SD = .033$). When examining the interaction between *content* and *position*, the profile diagrams show that the two variables interact significantly, but that the main effects of *content* and *position* are not disturbed and thus interpretable. This also applies to the

significant interactions between *room* and *position*. Again, there are no correlations between the two variables and the main effects can be clearly interpreted. Figure 6 shows for which conditions the values deviate significantly from 0, i.e. show a significant difference to the reference according to the criteria examined here.

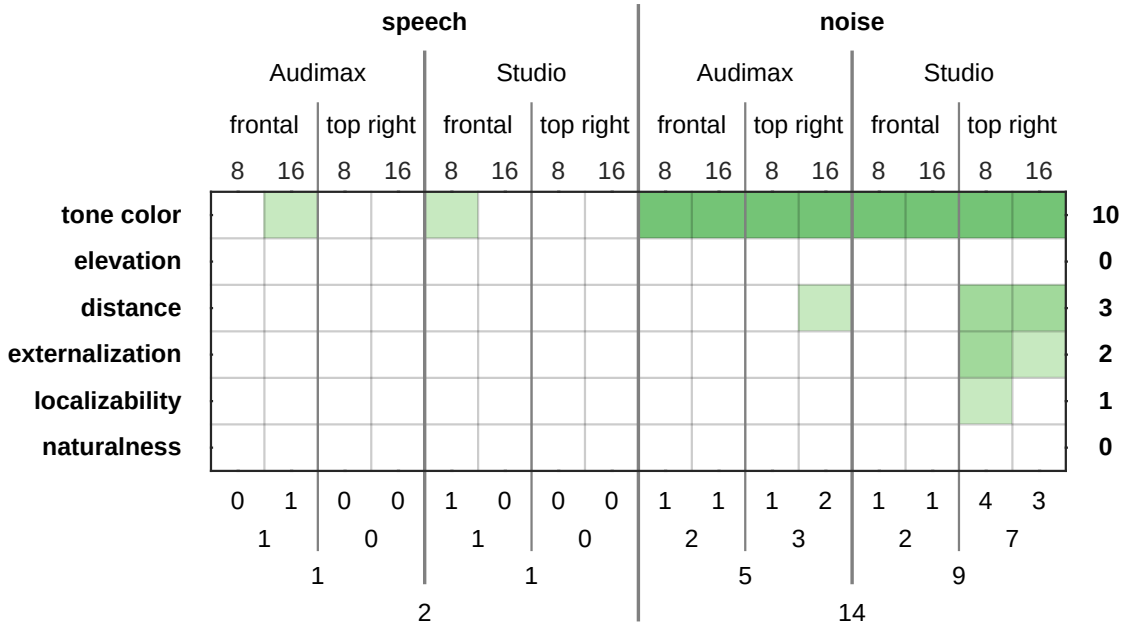


Figure 6: Deviations of the confidence intervals from 0 in graded shades of green (confidence intervals do not intersect with the zero-line). The darker the color, the greater the distance between the intervals and the zero-line.

4. Discussion

A Plausibility

In contrast to [11], which already investigated the perceived plausibility of MTB signals (in terms of interpolation algorithm, discretization and content), we used for our study real signals as reference. The subjects were aware that real and simulated stimuli were presented to them in randomized order. When listeners are asked if they have heard a real sound event or not, the detection rate of the simulated signals compared to pure guessing rate at 50 % is surprisingly slightly increased, at 62 %, if an input number of 16 channels and optimal interpolation is used in dynamic playback. On these conditions, the detection rate is just 12 % above the guess probability. A bit more frequently, the subjects were able to recognize the simulation when playing with 8 channels. In this case the detection rate is 66 %. The difference in mean values between the two test groups with respect to the sensory difference shows that a higher resolution in discretization resulted in a lower detection rate of the simulation, which can be regarded as an expected result. However, the difference is not significant, as can be also seen in the overlapping

confidence intervals in Figure 4. Due to the results in [11], where the perceived plausibility of discretization in MTB recordings was investigated, no better results can be expected with a further increase in the number of microphones. In comparison to the results of the MTB recordings, a re-synthesis of virtual acoustic scenes based on binaural room impulse responses achieved a detection rate of 51 %, as evaluated in [10]. However, this approach can not be used for live recording and live transmission of real sound fields. In [10] a plausibility criterion at a detection rate of $P_c = 0.55$ was assumed. That means, the detection rate of the simulation may exceed the guess probability by less than 5 %, if the system is to be considered as plausible. This criterion is very strict though and is below the absolute threshold of a psychometric function [10]. In a 2AFC paradigm, this value is at a detection rate of $P_c = 0.75$ and exceeds the detection rates of the 8 and 16-channel MTB reproduction. On the one hand, a certain latency of the system due to the block-by-block processing of the signals when the head was moving too fast, could have led to the detection of the simulation. Furthermore, the subjects were asked to complete a questionnaire after the listening test. On this basis, it could be figured out that artifacts were perceived during the playback of the simulated stimuli. These were, for example, comb filter effects, blurred sources and unstable localization of the sources, caused by missing spectral pinna cues and a divergence between the diameter of the MTB array and the head of the subjects. The results in [15] showed that the diameter difference between the sphere model and a humans' head causes the virtual source to move with or against the heads' rotation direction, depending on whether the sphere is smaller or larger than the head. In addition, a difference in tone coloration between real and simulated signals resulting from missing spectral influences of Pinna and Torso or rather diffuse field equalization could have led to the identification of the simulation. In conclusion, however, it should be noted that many of the subjects are experienced listeners. Some of them had already come into contact with the reconstruction of VAEs while studying or working. So they have been sensitized to corresponding system discrepancies by a trained ear.

B SAQI

For each investigated SAQI item we tested the perceived difference between an 8- and a 16-channel MTB recording (test signals) compared to a BRIR-based dynamic binaural synthesis as reference for two different source positions (frontal, top right), rooms (short, long reverberation time) and contents (speech, noise). Looking at the test signals in figure 5, it can be observed that the values for all investigated SAQI qualities are on a similar level. The fact that there are no significant differences between the test signals is also shown by the results of the MANOVA. Apart from the fact that the detection rate of the 16-channel playback while evaluating the plausibility, as discussed in section 4, was slightly lower (even when there was no significant difference), the 8-channel playback is not inferior to the qualities tested in the SAQI experiment described here. A recording device with just 8 channels can thus be a simplified alternative with qualitatively equivalent results in terms of the aspects investigated in this study.

In total, only a few significant differences between the test signals and the dynamic binaural synthesis were perceived. With only 16 out of 96 conditions, the values differ significantly from the value 0. These deviations occur especially in the case of the sensitive noise signal. The results regarding the tone color show that the test signals are perceived as much brighter compared to the reference. A possible explanation for this effect could be the diffuse field equalization filter of the MTB array. Since it probably does not provide an optimal equalization for the room acoustic environments tested in this study, it entails a certain tonal coloration. This is significantly more noticeable in the noise than in the speech signal, because it is a very sensitive signal and the diffuse field equalization filter causes a frequency increase especially in the high frequencies. Surprisingly, there was just a minor perceived difference in the MTB source positions compared to the reference source position on the median plane with small significant differences between the two rooms. The most important cues for the localization of sound sources on the median plane are the spectral cues induced by the pinna [6]. Despite the missing spectral informations of the pinna in MTB recordings, the subjects were able to perceive the elevation of the investigated sound sources. For static sound environments, the spherical head model provides only weak cues for the localization of elevated sound sources. However, motion cues induced by head movements, which cause changes in the ILD and ITD, can provide missing informations to improve the perception of elevation [24]. In the SAQI listening test, we presented dynamic sound scenes with head tracking. The presence of the motions cues may have led to the perception of the elevated sources. When evaluating the distance, the sources of the test signals with speech content were observed at about the same distance as the reference sources. Comparing these values with those of the noise signal, it can be seen that the test sources were not much, but significantly perceived closer to the listening position. This could be explained by the fact that the sound localization is characterized among other things by the spectral coloring. The further a sound event is away from the listener, the lower it sounds, as the high frequency part of the signal is more strongly absorbed by the air. Since the tone color of the MTB signals was noticeably brighter than the reference signal, especially in the case of noise, this could be a reason for the localization differences. A positive feature is the externalization of the sound sources in the test signals with only few deviations from the reference in almost all conditions. Only the elevated source in the Studio while listening to the noise signal was perceived as more internalized. Concerning the localizability of the sound sources, only small differences can be noticed here as well. As with the externalization, the median values of both test signals show that the elevated source position in the Studio deviates more strongly from the reference than the remaining median values in the noise signal. The localizability of this source position was indicated as more difficult. Particularly when rating the two last-named SAQI qualities, the subjects heard significant differences between the binaural reference and the test signals regarding to the elevated source in the Studio. The sensitivity of the noise and the perceived tonal difference of the test signals, in combination with the missing pinna, could have contributed to this effect. Regarding the naturalness of the MTB recordings, the test signals were nearly rated as natural as the reference signal.

5. Conclusion

In the course of our study, a recording device was developed in cooperation with Sennheiser working on the motion-tracked binaural technique, which was introduced by Algazi et al. [2] in 2004. To test its quality, we first investigated the plausibility of signals reproduced by the MTB method. In [11] the MTB sound quality regarding to the perceived plausibility was already evaluated. The study tested the variables microphone number, audio content and interpolation method without a reference. The results showed that the Spectral-Interpolation Restoration was the most plausible of the five interpolation algorithms presented in [2]. Using this algorithm, a varied number of microphones produced consistent effects in terms of plausibility. In contrast to this study, we verified the plausibility according to [10] with real signals as a reference and re-tested an MTB recording with 8 and 16 channels. In a 2AFC paradigm, we obtained a 62 % detection rate for 16-channel playback and a 66 % detection rate for 8-channel compared to a guess probability of 50 % using SDT with minimum effect hypothesis for the analysis.
In a second listening test, we investigated the quality of an 8- and a 16-channel MTB recording in comparison to a dynamic binaural synthesis with the BRIRs recorded by the artificial head FABIAN as reference by means of the items tone color, vertical direction, distance, externalization, localizability and naturalness from the SAQI. For this we tested the variables *room*, *source position* and *content*. For the evaluation of the ratings, the median values with 95 % bootstrap confidence intervals and significant interactions with a MANOVA were calculated. The results showed up differences between the MTB signals and the binaural reference for the speech stimulus just in a small change in tone color, while the elevation, externalization, localizability, and naturalness of the sound sources were nearly at the same level as the binaural reference. Slightly larger were the differences for the elevated source position in the noise signal, where there was a small, but significant decrease in externalization and localizability. In addition, the source was felt closer to the listening position. On the one hand, however, the results of the elevation assessment deserves a special mention. Despite the lack of spectral influences of the pinna and torso, subjects only noticed a minor difference in source position on the vertical plane. Concerning the degree of externalization on the other hand, the median values for most conditions show that only a small difference to the reference was perceived here as well. Since the dynamic binaural synthesis can not be used for live recording and live transmission of dynamic spatial sound fields, the MTB microphone thus offers a promising alternative.

PART II - Documentation

6. Hardware

The MTB recording device constists of a reverberant sphere with a diameter of $d = 17.6$ cm and 16 microphone capsules embedded in the surface of the horizontal circumference. Figure 7 shows the outer view on the microphone array on the left side. On the right side of the graphic you can see the inside view of the array from above. This section focuses on the hardware of the recording system. The first part describes the components of the sphere array more detailed. This is followed by the selection of the microphone capsules and the measurement results in the anechoic chamber. Finally, the preamplifiers, which are used especially for this MTB recording system are considered more closely.

Figure 7: Outer and inner view of the MTB recording device.

6.1. Sphere

The sphere is built of three main components: an upper shell, a bottom shell and a center ring with openings for the microphone capsules. The individual parts were produced by a generative manufacturing process named Selective Laser Sintering (SLS). They consist of the material polyamide and were subsequently painted black. The CAD model required for 'printing' the sphere was made by an external designer and can be seen in Figure 8.
On one side, the logo of the TU Berlin with lettering of the Audio Communication Group is engraved in the upper shell. On the other side, logo and lettering of the cooperation partner Sennheiser are imprinted. At the top of the shell is a recess in the material for a screw, holding the components together when assembled. After mounting the parts, the recess can be sealed with a small cap that locks into place. A thin slot screwdriver works well for taking off the cap.

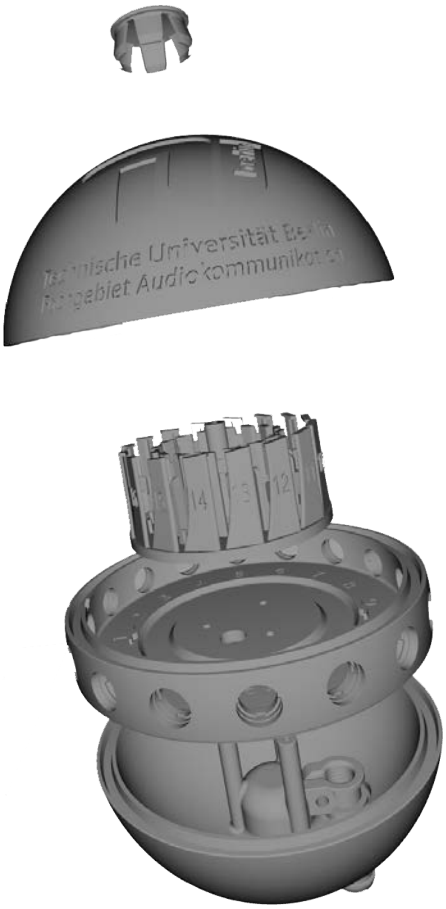


Figure 8: CAD model in exploded view.

Furthermore, a thin line is inserted on the shell, which connects the capsule positions 1 and 9. Position 1 is marked with a slight notch in the center ring (figure 9). It facilitates the alignment of the array. Capsule position 1 represents the frontal view (azimuth = 0°, elevation = 0°) of the listener.

Figure 9: Notch at capsule position 1.

On the lower side in the bottom shell, an opening is embedded, where the multicore cable can be passed. With a screw, the opening can be tightened inside to achieve a strain relief for the cable. In addition, a K&M 215 thread reducer with 3/8-inch internal thread is mounted in the lower shell so that the array can be screwed onto a standard microphone stand. For the microphone capsules, 16 openings were inserted in the surface of the center ring, which were labeled inside with the capsule positions to facilitate the assembly. Elastic rings (see A.1 for construction drawing) can be inserted here, serving as a holder for the microphone capsules. At capsule position 1, there is also a small recess inside, where the top shell can snap into place to support the proper assembly of the parts. The mount for the converter circuit boards was also manufactured in the SLS process. It is located between the upper shell and the center ring of the sphere and was screwed onto the circular area of the ring. It enables fixing the separate boards upright at a constant distance from each other. The slots are also numbered for easier assignment. To shield the converter boards, a laser cut steel sheet case (Figure 10, construction drawing can be found in A.1) with a thickness of 0.5 mm was placed around the mount. In the sidewall and at the bottom of the case there are circular openings for the cable routing as well as small pilot holes for the screws. After folding the sidewall, it is held together by the bottom and the cover. First, the steel sheet case without top cover was placed between the board mount and the center ring in the array, as shown on the right in Figure 7. After the wiring, the case cover can be added. It should be noted that the tabs for fixing the top cover do not need to be folded down, since at the end all components are held together anyway by the main screw. Otherwise, it is cumbersome to reach the converter circuit boards in the steel sheet case retrospectively, if a repair is necessary.

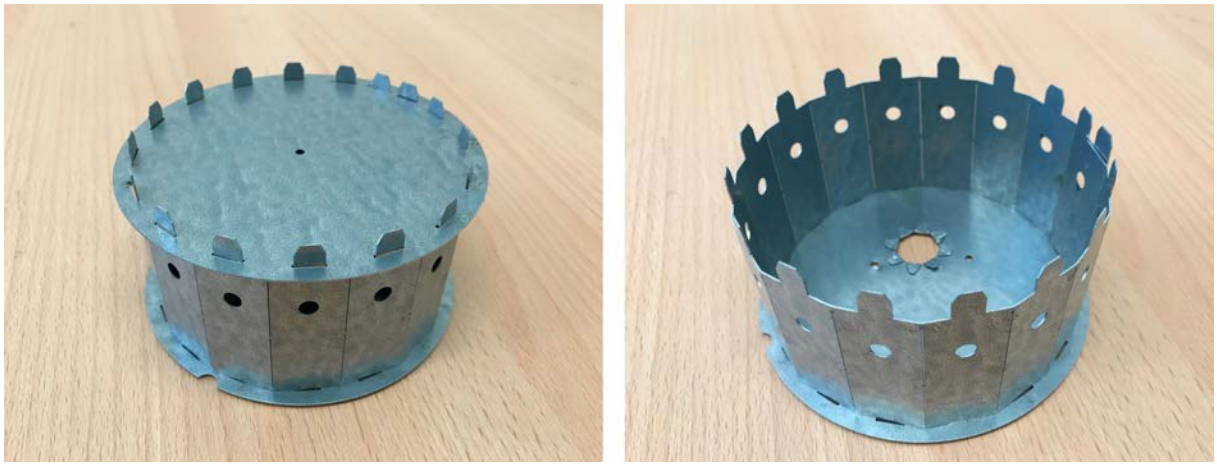


Figure 10: Steel sheet case for shielding the converter circuit boards.

To connect the capsules to the converter boards, the symmetric patch- and microphone cable Sommercable SC-Cicada[1] with a total diameter of 2.6 mm, two 0.14 mm^2 inner conductors

[1] http://shop.sommercable.com/Kabel/Meterware-Audio/Patch-Mikrofonkabel-SC-Cicada-SO-D14-200-0451.html#tab_attributes, last accessed on 17.04.2018

and copper spiral shielding as well as a tinned drain wire were used. The multicore cable is a Sommercable Multipair SC-Transfer AMCK[2] with a total diameter of 14.50 mm, 16 channels and two 0.14 mm^2 inner conductors, shielded in pairs with a copper spiral. The overall shield consists of a copper braid. For shielding the capsule boards and patch cables, a tinned copper braid was used. For the analog connection, 16 NC3MX Neutrik XLR cable connectors[3] (3-pin) have been installed.

6.2. Microphones

The omnidirectional electret condenser capsules KE14 (figure 11) and converter circuit boards (see section A.1 for the construction drawing with dimensions) from Sennheiser were used for the array. The capsules have a diameter of 14 mm and a height of 6 mm.

Figure 11: Sennheiser KE14 capsules.

The microphones are used in the amplifier configuration [25] according to Sennheiser. In the course of this, ground and phase inverted signal are merged on the capsule. This configuration, seen in Figure 12, results in a small amount of inherent noise. In order to use phantom power of 12-48 V, it is converted to the operating voltage of the capsules on the converter circuit boards. At this point, phase inverted signal and ground are also separated to obtain symmetric signal transmission.

6.2.1. Matching

The aim of the capsule matching should be to find 16 capsules that do not differ in amplitude over the entire frequency range by more than 1 dB. For this purpose, Sennheiser provided the

[2] http://shop.sommercable.com/Kabel/Meterware-Audio/Multipair-SC-Transfer-AMCK-100-0401.html, last accessed on 17.04.2018
[3] http://www.neutrik.de/de-de/xlr/x-serie/nc3mx, last accessed on 17.04.2018

FACTS ABOUT ELECTRET WIRING

Here are the technical descriptions of the two ways that an electret capsule and FET can be

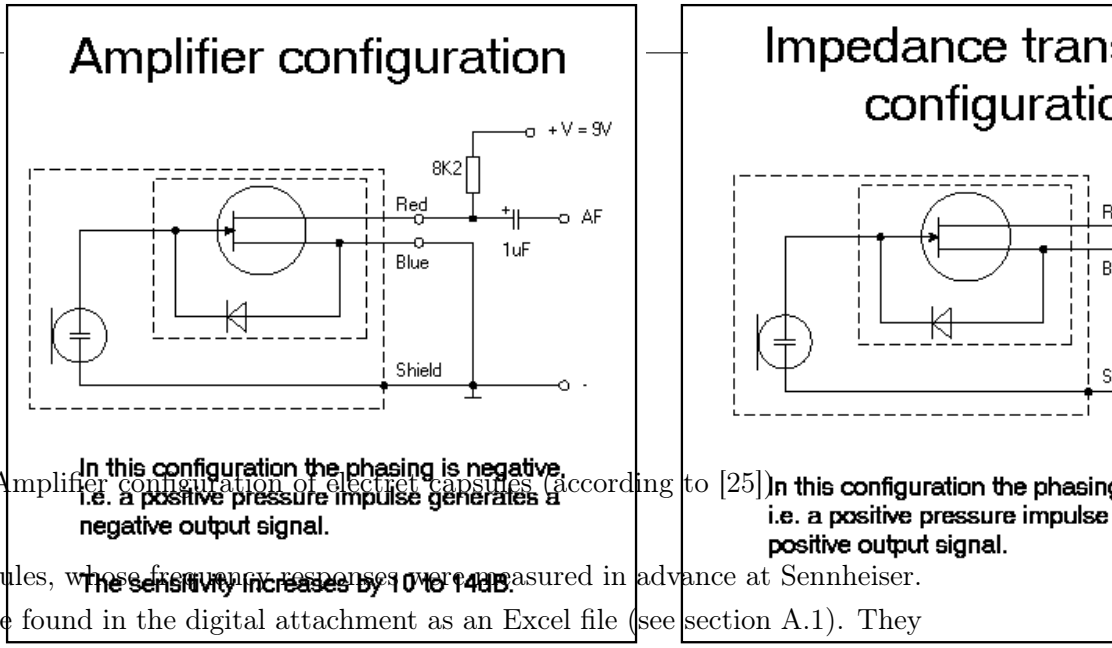


Figure 12: Amplifier configuration of electret capsules (according to [25])

All Sennheiser RF transmitters require the microphone to be wired in the amplifier or "two-wir

Because of this, many Sennheiser microphones such as the MKE2-4 or ME104 are built with t

shield tied together at the capsule. This results in the least amount of noise possible. These micr

be used in the amplifier configuration.

Microphones such as the MKE2-5 can be wired in either configuration depending upon the appli

continuity test between the shield and blue wire will indicate which configuration can be used.

In situations that require an electret microphone to be powered by 12 - 48V phantom powering, a

circuit must be used. Sennheiser can provide a "-PH" adapter which is the converter circuit insi

connector.

* When using a KA100 cable, substitute white for blue and brown for red.

6.2 Microphones

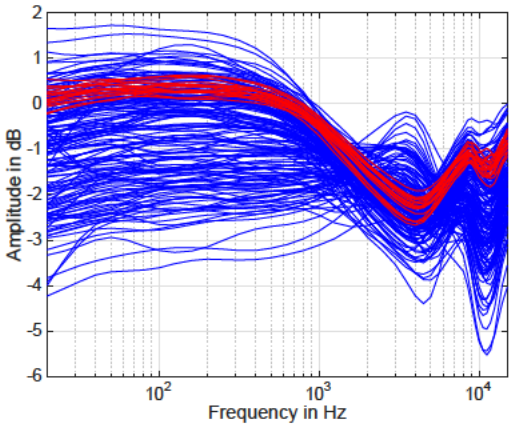
data of 162 KE14 capsules, whose frequency responses were measured in advance at Sennheiser. The value tables can be found in the digital attachment as an Excel file (see section A.1). They include the values of the amplitude responses once unnormalized and once normalized to 1 kHz as well as the serial numbers and sensitivities of the capsules. Optimally, a selection should be made from the unnormalized amplitude responses. Furthermore, the question arose about the methodology for calculating a best matching capsule set. This led to the idea of calculating either the difference or the euclidean distance between the amplitude responses. For decision making, the calculation was made using both methods. The investigation of the euclidean distance proved to be the better choice, since it produced a matching with smaller amplitude differences. The source code in 6.2.1 enables the calculation of both variants. Here the implementation of the 'difference' in line 59 is commented out. The set should contain a total of 18 capsules (two capsules as a replacement).

```
1  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
2  %%% Capsule Selection KE14 %%%
3  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
4
5  clear all
6  close all
7  clc
8
9  for ind = 1:4;
10
11     % read data
12     if ind == 1
13         A = xlsread('KE14normalizedSet1.xls');
14     elseif ind == 2
15         A = xlsread('KE14normalizedSet2.xls');
16     elseif ind == 3
17         A = xlsread('KE14unnormalizedSet1.xls');
18     else
19         A = xlsread('KE14unnormalizedSet2.xls');
```

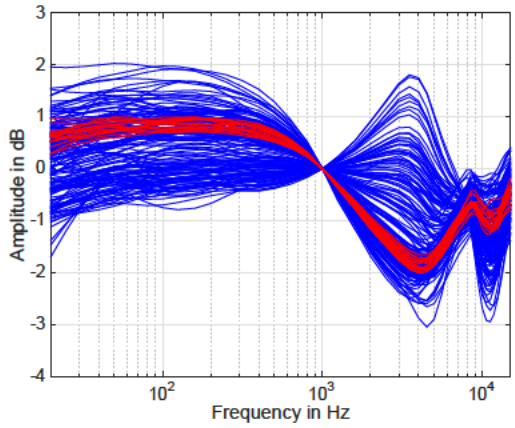
```
    end

    freq = A(1,:);

    % maximum difference allowed 1 dB
    differenceThreshold = 1;

    % clear up the data
    MicData = A(4:end,:);

    % number of capsule selection
    Selectionnumber = 18;

    Selectionnumber = Selectionnumber + 1;

    Diff = zeros(size(MicData));
    DiffSum = zeros(Selectionnumber,size(MicData,1));
    TempUltimate = zeros(Selectionnumber,size(MicData,1));

    for i = 1:size(MicData,1)

        % compare Mic Data
        for row_idx = 1:size(MicData,1)
            Diff(row_idx,:) = abs(MicData(i,:)-MicData(row_idx,:));
        end

        % find absolute differences above 1 dB
        x = Diff > differenceThreshold;
        % sum over frequency vectors to find pairs with values over 1 dB threshold
        y = sum(x, 2);
        Selection = find(y<1);

        % check if there are enough mics within threshold borders
        if length(Selection)<Selectionnumber
            fprintf('Skipping Mic Set number %i. Not enough mics within borders. ...
                \n', i);
            continue
        end

        % sum up frequency band difference of all mic pairs
        % TempDiffSum = sum(Diff(Selection,:),2);

        % Euclidian distance between frequency vectors
        TempDiffSum = sqrt(sum(Diff(Selection,:).^2,2));

        [TempDiffSum ,a] = sort(TempDiffSum);

        % save to matrix for all mic sets
        DiffSum(:,i) = TempDiffSum(1:Selectionnumber);
```

```
        TempUltimate(:,i) = Selection(a(1:Selectionnumber));

    end

    % find valid cols
    [~, validcols, DiffSumSelection] = find(sum(DiffSum,1));

    % valide cols in TempUltimate
    TempUltimate = TempUltimate(:,validcols);

    % best matching
    [DiffSumSelection, b] = sort(DiffSumSelection);

    % ultimate set
    resultingSet = TempUltimate(2:end,b(1));

    % plot
    figure
    semilogx(freq,MicData', 'b');
    set(gca,'fontsize', 16)
    if ind == 1
        title('Euclidian distance, normalized, Set 1');
    elseif ind == 2
        title('Euclidian distance, normalized, Set 2');
    elseif ind == 3
        title('Euclidian distance, unnormalized, Set 1');
    else
        title('Euclidian distance, unnormalized, Set 2');
    end
    xlabel('Frequency in Hz');
    ylabel('Amplitude in dB');
    hold on;
    semilogx(freq,MicData(resultingSet(1:end),:)', 'r');
    grid on;
    xlim([min(freq), max(freq)]);

end
```

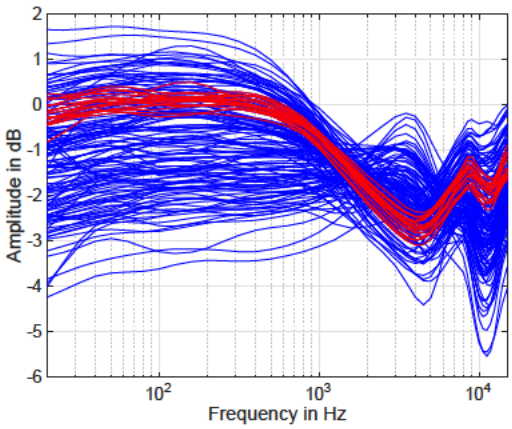
Figure 13 shows the amplitude responses of the 'ultimate' selection (set 1) for the unnormalized and normalized amplitude responses and in addition, a second replacement selection (set 2), represented by the red curves. The amplitude responses of the remaining capsules are shown in blue. The serial numbers of the capsule selections are listed in table 3.
Since it was possible to find a set among the unnormalized capsules within the given threshold (difference threshold < 1 dB), the decision fell on the selection in 13 (a). Actually installed in the microphone are the capsules (indication of serial numbers) in table 4. The table also shows the respective position in the array and the sensitivities of the capsules (unassembled and without amplification) in relation to 10 mV/Pa.



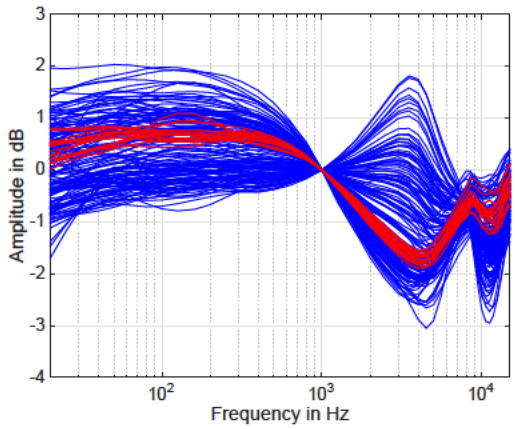
(a) Euclidian Distance, unnormalized, Set 1



(b) Euclidian Distance, normalized, Set 1



(c) Euclidian Distance, unnormalized, Set 2



(d) Euclidian Distance, normalized, Set 2

Figure 13: Matched sets of KE14 capsules

6.2.2. Frequency response measurement

After installing the microphones, a measurement of the frequency responses in the anechoic chamber of Georg Neumann GmbH Berlin was executed. For this, the frequency responses of each capsule were measured for all directions in a resolution of 2.8125°. The measured data is the basis for the subsequent diffuse field equalization. Figure 14 shows the amplitude responses of the mounted capsule at position 1 for the directions 0°, 90° and 180° cw.

To compare the capsules, the amplitude responses of all capsules are shown in Figure 15 in 0° direction. For the measurement data and the code for calculating the diffuse field equalization filter see section A.1.

Table 3: Capsule Selections (Serial numbers)

	Set no. 1		**Set no. 2**	
Position	normalized	unnorm.	normalized	unnorm.
1	161	116	126	115
2	122	135	93	20
3	9	136	85	45
4	162	85	118	53
5	14	122	154	152
6	123	148	99	82
7	148	126	8	57
8	98	93	117	98
9	2	142	7	97
10	97	76	112	86
11	86	2	19	47
12	128	100	1	161
13	22	112	11	49
14	95	8	159	3
15	136	128	142	95
16	82	153	76	22
17	37	13	32	37
18	135	91	124	110

Table 4: Sensitivity in relation to 10 mV/Pa

Position	Serial No. Capsule	Δ Sensitivity in mV/Pa
1	116	-0.49
2	135	-0.55
3	136	-0.43
4	85	-0.48
5	122	-0.50
6	148	-0.33
7	126	-0.55
8	93	-0.40
9	142	-0.39
10	76	-0.42
11	2	-0.59
12	100	-0.55
13	112	-0.27
14	8	-0.28
15	128	-0.35
16	153	-0.70

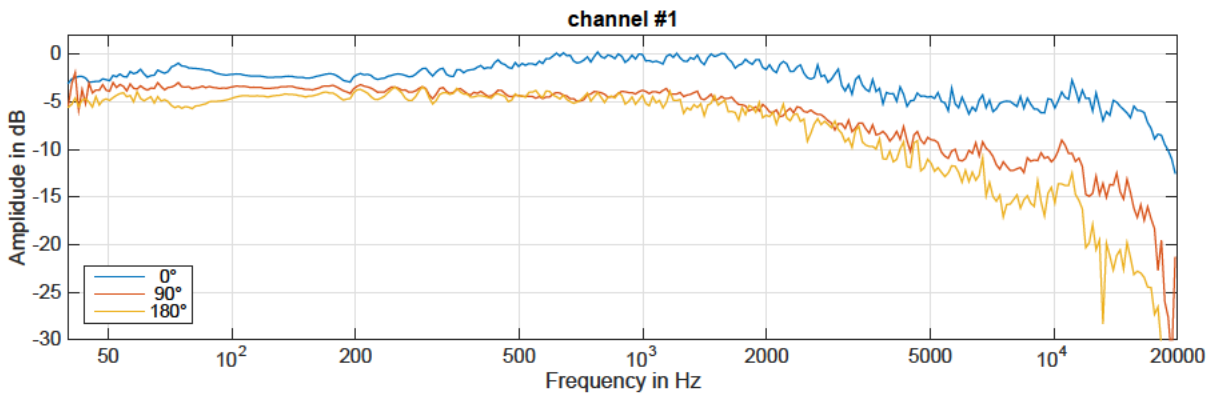


Figure 14: Amplitude response of the installed capsule at position 1 in 0°, 90° and 180° direction.

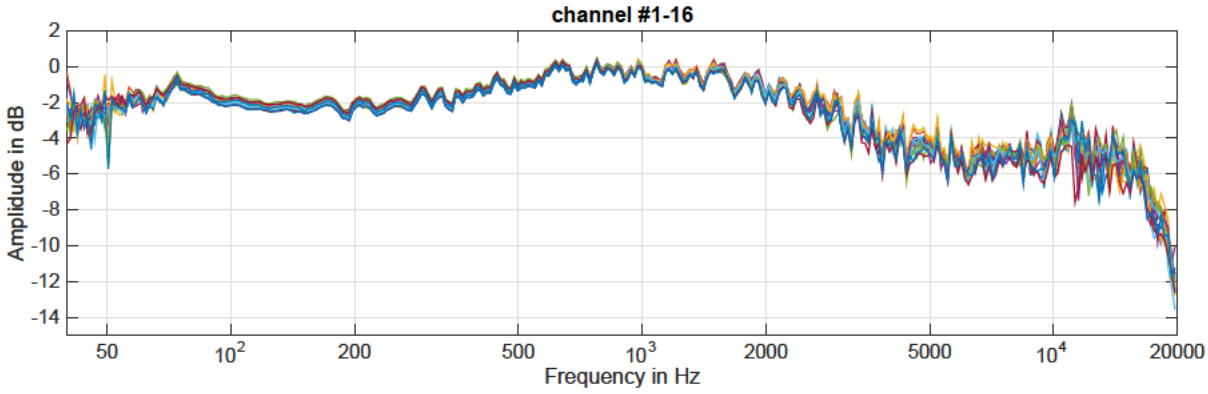


Figure 15: Amplitude response of all installed capsules in 0° direction.

6.3. Preamplifiers

Another part of the MTB recording system are the two Prisonus DigiMax DP88 preamplifiers and a/d/a converters. They are digitally controllable and have 8 analog XLR inputs each. To synchronize the preamps via ADAT we use a RME Fireface UFX audio interface. More about the configuration can be found in section 8.1. Beforehand, however, we checked both preamps for their gain control range and their inherent noise level to compare their gain behavior. For the measurement we used the Prism Sound dScope Series III. The SNR was measured for each of the 8 analog XLR input channels. For this purpose, the pure noise level of the devices was recorded for 0 dB and 30 dB gain. The values for both preamplifiers (with inventory numbers 2451 and 2452) are shown in the table 5.

To verify the amplification factor, a 1 kHz sine signal with -60 dBFS was generated and the output level of the preamps was measured for different amplification levels. For this, only the channel to be measured was amplified, while the other channels were set to 0 dB. The results of the measurement, which was performed for 0, 20 and 40 dB amplification, are shown in table 6. For a precise comparison, the analog input channel 1 was subjected to a more accurate measurement on both devices. Again, a 1 kHz sine at -60 dBFS was fed in and the output level was

Table 5: Signal-to-noise ratio of the Presonus Digimax DP88 with 0 dB and 30 dB gain

channel	**+0 dB**	**+30 dB**	
1	-102	-99	No. 2451
2	-102	-99	
3	-102	-99	
4	-102	-99	
5	-102	-99	
6	-101	-99	
7	-101	-99	
8	-100	-99	
1	-100	-97	No. 2452
2	-100	-97	
3	-101	-97	
4	-97	-97	
5	-96	-95	
6	-100	-98	
7	-100	-98	
8	-100	-98	

Table 6: Output level of the preamplifiers at 0, 20 and 40 dB gain

channel	**+0 dB**	**+20 dB**	**+40 dB**	
1	-61.4	-38.8	-21.5	No. 2451
2	-61.4	-38.8	-21.5	
3	-61.3	-38.6	-21.1	
4	-61.4	-38.7	-21.3	
5	-61.5	-38.8	-21.6	
6	-61.3	-38.7	-21.3	
7	-61.3	-38.6	-21.1	
8	-61.3	-38.6	-21.2	
1	-61.1	-38.4	-21.2	No. 2452
2	-61.1	-38.4	-21.1	
3	-61.2	-38.5	-21.0	
4	-61.2	-38.5	-21.1	
5	-61.3	-38.7	-21.3	
6	-61.4	-38.7	-21.3	
7	-61.3	-38.6	-21.0	
8	-61.4	-38.7	-21.2	

measured for all gain steps. Figure 16 shows the output of the preamps in dB. Both curves have a very similar amplification behavior. The values of this measurement are displayed in table 7.

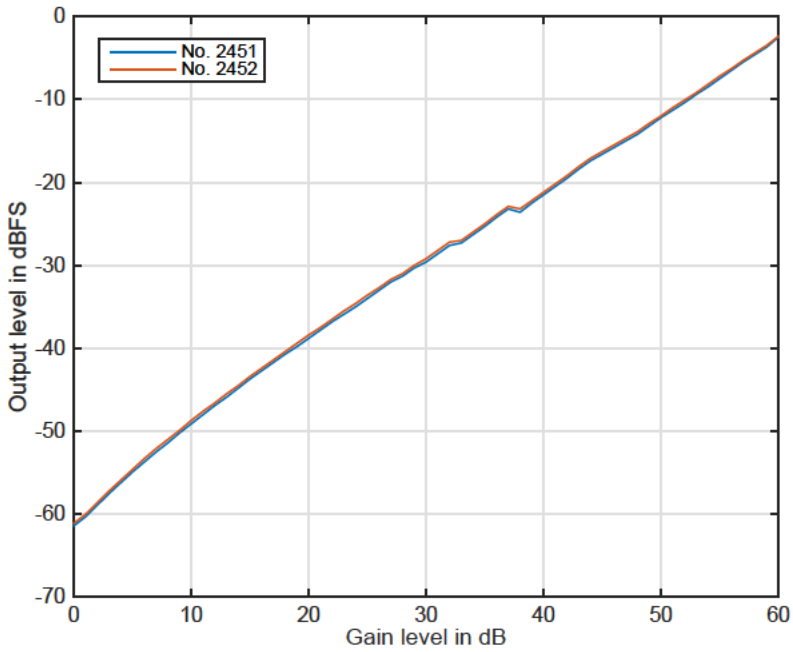


Figure 16: Amplification behavior of the preamplifiers with inventory numbers 2451 and 2452.

7. Evaluation

The following chapter focuses on the two listening tests and goes a little closer to the signal flow and further results analysis. The answers of the participants from the questionnaires can be found in A.1. The IDs match the subject numbers in the attached data. The questionnaire and the instructions for the subjects in the plausibility listening test were created on the basis of listening test materials by Alexander Lindau (see section A.1).

A Plausibility

To generate the stimuli for the plausibility listening test, the impulse responses for five various loudspeakers in the Electronic Studio of the TU Berlin were recorded. The exact coordinates and distances can be found in table 1. The MATLAB based software toolbox AKtools [26] was consulted for measuring the impulse responses with an 18th order sine sweep as excitation signal. The impulse responses are attached as wave files and can be found according to the instructions in section A.1. The anechoic audio material that was used to convolve with the impulse responses, was provided by the Audio Communication Group (see section A.1).
Figure 18 shows the signal flow in the plausibility listening test. After the participants entered their decision into the graphical user interface, the information about the signal content and the channel are sent to the Pure Data Patch via OSC. At this point the corresponding audio content and the output channel are selected. In case of a real signal output, the signal is transmitted to

Figure 17: Electronic Studio at the TU Berlin with MTB microphone at central position

the respective loudspeaker in the experimental room. If the signal is to be played via headphones, further intermediate steps must be made. First, the signal is convolved with the MTB recorded impulse responses for the corresponding channel. For this purpose, the software jconvolver, a Linux based convolution software for the JACK Audio Connection Kit, is used. The 8 or 16 resulting signals (depending on whether an MTB playback should be reproduced with 8 or 16 channels) are transferred to the MTB Renderer. In the next step the rendered stereo signal is modified by the headphone compensation filter of the extraaural headphone to adjust the spectral influences. Afterwards the signal is spectrally linearized with the diffuse field filter of the MTB microphone and also equalized with the CTF of the artificial head FABIAN, which leads to a slight increase in the high frequencies. The stereo signal then flows into the subjects' headphones with an attached head tracking sensor. The motion data is transmitted to the MTB Renderer and the participant can thus experience a virtual acoustic scene.
The evaluation of the data was conducted with a script implemented by Alexander Lindau. See A.1 to find the code and measuring data. The results can be found in section 3.

B SAQI

To return to the evaluation method of descriptive statistics, it will be briefly discussed why the median as a measure of central tendency and bootstrap confidence intervals are used. The median value represents the halving of a frequency distribution and contributes to a minimum of the absolute deviation of a sought value. This means that outliers are of less importance than in the calculation of the mean. Since outliers are among the ratings for almost all SAQI items (this can have different reasons, for example a misunderstanding of the task), it is advisable to use the median in this case. The bootstrap method can be used if the distribution of the sample of a study is unknown (not normally distributed). The method empirically calculates a distribution function that relates to one concrete, empirical investigation. In the process, many more 'boostrap samples' (here $n_{boot} = 2000$) are drawn from the original sample with replacement, which are used to calculate the value of the test statistic. Thus, even without a normal distribution of residuals (as in this case), the bootstrap confidence intervals allow a representative estimate of the variance. [27]

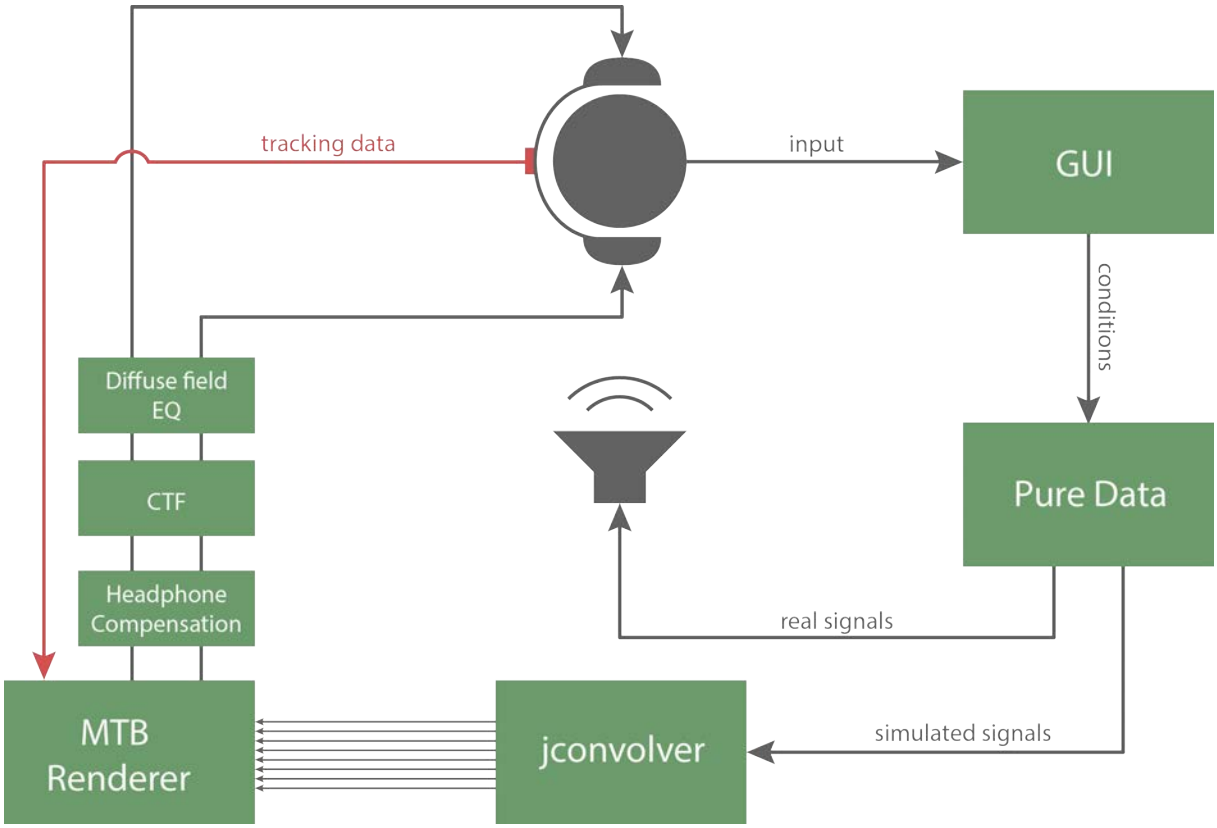


Figure 18: Signal flow in the plausibility listening test

The attached MATLAB script `SAQI_plots.m` (section A.1) computes the resulting graphs with median values and 95 % bootstrap confidence intervals for the reasons just mentioned. However, it also allows a representation of the means with 95 % bootstrap confidence intervals. The appendix (A.1) also contains the plots of residuals' distributions for each SAQI item, the correlation between the SAQI items, the deviations of the values from 0, the evaluation of the normal distribution for each condition and the result graphics. The results of the MANOVA calculation with the software SPSS including all value tables and profile plots as well as the syntax are also attached (see A.1: `Output_SAQI_Paper.spv`).

In addition to the MTB microphone and the artificial head FABIAN, a recording with a Behringer measuring microphone was investigated as an anchor signal when evaluating with the SAQI. The analysis of the anchor data and three further SAQI items (difference, front-back position, horizontal direction) are not included in the paper and will now be evaluated in this section.

For the listening test, the impulse responses were recorded for two speaker positions in the Electronic Studio (V = 140 m^2, RT $\approx$ 0.26 s) and the Audimax (V = 8500 m^2, RT $\approx$ 2 s) of the TU Berlin. The coordinates of the positions can be found in table 2. At the receiver position the impulse responses with the MTB array (test device) and a Behringer measuring microphone (anchor) as well as the BRIRs with FABIAN were recorded. As with the first listening test, a script from AKtools [26] was used for the measuring and the excitation signal was again a sine

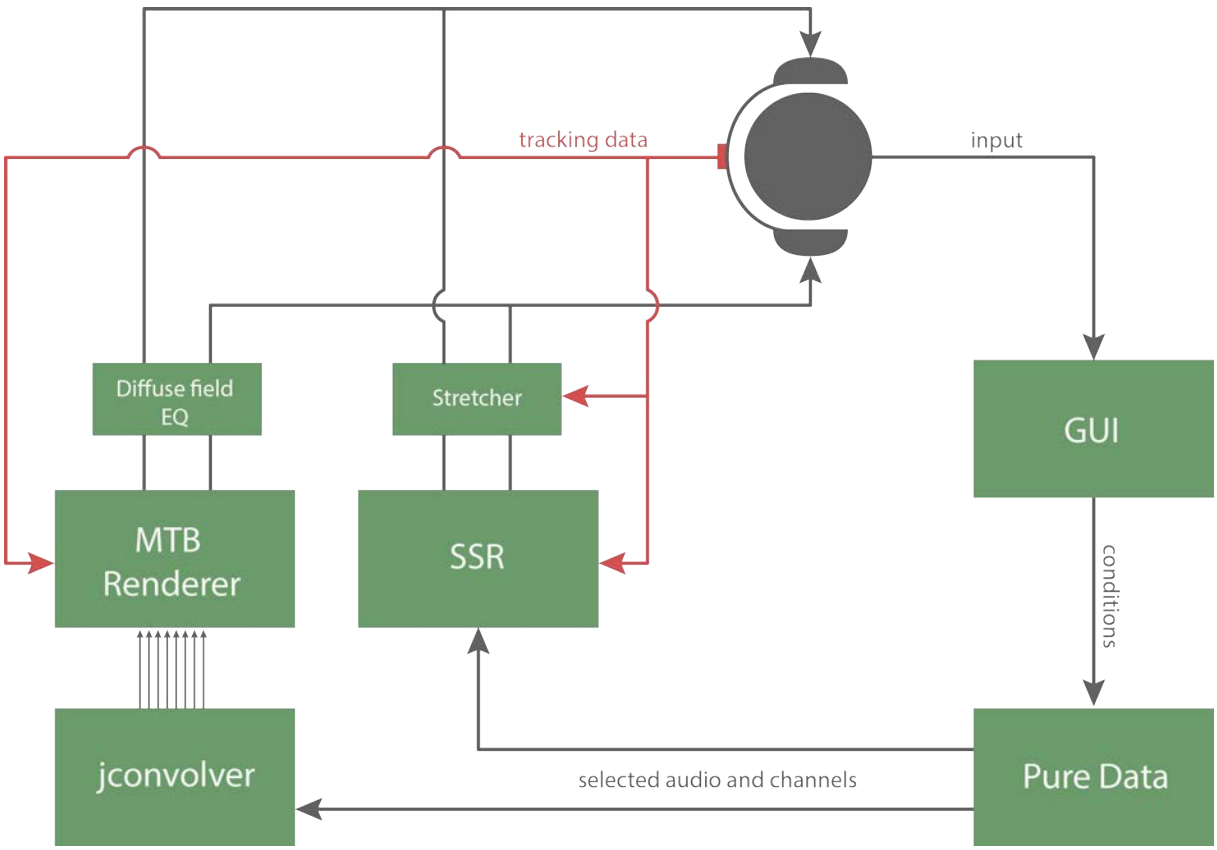


Figure 19: Signal flow in the SAQI listening test.

sweep of the 18th order. The impulse responses and BRIRs can be found in the attachment (section A.1). The anechoic speech and noise signal for the reproduction of the stimuli was also provided by the Audio Communication Group (see section A.1).

In the SAQI listening test the MTB signals were not compared to real signals, but to a dynamic binaural synthesis. The experimental setup is shown in figure 19. After entering the ratings in the whisPER GUI, the selected conditions will be transmitted via OSC to the Pure Data Patch. Here, the corresponding channels and audio signals are selected. If the reference signal is to be played, the convolution of the BRIRs, provided in SOFA files, with the anechoic audio signal and the rendering is conducted in the SoundScape Renderer. The resulting stereo file is transmitted to the ITD-Stretcher software (see section A.1), which adjusts the interaural time difference to the individual head size of the subjects. Thus, an optimal dynamic binaural synthesis can be applied as reference. The SSR and the Stretcher software are supplied by the user's tracking data. In case of a test signal transmission, the jconvolver is used for convolving the impulse responses with the anechoic audio material. After that the 8 or 16 MTB signals are transmitted to the MTB Renderer, which is also supplied with the tracking data of the person's head movements. Then the stereo signal is equalized with the MTB diffuse field filter and transferred to the already diffuse field equalized headphones.

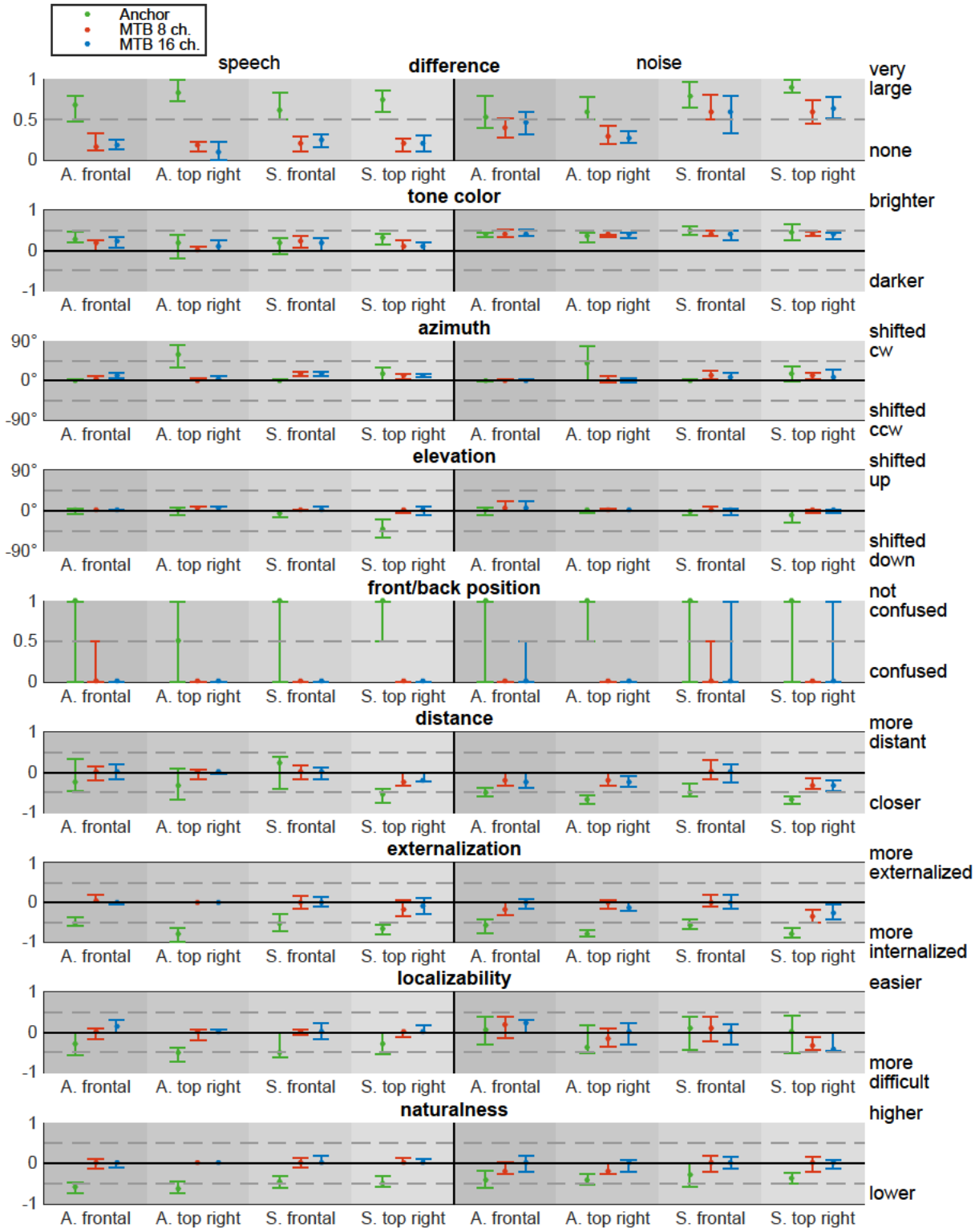


Figure 20: Median values with 95% bootstrap confidence intervals. A = Audimax TU Berlin, S = Electronic Studio TU Berlin.

The input data of the test persons used for the following evaluation is included in the appendix (section A.1). The edited data is also available as Excel file for further processing. For the analysis, the median values with 95% bootstrap confidence intervals were calculated for each condition with the MATLAB script `SAQI_plots.m` (see section A.1). Figure 20 shows the test results.

A detailed written evaluation of the results should not take place at this point. However, below is an explanation why the missing values in the paper were not included. Three SAQI items were excluded for publishing in the paper. For one thing, this is the difference. This item served as an introduction to the listening test, just to verify, if there was any difference between the test signals and the reference. With respect to the source position on the azimuthal plane, a larger difference was perceived between the test signals and the reference. When inspecting the data, a measurement error was detected in the recording of the impulse responses. The artificial head and the MTB microphone had an offset of 6° on the horizontal plane. After adjusting the data, a difference in position up to a median value of 15° cw (frontal source in the Studio, speech signal, both test signals) was still existing. This may have been due to the fact that the subjects were unable to estimate this difference to an exact degree and probably tended to have a higher rate. These results are therefore not representative of the test signals. When investigating the front-back position, the results also indicate a measurement error. The values vary widely and are contrary to the results of horizontal and vertical direction. In a conversation with the subjects after the listening test, it became apparent that the meaning of this SAQI item was partly misunderstood. Because of this, no reliable statement about the front-back position can be made here either. Furthermore, the anchor signal ratings were excluded because they are of no relevance for evaluating the quality of the test signals.

8. How to use the MTB system?

8.1. Hardware

The MTB array consists of a reverberant sphere, similar to a human head, with a diameter of d = 17.6 cm and 16 microphones mounted on the horizontal circumference. At the bottom of the sphere, a multicore cable with 16 XLR connectors is installed. There is also a 3/8-inch thread, which allows the MTB microphone to be screwed onto a microphone stand. In addition, a ball joint is available, which can be set between the microphone stand and the MTB Array.

As preamplifiers two Prisonus DigiMax DP88 devices are used. They supply the microphones with phantom power and convert (a/d) the incoming signals. The XLR connectors of the MTB array have to be be inserted into the 8 analog XLR inputs of each preamp. The channels 1-8 are inserted in the upper, the channels 9-16 in the lower amplifier from right to left. For synchronization, an RME Fireface UFX is used as master device. In this configuration the upper amplifier must be connected via ADAT 1 (In/Out) and the lower amplifier via ADAT 2 (In/Out) to the Fireface. The setting of the gain level is done digitally at the preamps and thus allows

a very exact configuration. The Fireface can be connected via USB to a laptop where further settings can be made with the RME software. The incoming signals can be recorded by a Digital Work Station (DAW).

8.2. Rendering software

The MTB Renderer is a C++ based software implemented for Linux systems. Detailed instructions on how to install the software and an explanation of the setting options are given in [28]. As head tracking system, however, a Polhemus head tracker was used here. For the MTB microphone presented in this document, the following settings are optimally used in the MTB Renderer. Since the array is limited to a microphone number of 16, either 8 or 16 input channels can be selected here. The qualitative difference is small and can be retraced in section 3. The interpolation method used is the Spectral-Interpolation Restoration (LP Linear, HP STFT), which achieved in [11] the highest quality from the five selectable algorithms regarding to the perceived plausibility. The settings for the STFT are 128 samples block size and a Hanning window. Depending on the number of input channels, different cutoff frequencies must be entered according to equation 3. For $N = 8$ you get the cutoff frequency $f_{max} = 1250$ Hz, for $N = 16$ a cutoff frequency of $f_{max} = 2500$ Hz. It should be noted that the MTB Renderer can only assimilate audio signals with a sample rate of 44.1 kHz. If the recording system described in this section is to be used for live streaming, it should also be noted that the Fireface UFX operates only in CC mode when connected to a Linux system. However, this mode only allows an ADAT transmission of 8 channels.

List of Figures

List of Tables

References

[1] Wightman, F. L. and Kistler, D. J. (1990): "Hearing in three dimensions: Sound localization." In: *Proc. of the AES 8th International Conference*, Washington, pp. 21-26.

[2] Algazi, V. R.; Duda, R. O. and Thompson, D. M. (2004): "Motion-Tracked Binaural Sound." In: *J. Audio Eng. Soc.*, **52**(11), pp. 1142-1156.

[3] Blauert, J. (1997): *Spatial Hearing: The Psychophysics of Human Sound Localization.* Cambridge, MA, USA: MIT Press.

[4] Wightman, F. L. and Kistler, D. J. (1992): "The dominant role of low-frequency interaural time differences in sound localization." In: *J. Acoust. Soc. Am.*, **91**(3), pp. 1648-1661.

[5] Algazi, V. R. and Duda, R. O. (2002): "Approximating the head-related transfer function using simple geometric models of the head and torso." In: *The Journal of the Acoustical Society of America*, **112**(5), pp. 2053-2064.

[6] Algazi, V. R.; Avendano, C. and Duda, R. O. (2001): "Elevation localization and head-related transfer function analysis at low frequencies." In: *The Journal of the Acoustical Society of America*, **109**(3), pp. 1110-1122.

[7] McAnally, K.; Martin, R. (2014): "Sound localization with head movement: Implications for 3-d audio displays." In: *Frontiers in Neuroscience*, **8**, pp. 1-6.

[8] Møller, H. (1992): "Fundamentals of binaural technology." In: *Applied Acoustics*, **36**, pp. 171-218.

[9] Lindau, A.; Hohn, T. and Weinzierl, S. (2007): "Binaural resynthesis for comparative studies of acoustical environments." In: *Proc. of the 122nd AES Convention*, Vienna, preprint no. 7032, 10 pp.

[10] Lindau, A. and Weinzierl, S. (2012): "Assessing the Plausibility of Virtual Acoustic Environments." *Acta Acustica united with Acustica*, **98**(5), pp. 804-810.

[11] Lindau, A. and Roos, S. (2010): "Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings. *Proc. of the 26th Tonmeistertagung*, Leipzig, pp. 680-701.

[12] Lindau, A.; Erbes, V.; Lepa, S.; Maempel, H. J.; Brinkman, F. and Weinzierl, S. (2014): "A spatial audio quality inventory (SAQI)." *Acta Acustica united with Acustica*, **100**(5), pp. 984-994.

[13] Algazi, V. R.; Avendano C. and Duda, R. O. (2001): "Estimation of a Spherical-Head Model from Anthropometry." In: *J. Audio Eng. Soc.*, **49**(6), pp. 472-479.

[14] DIN 33402-2 (2005): "Ergonomics – Human body dimensions – Part 2: Values." Deutsches Institut für Normung.

[15] Melick, J. B. et al. (2004): "Customization for Personalized Rendering of Motion-Tracked Binaural Sound." In: *Proc. of the 117th AES Convention.* San Francisco, preprint no. 6225, 20 pp.

[16] Kumar, S. (2003): "Selective Laser Sintering: A Qualitative and Objective Approach." In: *The Journal of The Minerals, Metals & Materials Society*, **55**(10), pp. 43-47.

[17] Hom, R. C.-M.; Algazi, V. R. and Duda, R. O. (2006): "High-Frequency Interpolation for Motion-Tracked Binaural Sound." In: *Proc. of the 121st AES Convention.* San Francisco, preprint no. 6963, 13 pp.

[18] Lindau, A. and Weinzierl, S. (2006): "FABIAN – An instrument for softwarebased measurement of binaural room impulse responses in multiple degrees of freedom." In: *24. Tonmeistertagung - VDT International Convention*, Leipzig, 5 pp.

[19] Schultz, F.; Lindau, A.; Makarski, M. and Weinzierl, S. (2011): "An extraaural headphone for optimized binaural reproduction.*Proc. of the 26th Tonmeistertagung*, Leipzig, pp. 702-714.

[20] Ciba, S.; A. Wlodarski and H. J. Maempel (2009): "WhisPER – A new tool for performing listening tests." In: *Proc. of the 126th AES Convention*, Munich, 8 pp.

[21] Lindau, A.; Estrella, J. and Weinzierl, S. (2010): "Individualization of dynamic binaural synthesis by real time manipulation of ITD." In: *Proc. of the 128th AES Convention.* London, preprint no. 8088, 10 pp.

[22] Lindau, A.; Kosanke, L. and Weinzierl, S. (2010): "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses." In: *Audio Engineering Society Convention 128*, London, 17 pp.

[23] Geier, M.; Ahrens, J. and Spors, S. (2008): "The SoundScape Renderer: A Unied Spatial Audio Reproduction Framework for Arbitrary Rendering Methods." In: *Proceedings of the 124th AES Convention*, Amsterdam, 6 pp.

[24] Bögelein, S.; Brinkmann, F.; Ackermann, D. and Weinzierl, S. (2018): "Localization cues of a spherical head model." In: *Fortschritte der Akustik - DAGA 2018* , pp. 347-350.

[25] Sennheiser Electronic Corporation: "Facts about electret wiring." https://cs.uwaterloo.ca/ mannr/Export/mke2wiring.pdf (last accessed on 14.05.2018)

[26] Brinkmann, F. and Weinzierl, S. (2017): "AKtools – eine offene Software zur Erhebung, Verarbeitung und Inspektion akustischer Signale. In: *Fortschritte der Akustik - DAGA 2017*, pp. 210-213.

[27] Bortz, J. and Schuster, C. (2005): *Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien.* Berlin, Heidelberg: Springer.

[28] Roos, S. (2011): "MTB - Ein System zur richtungstreuen Übertragung von akustischen Szenen. Dokumentation." Version 1.2.2, 9. February 2011.

A. Appendix

A.1. Contents of the attached DVD

Folder 01_Hardware

.../Capsules: This folder contains the MATLAB script `Selection_KE14.m` for matching the KE14 capsules. The raw data of the 162 capsules, received from Sennheiser, can be found in the Excel file `Data_162xKE14.xls`. The remaining Excel files contain the prepared data for the MATLAB Script to calculate the normalized and unnormalized capsule Sets. Furthermore the construction drawings of the converter circuit board and the elastic capsule mount as well as the electret wiring configuration are included.

.../Data_Sets_Sphere: In this folder the manufacturing data sets for the Selective Laser Sintering of the sphere and for the laser cutting of the steel sheet case, created by Frank Kohlrusch, are located.

Folder 02_Diffuse_field_eq

The script `Specs.m` calculates the frequency spectra from the sweeps in **.../MTB_Chirps** recorded with the MTB array. The `Diffusefieldequalization.m` script computes the desired diffuse field equalization filter for the MTB Array.

.../MATLAB_functions: The folder includes functions used by the MATLAB scripts.

.../MATLAB_tables: The script `Diffusefieldequalization.m` uses the MATLAB tables `CalMicSpecindB.mat` (frequency response of calibration microphone) and `CalSweepSpecindB.mat` (frequency response of the anechoic chamber, recorded with the calibration microphone). It also uses `fspec.mat` (MATLAB table with frequency bins) and `Specs_linear_0to180` (contains the frequency spectra of all capsules for all directions from 0° to 180° relating to the sound incidence direction in 2.8125° resolution), created by `Specs.m`.

.../MTB_Chirps: The included wave files contain the continuous sine sweeps recorded with the MTB microphone in the anechoic chamber for all capsules with a resolution of 2.8125°.

.../Wavefiles: Sample audio files that can be used by `Diffusefieldequalization.m` to create filtered audio examples.

Folder 03_Listening_tests

.../Backup_experimental_PC: The folder includes a backup of the data on the external computer used for the listening tests. It contains all shell scripts, audio material, configura-

tion files, impulse responses, equalization filters and software that were used for both listening tests. See the subfolders **.../Dokumente/MTB** for the plausibility listening test and **.../Dokumente/MTB_SAQI** for the SAQI listening test.

.../Plausibility: The MATLAB script `plot_hv_results_2017_hit_MTB.m`, implemented by Alexander Lindau, is used for analysing the measured values (subfolder **.../results**) in the plausibility listening test. Furthermore the result figures and the instruction for the subjects can be found here.

.../Questionnaire: The folder contains the questionnaire for the subjects as well as a table with the answers.

.../SAQI: The MATLAB script `SAQI_plots.m` evaluates the measured data of the SAQI listening test using the MATLAB tables `saqi_values_documentation.mat` and`saqi_values_paper.mat` and creating the also included result figures. The Excel file `results.xlsx` contains all ratings obtained in the listening test. The instructions for the participants can also be found here. The subfolder **.../TSD** includes the required information from the whisPER TSD files as MATLAB tables. The subfolder **.../pretest** contains diagrams created by the Matlab scripts `d_SAQI_1.m` and `d_SAQI_2`, implemented by Fabian Brinkmann. On the one hand, there are figures of item correlation and zero deviation. On the other hand, there are the graphics for checking the normal distribution of all conditions for each item as well as the graphics for checking the residual distribution for each item.

.../SPSS: The folder includes the SPSS value tables and output files for the plausibility t-test and the SAQI MANOVA.

Folder 04_Literature

The folder includes all literature used for the master thesis as well as the bibliography as .txt file.

Folder 05_Pictures

For documentation and to get an insight into the application of the MTB microphone, this folder contains pictures of the hardware, measurement setups and recordings settings.

Table 7: Gain control range of the preamplifiers

	Preamp No. 2451	**Preamp No. 2452**		**Preamp No. 2451**	**Preamp No. 2452**
Gain value [dB]	Output [dBFS]	Output [dBFS]	Gain value [dB]	Output [dBFS]	Output [dBFS]
0	-61,4	-61,1	31	-28,6	-28,2
1	-60,3	-60,0	32	-27,6	-27,2
2	-58,9	-58,6	33	-27,3	-27,0
3	-57,5	-57,2	34	-26,3	-26,0
4	-56,2	-55,9	35	-25,3	-25,0
5	-54,9	-54,6	36	-24,2	-23,9
6	-53,7	-53,3	37	-23,2	-22,9
7	-52,5	-52,1	38	-23,6	-23,2
8	-51,4	-51,0	39	-22,5	-22,2
9	-50,2	-49,9	40	-21,5	-21,2
10	-49,1	-48,7	41	-20,5	-20,2
11	-48,0	-47,6	42	-19,5	-19,2
12	-46,9	-46,6	43	-18,4	-18,1
13	-45,9	-45,5	44	-17,4	-17,1
14	-44,8	-44,5	45	-16,6	-16,3
15	-43,7	-43,4	46	-15,8	-15,5
16	-42,7	-42,4	47	-15,0	-14,7
17	-41,7	-41,4	48	-14,2	-13,9
18	-40,7	-40,4	49	-13,2	-12,9
19	-39,8	-39,4	50	-12,2	-12,0
20	-38,8	-38,4	51	-11,3	-11,0
21	-37,8	-37,5	52	-10,4	-10,1
22	-36,8	-36,5	53	-9,4	-9,2
23	-35,9	-35,5	54	-8,5	-8,2
24	-35,0	-34,6	55	-7,5	-7,2
25	-34,0	-33,6	56	-6,5	-6,3
26	-33,0	-32,7	57	-5,5	-5,3
27	-32,0	-31,7	58	-4,6	-4,4
28	-31,3	-31,0	59	-3,7	-3,5
29	-30,3	-30,0	60	-2,5	-2,4
30	-29,6	-29,2			