Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings

(Perzeptive Evaluation von Diskretisierungs- und Interpolationsansätzen für 'motion-tracked binaural' (MTB-)Aufnahmen)

Alexander Lindau, Sebastian Roos***

* TU Berlin, Fachgebiet Audiokommunikation, alexander.lindau@gmx.de
** TU Berlin, Fachgebiet Audiokommunikation, seba@lidsquid.de

Abstract

In 2004, motion-tracked binaural (MTB) sound was introduced as a new method for capturing, recording, and reproducing spatial sound. MTB uses a circular array of microphones on a rigid sphere with the diameter of an average head. Recordings are played back via headphones while tracking the listener's head movements in the horizontal plane. We conducted a listening test with 26 subjects assessing the reproduction's plausibility. Using a MUSHRA-like comparative test paradigm we tested the effect of the number of microphones (8, 16, 24, 32), interpolation scheme (5 methods), and audio content (noise, music, speech). Although there was a clearly superior configuration, the plausibility of MTB reproduction was found to be highly interdependent on all three parameters.

1. Introduction

In the context of virtual acoustic environments (VAEs) immersive recordings of ambient sounds are becoming more relevant. The 'acoustic atmosphere' of VAEs suffers from a lack of naturalism as typically the sound from a surrounding audience, nature events or sounds of technical origin (traffic, HVAC) – which one would expect to occur in reality – are not rendered. Thus, listeners sporadically complain about an 'aseptic feel' of the acoustic simulation. Moreover, in a recent study [1], indications for the particular relevance of social and emotional aspects for the experience of live music performances were found. An authentic simulation of environmental noise – of both natural and technical origin – might thus increase perceived realism of VAEs.

With common approaches to holophonic sound field reproduction, as for instance wave field synthesis (WFS) and higher-order Ambisonics (HOA) [2], the physically authentic reproduction of arbitrary ambient sound fields over extended listening areas would require densely spaced microphone and loudspeaker arrays. Therefore, the costs to render acoustic sceneries or 'soundscapes' in an ecologically valid manner would be considerable. For WFS an approach to represent ambient scenes aiming primarily at plausibility than at physical

ISBN 978-3-9812830-1-3

authenticity has been discussed in [3]. There, ambient scenes are constructed from a combination of plane waves and focused point sources. Plane waves are used to render a perceptually diffuse sound field (i.e. a 'background atmosphere'), whereas the focused sources, which are virtually distributed inside the listening, area are used to render 'Sound Particles' i.e. singular sound events as handclaps, raindrops or footsteps. Results from perceptual evaluation have though not been presented so far.

An alternative to holophonic sound field synthesis is dynamic binaural synthesis ([4], [5]). It is based on sets of binaural impulse responses (BIRs) which have been collected for several densely spaced head orientations using an adjustable head and torso simulator (HATS). These impulse responses completely describe how an acoustic signal is transformed on its way from the sound source to the entrance of the ear canals. The BIRs therefore contain all relevant monaural and binaural localization cues as e.g. spectral coloration, and frequency dependent inter-channel phase and magnitude differences. If these impulse responses are convolved with anechoic audio the resulting binaural signals will induce a realistic impression of the original sound field. By means of a time variant convolution algorithm the impulse responses can be updated according to the listeners current head orientation. A head tracking device is then needed to monitor the listener's current head orientation. Adapting to the listener's head orientation provokes several perceptual advantages: Firstly, perceived source positions remain stable with head movements as with normal hearing, and secondly, responsiveness to head movements efficiently prevents front-back-confusion and in-head-localization which are frequently observed during playback of static binaural recordings [6]. Binaural signals are mostly reproduced using headphones, though – by use of cross talk cancellation – transaural reproduction alternatives do exist [7]. When headphones are used, care has to be taken to avoid spectral coloration by applying appropriate frequency response compensation ([8], [9]).

Using the impulse response based approach applied in dynamic binaural synthesis also for rendering a binaural ambience would imply an unfeasible effort, as a) each singular source in the ambient scene would have to be modeled by binaural transfer functions of e.g. a loudspeaker, and b) each individual source's signal would have to be available as anechoic recording. A more pragmatic approach to incorporate ambient sounds in a dynamic binaural simulation environment was presented for instance in [10]. There, dynamic binaural ambiences were produced for a car sound simulator. A real car ride was recorded several times using a dummy head with an adjustable head, which was panned to a different horizontal head angle at each ride. This way a set of binaural recordings was obtained for a coarse grid of head orientations. For smooth interactive playback (involving cross fading between these discrete tracks) each car ride needed to be repeated under nearly identical conditions (speed, weather, road surface, interfering traffic), which in praxis was only possible using a dedicated test track. Recordings conducted this way would have to be synchronized before playback, which can be assumed a tedious task. For convenience in [10] though, only looped 30 second excerpts were used, during which the car noise was assumed to be stationary. In sum, this approach appears to be rather impractical for the representation of dynamic ambiences in a binaural VAE.

In 2004 though, Algazi et al. [11] proposed a new binaural recording technique – motion tracked binaural sound (MTB) – which promises an efficient solution for the recording and rendering of dynamic binaural ambient scenes. The MTB recording device consists of a rigid sphere with the diameter of an average head. At the horizontal circumference of the sphere a number of microphones are mounted evenly distributed on the surface (cf. Fig. 1). By means

of interpolation (i.e. cross-fading) algorithms, the multichannel signal recorded with the MTB array can be used to reconstruct two audio signals in a way as if they were recorded approximately at a listener's ear position. The rigid sphere of the MTB array acts as an obstacle for sound propagation introducing frequency dependent interaural level (ILD) and time (ITD) differences which are similar to those occurring in binaural hearing. The MTB reproduction can graphically be thought of as listening through headphones to the Schoeps spherical microphone [12] being rotated according to one's current head orientation. Depending on the complexity of the interpolation algorithm used, the MTB signal can be played back in real-time adapting to head orientation, leading to the same perceptual advantages as outlined for dynamic binaural synthesis.
By allowing recording directly 'in the field', MTB avoids the inherent problems of the impulse response based approach in capturing ambient scenes. This advantage though has to be traded against a limited flexibility, as audio content and acoustic scene setup cannot be manipulated after recording.

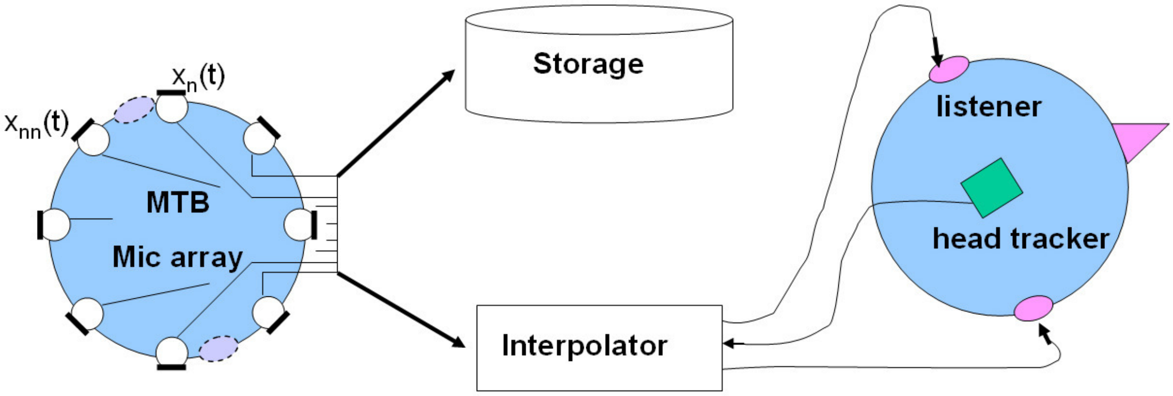


Fig. 1: Sketch of MTB reproduction principle: a head tracker is used to determine the position of the listener's ears and the two microphone signals $x_n(t)$ and $x_{nn}(t)$ that are closest to that position. The ear signals are constructed by interpolating between these microphones while following head movements. Ear signals can be reconstructed instantaneously or off-line using stored MTB recordings (after [13]).

A thorough analysis of physical and psychoacoustic properties of MTB approach has been given in [11]. For better understanding a review of some relevant points will be given though. Algazi et al. [11] summarize important auditory cues exploited for spatial hearing:

- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Spectral cues introduced by the pinna
- Torso reflection and diffraction cues
- Direct to reverberant ratio
- Changes of the above mentioned cues due to head movements
- Familiarity with the sound of the source

Additionally, the above mentioned acoustic cues do vary with azimuth and elevation of source and receiver, as well as with range and with frequency.

The most important interaural cues exploited for localization ITD and ILD have been described within Lord Rayleigh's duplex theory [14]. Accordingly, there is a consensus that ITD is evaluated below approximately 1.5 kHz for determining horizontal source localization, whereas above that frequency, mostly ILD cues are evaluated. For wide band stimuli containing conflicting ITD and ILD cues, Wightman and Kistler [15] showed the dominance of the ITD in determining source localization.

Elevation of sound sources is detected by evaluating monaural spectral differences introduced by torso and pinna [15]. A basic familiarity with the source sound spectral content is though a prerequisite for correct perception of elevation. Therefore, if pinna geometry differs from that of the listener, or, as in the case of MTB pinnae are completely absent, errors in determining source elevation can be expected (for more details on perceptual properties of MTB see sect. 1.2).

1.1. MTB discretization and interpolation

MTB exhibits some obvious major system parameters that can be chosen more or less deliberately. The most relevant parameters are: 1) the number of microphones, 2) the interpolation algorithm, 3) the array geometry, and 4) the angular position of the interpolated microphone signals. The impact of these parameters have been discussed by Algazi et al. [11] using a quantitative approach. With the help of graphs they discussed perceptual effects of five different approaches to interpolation while using 8, 16, or 32 microphones respectively. Perceptual ratings were though derived mostly informally. The five interpolation algorithms proposed in [11] will now be shortly explained following an order of increasing complexity.

1.1.1. Full Range Nearest Microphone Selection (FR-NM)

The probably easiest way to reconstruct the ear signals would be to simply select the signals of the microphones which are located next to the listener's ear positions. As a result, the acoustic space will be fragmented into N (N being the number of available microphones) angular sectors each of a size of $\Delta\varphi$ = 360°/N in which the signal is not adapted to the listeners head orientation. Thus, the acoustic image will 'jump' at the sectors' boundaries. Additionally, switching artifacts can be expected to become audible each time when crossing such a boundary.

1.1.2. Full Range Linear Interpolation (FR-LI)

The discontinuous behavior of the FR-NM method can be avoided by linearly cross fading between two adjacent microphones' signals. Then the signal x(t) at the ear's position can be interpolated from $x_n(t)$ being the output of the nearest microphone and $x_{nn}(t)$ being the output of the next nearest microphone (cf. Fig. 1) by (from [11])

$$x(t) = (1-w)x_n(t) + wx_{nn}(t)\,, \qquad (1)$$

where the interpolation weight w is determined as the ratio of the angle between the ear and the (currently) nearest microphone β and the average angular microphone distance $\Delta\varphi$:

$$w = \beta/\Delta\varphi\,. \qquad (2)$$

Linearly combining signals from closely aligned microphones is prone to comb filtering. Moreover, spectral coloration will vary with head movements and direction of sound incidence. For grazing sound incidence from an analysis of the resulting comb filter system Algazi et al. [11] derived a criterion for the minimum number of microphones needed to keep MTB's magnitude response within ±3dB deviation below a certain frequency f_{up}. Assuming an MTB radius of $r_{MTB} = 87.5$ mm and a sound velocity c_0, it states

$$N_{\min} = 8\pi r_{MTB} f_{up} / c_0 \ . \qquad (3)$$

Thus, when setting f_{up} to 20 kHz one would need a number of at least 128 microphones. As this is a rather large number, the next three interpolation methods exploit aspects of human hearing to achieve better sound quality in a more efficient manner.

According to equation 3, when using a number of at least eight microphones most severe comb filtering artifacts can be pushed beyond spectral regions of approximately 1200 Hz. Applying a low pass (anti-aliasing) filter could thus be used to eliminate comb filter artifacts from the interpolated MTB signal. Perceptually, this approach has the advantage that low frequency ITD cues dominating localization in the horizontal plane remain uncorrupted. High frequency ILD cues are though important, too. The following three proposed approaches employ such a two-band treatment but differ in the way the high frequency components are restored.

1.1.3. Two Band Fixed Microphone Interpolation (TB-FM)

A simple but quite crude approach would be to reproduce the higher frequencies from a high pass filtered fixed omnidirectional microphone ('complementary microphone'). This way ITD and ILD become zero above the crossover frequency. High frequency spectral energy is restored, but spatial information will be lost. The loss of high frequency ITD might be less disturbing as sensitiveness to high frequency phase differences is low. In contrast, erroneous high frequency ILDs will strongly disturb spatial auditory perception. Algazi et al. [11] informally reported the perception of 'split' acoustic images, with low frequency content reproduced correctly, while high frequency content is located inside the head. Additionally, the question arises, where the complementary microphone should be located. If it is mounted on the MTB sphere, it will not be omnidirectional. Depending on the composition of the acoustic scene and the amount of diffuse reverberation its exact location can have a strong effect on the overall sound color.

1.1.4. Two Band Nearest Microphone Selection (TB-NM)

High frequency ILD can be restored by applying the nearest microphone selection procedure (FR-NM) explained above to the high frequency audio range (cf. Fig. 2). This approach can be expected to improve stability of source localization, as high frequency ITD information is weak and ILD is not the dominating localization cue. The perceptual severity of high frequency sectoral switching will though be dependent on the audio content and the number of microphones used, as was confirmed also by our listening test results.

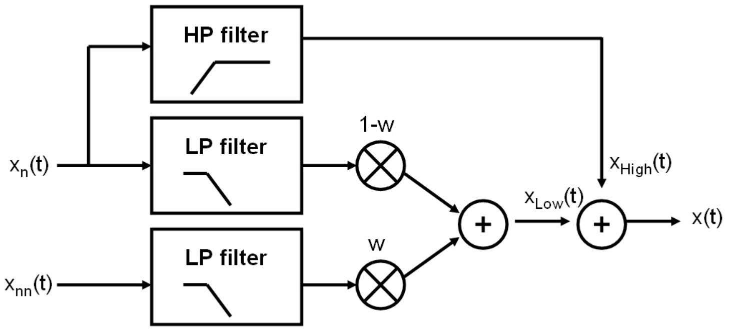


Fig. 2: Block diagram of TB-NM algorithm: Low frequency content is derived from continuously interpolating between nearest and next nearest microphone signals, $x_n(t)$ and $x_{nn}(t)$. High frequency content is derived through switching to nearest microphone signal $x_n(t)$ (after [13]).

1.1.5. Two Band Spectral-Interpolation Restoration (TB-SI)

Using short-time (fast) Fourier transform, linear interpolation can also be conducted in real-time in the spectral domain. With $M_n(\omega)$ and $M_{nn}(\omega)$ being the magnitudes of the short-time Fourier transform of the high frequency content of the microphone signals $x_n(t)$, and $x_{nn}(t)$ respectively (cf. Fig. 3), equation 1 becomes

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

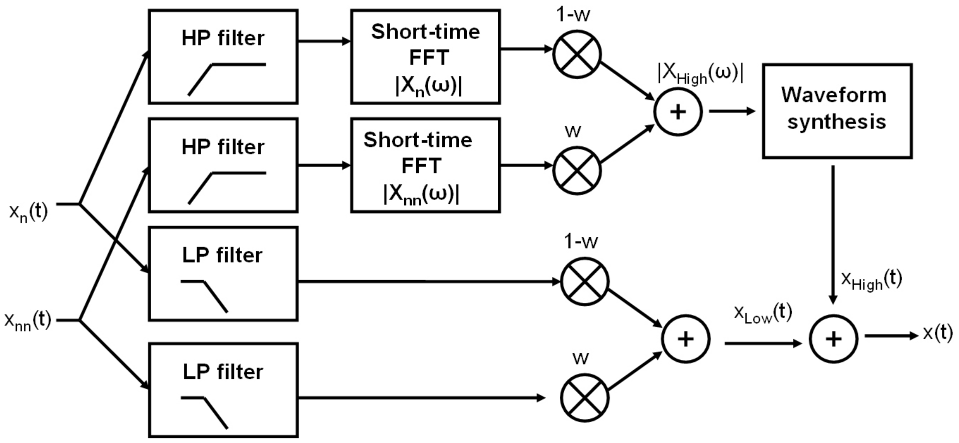


Fig. 3: Block diagram of TB-SI algorithm: Low frequency content is derived from continuously interpolating between nearest and next nearest microphone signals, $x_n(t)$ and $x_{nn}(t)$. High frequency content is derived through interpolation of short-time magnitude spectra and phase reconstruction during waveform synthesis (after [13]).

Three different procedures for how $x_{High}(t)$ (cf. Fig. 3) can be recovered are discussed thoroughly in [13]. The approaches have been discussed regarding their abilities to fulfill continuous block wise processing necessary for real-time operation mode, and how reconstruction of a temporal waveform is achieved. Hom et al. [13] evaluated two real-time methods: 1) weighted-overlap-and-add (WOLA), 2) a least-squares-error-estimation method applied on the short-time magnitude spectrum (LSEE-MSTFT, [16]) and 3) an offline method (LSEE-MSTFTM). All methods share the same analysis step: The short-time Fourier transform (STFT) is calculated for overlapping successive time segments using triangular or Hanning windows. The synthesis step (recreation of the time signal) though differs for all three approaches. Moreover, the two real-time methods only create an interpolated short-time magnitude spectrum, so a suitable phase spectrum has to be derived somehow. A simple but elegant approach pursued in [13] is to select the phase of the nearest microphone ('phase switching'). Hence, a MTB signal can be produced whose frequency response now varies continuously with angle.

1.2. Perceptual evaluation of MTB

Summarizing in how far each of the five interpolation approaches considers different psychoacoustic aspects (ITD and ILD reproduction, split acoustic imagery, spatial continuity with head movements), Algazi et al. [11] derived a perceptual ranking. From that they concluded that 'spectral interpolation' method should result in perceptively best reproduction, followed by 'two band nearest-microphone' method, whereas 'full band interpolation' is seen on the third rank. 'Full band nearest microphone' and 'two-band fixed microphone' methods both are considered worse but efficient in terms of bandwidth and computational effort. Authors did not formally assess this ranking, for instance in terms of an overall impression score (i.e. for plausibility, authenticity, or preference). Instead they referred to informal listening test results. Differentiating between speech and music applications they concluded that acceptability of the interpolation approaches would depend on the targeted content or application and would have to be traded against available bandwidth. Regarding the number N of microphones needed, they state generalizing that N does not have to be large to achieve a 'strong sense of realism and presence' [11]. Though, if N was small, interpolation artifacts might become objectionable. Moreover, spatial discontinuities were assumed to make reproduction unacceptable for musical content. They argued that, if the required sound quality was low and available bandwidth was limited, a number of N = 8 might suffice.

Further perceptual shortcomings of the MTB approach have been discussed in [11]. Possible remedies and means for the individualization have been explained by Melick et al. ([17], table 2). Most relevant issues were found to be due to a) the missing pinnae, 2) mismatch in head and array diameter, 3) shortcomings of the interpolation algorithms, 4) and mismatch of microphone and ear location. Effectiveness of remedies was though reported merely based on informal listening tests.

The missing pinnae result in both erroneous monaural high frequency spectral detail and high frequency ILD cues. Thus, in comparison to a head with pinnae, MTB signals will produce spectral coloration, erroneous elevation cues and horizontal localization mismatches. The missing pinnae though offer the advantage that an MTB array has no preferred spatial orientation within the horizontal plane. Hence, from a single MTB recording, ear signals for a plurality of listeners with individual head orientations can be rendered synchronously.

A mismatch between the MTB's and a listener's head diameter will produce erroneous ITD cues. As a result, horizontal source location will not be perceived as stable. If the listener's head diameter is smaller than that of the array, ITD cues are larger than natural, resulting in a perceived motion of the source in retrograde direction of the listeners head movements. If the head size is larger, the inverse effect occurs and is perceived as the source 'lagging behind' the head movements [11]. A treatment for this issue occurring also in impulse response based binaural rendering has been presented lately in [18]. Therefore, the choice of

a generic MTB diameter is crucial. Algazi et al. [11] proposed to use the 'traditional average value' of 175 mm, from which 98% of the U.S. American population deviate by ±15 %.
The possibility to interpolate the MTB signal at arbitrary in-between microphone positions gives a means to manipulate both the ITD and spectral content as both ears' virtual positions can be shifted forward and backward on the circumference. Withdrawing the ears from the position of maximum diameter (i.e. from the frontal plane) will decrease the low frequency ITD in the MTB signal [17]. When considering frontal sound incidence, moving ear positions backwards will lead to high frequency shadowing, while moving forward will increase pressure stasis thus leading to high frequency boosting. For the two-band interpolation algorithms Melick et al. [17] proposed the interesting approach to 1) move the virtual ear position for the low pass filtered path backwards on the circumference, this way achieving a – when compared to human physiology – more natural ear position, while 2) shifting the ear position for the high pass filtered path slightly forward compensating the missing pinnae with the pressure stasis effect. Theile [12] exploited the same effect when he proposed moving the Schoeps spherical microphone's capsule positions 10° backward from the position of maximum diameter to achieve both a flat diffuse and free field response.

1.3. Aims of this study

So far we have introduced MTB as a method for recording and interactive rendering of spatial acoustic scenes. We explained its basic technology, reviewed its perceptual shortcomings and remedies for them. We also showed that – though theoretically inferred perceptual impacts are mostly very plausible – results from perceptual evaluation of even the most relevant system parameters (number of microphones, interpolation strategies, and interaction with different audio content) are sparse and mostly informal. Therefore, in our study we wanted to assess MTB reproduction quality in an integrative manner and as a function of the number of microphones, the type of interpolation algorithm and the audio content.

2. Methods

To allow for a perceptual comparison in a real-time capable listening test environment several steps had to be taken. In order to render MTB live streams or recordings acquired with different numbers of microphones, two applications, a 'MTB renderer', and a 'MTB player' had to be programmed (cf. sect. 2.1). Secondly, we needed authentic MTB ambience recordings conducted with different numbers of microphones. These were synthesized from impulse responses measured using a virtual MTB microphone array and an acoustic scene made up from an arrangement of several loudspeakers (cf. sect. 2.2).

2.1. Implementation of parametrical MTB software

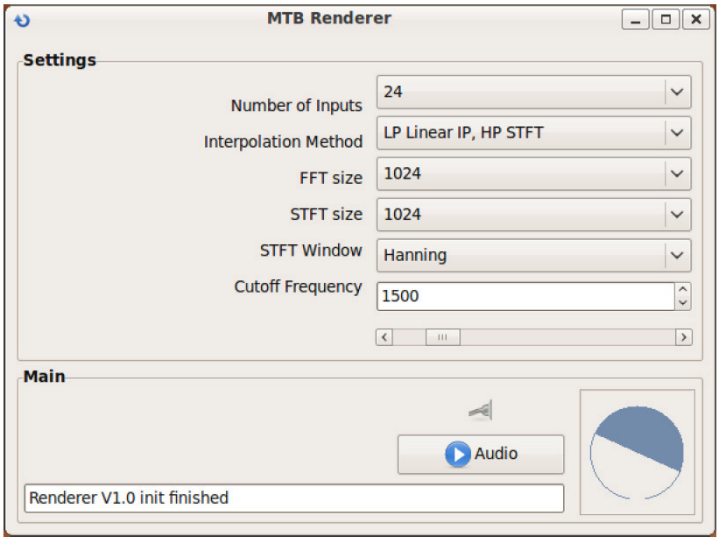
Our MTB renderer uses the Jack audio connection kit (http://jackaudio.org) as backbone for real-time audio streaming. After choosing the correct number of microphones, a suitable interpolation algorithm and, if needed, a crossover frequency, it can be used to instantly render audio signals from MTB arrays which are interfaced to the rendering computer. For head tracking we used an Intersense InertiaCube. The renderer can be controlled to instantly change the most relevant parameters (cf. Fig. 4) using a graphical user interface or the OSC remote control protocol [19]. Until now, it implements the five interpolation algorithms described in section 1.1:

▪ Full Range Nearest Microphone Selection:	FR-NM,
▪ Full Range Linear Interpolation:	FR-LI,
▪ Two Band Fixed-Microphone Interpolation:	TB-FM,
▪ Two Band Nearest Microphone Selection:	TB-NM, and
▪ Two Band Spectral-Interpolation Restoration:	TB-SI,

while supporting MTB arrays with four different numbers of microphones, namely 8, 16, 24, and 32.

As there was no clear recommendation in [11], for the fixed microphone (TB-MF) method we deliberately chose the microphone pointing into the array's frontal orientation for high frequency reconstruction. However, depending on the acoustic scene (i.e. within an outdoor soundscape) a 'frontal' orientation might not be definable. For convenience, when doing the recordings for our listening test (cf. sect. 2.2) we arranged the MTB array with the frontal microphone pointing at the stage end of the hall.

For the spectral interpolation method we chose the weighted overlap and add (WOLA) method with phase switching described in [13] as it resembles a straight forward implementation of the spectral interpolation as originally proposed by Algazi et al. [11] (cf. sect. 1.1). Being a block based implementation of the continuous spectral interpolation different analysis/synthesis windows widths and sizes can be chosen from (rectangular, triangular, Hanning, Blackman, Hamming). For our listening test we used WOLA with a 128 taps Hanning window and with 75% overlap.



MTB Renderer

Settings

Number of Inputs: 24

Interpolation Method: LP Linear IP, HP STFT

FFT size: 1024

STFT size: 1024

STFT Window: Hanning

Cutoff Frequency: 1500

Main

Renderer V1.0 init finished

Fig. 4: Graphical user interface of the MTB renderer application. Relevant MTB parameters can be chosen manually or via OSC control. The circular symbol on the lower right gives an instant visual feedback on received head tracking data.

As our listening test was intended to use pre-created MTB recordings to be directly comparable to each other, another dedicated application for feeding the MTB renderer, an MTB player, had to be implemented. The player software is also remotely controllable via OSC (OSC commands: load audio files, start/stop audio playback) as far as needed for our listening test.

2.2. Synthesizing MTB Recordings

To create the MTB recording for our listening test we chose the HE 101 lecture hall of the TU Berlin (RT = 1.67 s, V = 5200 m³). Basically it is a rectangular shaped room with a sloped seating area, which rises softly towards the back wall. The MTB microphone was placed centrally in the hall at approximately two meters above floor level at the location of a former front-of-house desk (cf. Fig. 5, left). To imitate an ambient sound scene, we arranged eight small active wideband loudspeakers (Fostex 6301, 10 cm diaphragm) more or less randomly in the hall. Loudspeakers were placed in different distances and heights regarding the MTB-microphone's position, each pointing at a different direction (cf. Fig. 5, middle). From this setup we would be able to create both simple and more complex acoustic scenes.

Instead of building four different MTB arrays we constructed a virtual MTB microphone using a singular microphone[1] attached at the circumference of a rigid plastic sphere of 180 mm in diameter[2] (cf. Fig. 5, right). This single-microphone array was mounted on a base that could be rotated with high angular precision using a servo motor device [20]. With this setup we measured the impulse responses for all eight loudspeakers to the MTB array. Each time the eight measurements were completed, the microphone was rotated another angular step with a step size according to the intended MTB array solution (i.e. by 360/8 degrees for the virtual 8-microphone-array) and the measurements were repeated. For a sound pressure level to give a reasonable signal to noise ratio, we had to high pass filter our measurement signal at 200 Hz to prevent damage of the rather small Fostex monitors. In the end we collected the impulse responses of all eight loudspeakers to all possible microphone positions of the four virtual MTB arrays. The measurement duration could be reduced as the equiangular 8 and 16 channel microphone arrays form a symmetric subset of the 32 channel MTB.

With the help of the Matlab [21] software the four sets of multichannel MTB impulse responses were (independently for each loudspeaker) convolved with selected anechoic audio stimuli. Thus we obtained 'virtual' multichannel MTB recordings which could instantly be rendered using the MTB player and MTB renderer applications. In summing up the corresponding MTB microphone channels of these virtual recordings for different loudspeakers, we could easily generate MTB recordings of several loudspeakers playing different audio material simultaneously. From a pool of stimuli we created this way, the following three scenarios were chosen for the listening test (cf. Fig. 5, middle):

1. a series of pink noise bursts of 4.5 s length emitted from loudspeaker 1,
2. some sentences of a German male speaker emitted from loudspeaker 1, and
3. a string quartet playing an excerpt from a modern tango piece (1st violin from loudspeaker 8, 2nd violin from loudspeaker 2, viola from loudspeaker 5, violoncello from loudspeaker 4).

Admittedly, the last scenario is not a very realistic setting, but the recordings of anechoic string quartet were at hand. It was though thought to merely represent a spatially distributed scene of multiple synchronously playing sources.

[1] Panasonic WM-61A omnidirectional back electret miniature microphone (20 - 20.000 Hz)

[2] This diameter was supposed to be reasonably close to Algazi 's [11] proposal of 175 mm.

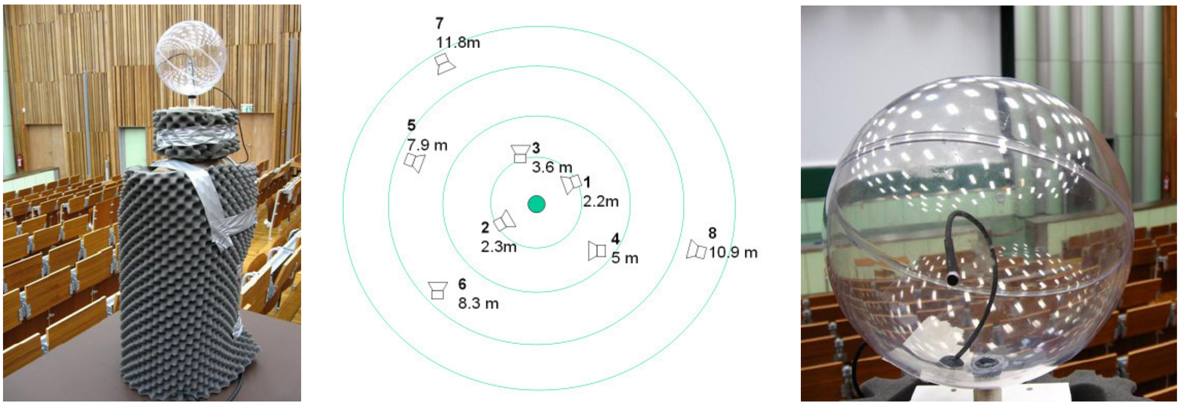


Fig. 5: Left: Virtual MTB mounted on a remotely rotatable basement in the middle of a lecture hall, middle: loudspeaker arrangement for impulse response measurements with the virtual MTB array (distances are given relative to virtual MTB's location). Right: Close up of the virtual MTB array showing the singular microphone mounted in the plastic sphere.

3. Listening Test

3.1. Operationalization of MTB sound quality

To assess the quality of MTB recordings, different criteria can be thought of. Typically, audio quality is defined as degradation or improvement with regard to a reference. If this reference can be provided externally, for instance if a reproduction can be directly compared to the corresponding real sound field, listeners are said to evaluate the *authenticity* of the stimulus, i.e. whether they perceive a difference between reference and reproduction ([22], pp. 373). Differences can be rated regarding either specific quality aspects (e.g. ratings of localization accuracy, perceived sound coloration, or latency) or in an integral manner (e.g. rating of overall preference). In the strictest sense, for us it would have meant to let subjects compare MTB recordings (presented via headphones) directly to the real sound field as perceived with their own ears. Two problems emerge with this approach. First, we would have had to conduct the listening test completely in the lecture hall, which was not possible, and secondly, subjects would have had to take the headphones off when listening to the real sound field making blind testing impossible. The only way for assessing authenticity of MTB recordings in a double-blind listening test would be to provide the acoustic reality via individual dynamic[1] binaural recordings. Only then, both, individual reality and MTB recordings could be compared instantaneously and in a controlled laboratory environment. Individual binaural measurements obviously represent a very time-consuming procedure but might become realizable in the future, especially if a suitable spatial interpolation is utilized [23]. Alternatively, we could have also used our adjustable FABIAN dummy head [20] to binaurally record the sound field for later off-scene comparisons. However, due to individual morphologic differences between FABIAN and each listener such so-called non-individual

[1] 'Dynamic' meaning individual measurements of binaural impulse responses conducted for a fine angular discretization of head orientations, suitable for interactive binaural re-rendering using a time-variant fast convolution algorithm [5].

recordings are known to induce deviation in perceived localization and timbre and would thus disturb an objective assessment of MTB recordings' authenticity.

Instead, we decided to test MTB quality using the less strict quality criterion of *plausibility*. Here, the listener is urged to relate his sensory impression of a stimulus to an internal reference, for instance to that of an equivalent natural acoustic event ([22], pp. 389). Plausibility is much easier to assess in a listening test as stimuli can be presented without the need to accompany them with a physically authentic reference. Therefore of course, ratings of plausibility can only give an indication of how close a reproduction is to reality. However, as at this point, we were mainly interested in quality differences *between* the alternatives for interpolation and discretization of MTB recordings, plausibility was regarded a sufficient criterion. In the end, with this approach we should be able to find the perceptually superior combinations of interpolation and discretization methods, which then, in the future, can be assessed again for authenticity using the above described method.

3.2. Listening test design

For the assessment of MTB's plausibility we chose a MUSHRA-like (multiple stimulus with hidden references an anchors, [24]) listening test design. As said before, the test involved three independent variables:

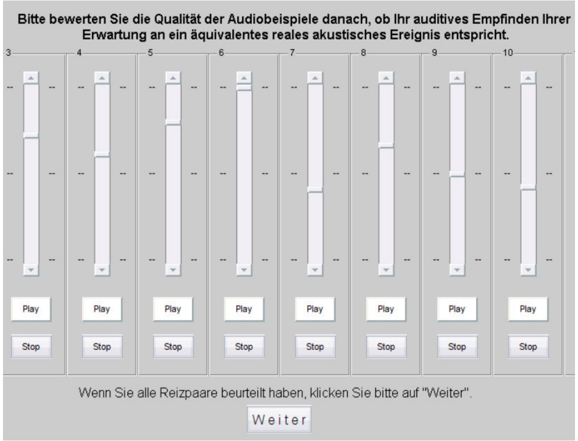
1) *DISCRETIZATION:* number of microphones (8, 16, 24, 32),
2) *INTERPOLATION:* type of interpolation algorithm (TB-FM, FR-NM, FR-LI, TB-NM, TB-SI), and
3) *CONTENT*: audio content and spatial scene composition (pink noise, male speech, string quartet).

The spatial arrangements related to the audio contents were described already in sect. 2.2. In a repeated measures design all 4 (*DISCRETIZATION*) x 5 (*INTERPOLATION.*) x 3 (*CONTENT*) = 60 stimulus combinations were assessed by each subject. Our design deviates from the original MUSHRA test in so far as a reference is not provided (cf. sect. 3.1). For a proper scale usage though, stimuli were included which were thought to serve as either very low[1] or high[2] audio quality anchors, respectively.

As the listening test involved only headphone presentation, it could be conducted in a typical seminar room lacking any special acoustic treatment. The used headphone model was a Sennheiser HD 800; a non-individual headphone compensation created from measurements on our FABIAN dummy head was applied. A head tracker (Intersense InertiaCube) was taped to the top of the headphones and connected with the rendering computer. Subjects were placed in front of a laptop displaying the listening test GUI (cf. Fig. 6) realized in Matlab [21]. All audio processing was done on a dedicated rendering computer (8-threaded IntelCore i7, Linux OS, 12 GB RAM). The MTB renderer was configured to realize the virtual ear positions at diametrically opposed positions of maximum sphere diameter. OSC messages controlling the listening test progress were sent via Ethernet from the laptop to the rendering computer. Audio signals were played back using a M-AUDIO Delta Audiophile 192 sound card.

[1] 8 channel version of Two Band Fixed-Microphone Interpolation (TB-FM)

[2] 32 channel version of Two Band Spectral-Interpolation Restoration (TB-SI)



Bitte bewerten Sie die Qualität der Audiobeispiele danach, ob Ihr auditives Empfinden Ihrer Erwartung an ein äquivalentes reales akustisches Ereignis entspricht.

Wenn Sie alle Reizpaare beurteilt haben, klicken Sie bitte auf "Weiter".

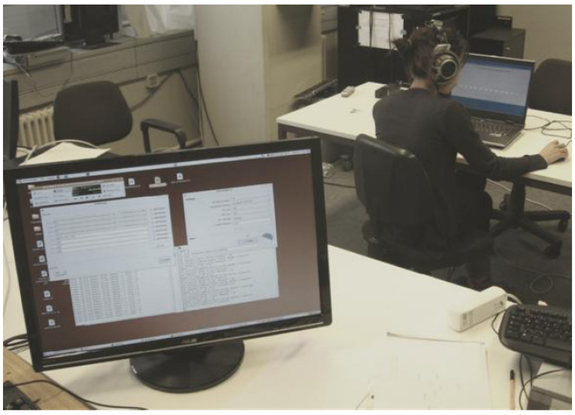


Fig. 6: Left: Graphical user interface for of MUSHRA test, the briefing sentence (in German) was always displayed at the top of the panel of sliders. Right: Listening test setup as seen from the operator's place.

The GUI consisted of a row of continuous sliders and 'play' and 'stop' buttons for each stimulus. Subjects could switch instantly between different stimuli, taking their time for rating at will. Stimuli were presented in six successive panels each displaying 12 sliders (in sum 72 presentations). Within two successive presented panels of sliders the audio content was kept constant. Panel order and slider assignment was randomized for each subject. The number of 72 presentations comes off as the 'high' and 'low' quality anchor stimuli were additionally included in each panel. At the beginning of the listening test, a training panel was presented to make each subject known to the variety of stimuli within the listening test.

Internal references can be strongly depending on individual expectations towards the presented content as well as on function and form of the reproduction system. Therefore the internal reference has to be carefully established by explicit briefing of test subjects. Using a written German text, plausibility was introduced as the amount a stimulus corresponded to the expectation of an equivalent real acoustic event. Listeners were instructed to rate each stimulus' plausibility in an integral manner using the sliders labeled from 'hervorragend' (excellent) to 'sehr schlecht' (very bad) at the corresponding ends (cf. Fig. 6, left). A briefing sentence repeating the definition of plausibility was displayed throughout the whole test procedure above the sliders of the rating panel (cf. Fig. 6, left). During training it was additionally emphasized that, as far as possible, before switching to the next panel, an order of plausibility should have been established between the stimuli in a singular panel.

3.3. A-priori hypotheses and determination of sample size

Basically we expected the following main effects: At first – as cross fade artefacts should be less audible – an increase in plausibility with decreasing stimulus bandwidth (i.e. plausibility should increase from noise stimulus over the string quartet to the speech stimulus), secondly, a plausibility increasing with the number of microphones and thirdly (in accordance with [11], cf. sect. 1.2) an increasing plausibility for the following order of interpolation algorithms: TB-FM, FB-NM, FB-LI, TB-NM, TB-SI (in the following also designated algorithms 1 to 5). We though also expected some interaction effects, for instance by assuming that for certain combinations of stimulus and interpolation algorithm from a certain point on an increasing number of microphones will not increase plausibility any further (saturation).

Using the GPower software [25] we calculated the sample size needed in the repeated measures design to test effects at a 5% type-1 error level with a power of 80% [26]. When assuming an average intercorrelation between subjects of $\rho = 0.25$, to test a small effect for least gradated main effect (*CONTENT*) 13 subjects would be needed. To test a small effect for the highest order interaction (*DISCRETIZATION* x *INTERPOLATION* x *CONTENT*) 29 subjects would be needed. Finally 26 subjects (84 % male) of an average age of 29.5 years took part in our test. Most of them had prior experience with listening tests, and a musical education of in average more than 10 years. The average test duration was approximately 40 minutes.

4. Results

Following recommendations in [24] we post screened our raw data. In order to visually identify subjects with ratings strongly deviating from the group's average, we boxplotted results independently for all 60 tested stimuli. Additionally, after checking normality of stimuli's ratings using Matlab's Lilliefors test, we conducted two-sided Grubb's outlier tests. Both tests were done at a 5 % type-1 error level. Subsequently we checked the distributions of individual ratings to identify subjects lacking variability in their ratings or showing clusters of extreme ratings. The Grubb's test found an increased number of outlier ratings for three subjects (8, 8 and 5 times) but as there were no other problems evident with them and 8 was considered a rather small fraction of 60 ratings, we concluded that there was no reason to exclude any subject from further analysis.

With the help of the SPSS software package we calculated the intraclass correlation [27] ICC(2, k) = 0.919 for the raw ratings as a measure of our subjects' agreement. This is a fairly high value; together with the differentiated picture we got from the test results, it serves as an indication, that in combination with our instructions subjects were able to rate stimuli's plausibility in a consistent manner. Internal references appeared to have converged fairly well across subjects.

To obtain an impression of the sensitivity of our test we calculated posterior effect sizes using the GPower [25] software. With ratings from 26 subjects sharing an average intercorrelation of $\rho_{emp} = 0.17$ we were able to test an effect size of E = 0.08 for the least gradated main effect (content). For the highest order interaction (*DISCRETIZATION* x *INTERPOLATION* x *CONTENT*) we were able to test effect sizes of E = 0.11. Hence, we achieved our aim of testing small effect sizes.

As our stimuli contained no intermediate quality anchor we followed recommendations from [28] to standardize individual ratings. All further results from inferential statistics were derived from these standardized ratings. Eventually we though found that statistical results from both raw and standardized data were nearly identical.

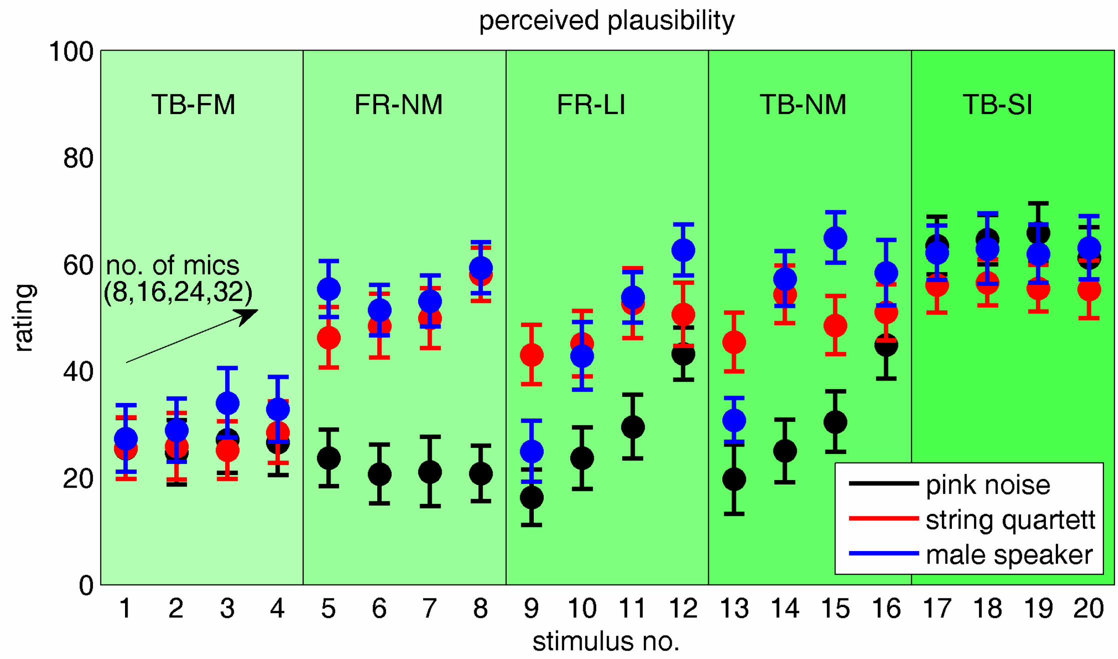


Fig. 7: Results from listening test assessing plausibility of different MTB recordings. Raw ratings are displayed as means and 95 % confidence intervals. Results are ordered for the three independent variables 1) number of microphones, 2) interpolation algorithm and 3) audio content. Results are ordered from interpolation algorithm 1 to 5 in the greenish sections. Within each section the number of microphones increases.

Results (raw ratings) are displayed in Fig. 7. For each stimulus values of average perceived plausibility are displayed together with 95% confidence intervals. For both the *INTERPOLATION* and *DISCRETIZATION* (no. of microphones) factor, stimuli have been ordered from left to right in the direction of increasing plausibility expected from our prior hypotheses (cf. sect 3.3). The effect of the *CONTENT* factor is illustrated using colour-coding.

5. Discussion

Firstly, if our main hypotheses were true, ratings should in general increase from left to right. Approximately this seems to be the case. Additionally, when averaging over all three contents, our anchor stimuli (no. 1 and no. 20) were indeed rated similar as the worst or best stimuli, respectively. They thus can be expected to have served as expected.

From non-aggregated data (not shown) we found that the scale's range was fairly well exploited by subjects, though there was a tendency for ratings to cluster at the bottom end. Besides, the average ratings of MTB's plausibility appear limited at the top end, they seldom exceeded the 60 % mark (remember the scale end labels 'excellent' and 'very bad'). As a true reference stimulus – the individual experience of acoustic reality – was not provided in our test, this fact should be though interpreted carefully. Subjects were aware of listening to recordings only; they did not expect to hear a stimulus being identical to their expectation and therefore the upper scale end might have been avoided. Besides, in written comments collected after the listening test several subjects also mentioned a permanent perception of elevation and of unstable source localization. From section 1.2 we know, that these artefacts are due to missing pinnae and a mismatch in sphere diameter resulting in erroneous ITD cues. These artefacts may very well have caused an inherent limitation of MTB's plausibility.

To support our further statements with results from inferential statistics a repeated-measures ANOVA was conducted. Data requirements for ANOVA were verified using Mauchly's test of sphericity. Accordingly, significance ratings were read after correction of degrees of freedom where needed. It was found that all main effects and all first order interactions were highly significant. Defining suitable directive contrasts for the main effects all a-priori hypotheses could be approved at a highly significant level. That means, on average, plausibility rises with number of microphones, with decreasing stimulus bandwidth, and following the perceptually motivated order of interpolation algorithms proposed by Algazi et al. [11] and explained in sect. 1.2. Post hoc pair wise comparisons (with Bonferroni adjustment for multiple comparisons) though put these results into perspective, as 1) the *CONTENT* effect is governed mainly by the noise stimuli's ratings differing from that of the other two stimuli, 2) the *INTERPOLATION* effect is formed by threefold grouping of the five algorithms (algorithm 1 is rated worse than all others, followed by algorithms 2 to 4 evaluated to be of similar performance[1] and algorithm 5 being rated best), and 3) as the 8-microphone condition (*DISCRETIZATION)* was rated worse than all other combinations, while both 24 and 32 microphones were on average rated similarly well.

As all first order interactions (*DISCRETIZATION* x *INTERPOLATION, DISCRETIZATION* x *CONTENT,* and *INTERPOLATION* x *CONTENT*) were significant too, they have to be analyzed for contradictions before accepting the main effects.

We discuss the *DISCRETIZATION* x *INTERPOLATION* interaction first, as it is both demonstrative and most interesting as seen from the scope of our study. It is expressed by the group wise similar ratings of interpolation algorithms identifiable in Fig. 7. With increasing number of microphones plausibility ratings remained nearly constant for algorithms 1, 2, and 5 (TB-FM, FR-NM, TB-SI), whereas for algorithms 3 and 4 (FR-LI, TB-NM) ratings increase with the number of microphones as expected. Additionally, for algorithms 1 and 5 (TB-FM, TB-SI) ratings were nearly independent from audio content and additionally either pretty worse or rather good. These are good news as we instantly can define both a clear overall 'looser' and a 'winner' algorithm. Two Band Spectral-Interpolation Restoration (TB-SI) performed as good as the best remaining combinations of microphones and algorithms. But what is more, it does that equally well even for most critical signals and a minimum number of microphones. In contrast, Two Band Fixed-Microphone Interpolation (TB-FM) was rated inferior under all conditions. This is most probably due to the monaural high frequency components leading instantly to in-head-localization and a 'split' perception of low and high frequency content. For Full Range Nearest Microphone Selection (FR-NM) ratings were also statistically independent from the number of microphones (approved by post-hoc pair wise comparisons), but opposed to TB-FM and TB-SI, for the noise stimulus ratings were much worse. So, for all stimuli the annoyance of FR-NM's full range switching artifacts – whether on average obvious (for noise) or faint (for the natural stimuli) – was not improved by increasing the number of microphones. Although the number of microphones influences the frequency of switching events during head movements, the velocity of head movement and in turn the switching frequency will seldom be constant. It might therefore be independent from the absolute

[1] Whereas algorithm 3 (FR-LI) was on average rated even slightly worse than algorithm 2 (FR-NM).

number of microphones. Secondly, the 'gestalt' of the switching artifacts does not change with frequency of occurrence and thus remains equally disturbing under all microphone conditions[1]. As mentioned, algorithms 3 and 4 (FR-LI, TB-NM) form a second group. Both were rated in a similar manner as perceived plausibility (mostly) increased with the number of microphones. The gain from an increasing number of microphones is explainable for FR-LI as cross fading signals from continuously more closely neighbored spatial sectors pushes comb filter artifacts into higher frequency regions. Perceptual performance of TB-NM (high frequency switching) gains even more from a higher number of microphones, probably as discretization of high frequency spectral cues does not affect localization and switching artifacts might become more subtle with increasingly refined spatial sampling. Moreover, for algorithm 3 and 4 we found trends to second order interaction, which are discussed at the end of this section.

A symptom of the second significant interaction *INTERPOLATION* x *CONTENT* was that, while interpolation algorithms were – as expected – mostly rated increasingly better, only for the critical noise stimulus, all algorithms were rated constantly bad despite for the superior TB-SI interpolation. Moreover interaction is seen within the content related differences of ratings of algorithm 2 (FR-NM, full range switching), when it was rated above average for speech and strings but below average for the noise stimulus, indicating that the annoyance of (full range) switching artefacts was strongly dependent on content. Indeed, with modulated signals as speech and music, switching was sometimes hard to hear at all, whereas for steady state noise it was always clearly perceivable.

In contrast to the *CONTENT* main effect there was a trend to rate the string quartet worst with TB-SI interpolation. This can be considered an artifact of the test design, as consequently, as we established the internal references to be an equivalent natural experience, subjects often criticized the unusual spatial arrangement of the strings as being implausible. Interestingly, this 'misalignment' within the string quartet stimulus showed off only in the case of the best algorithm (TB-LI) where ratings where otherwise independent from discretization and interpolation. This difference in ratings might therefore directly be interpreted to stand for the amount of implausibility of the unusual spatial arrangement. It can be supposed that, if the string quartet was arranged in a natural way, ratings would have been similar to those for speaker and noise.

The third and last significant interaction *DISCRETIZATION* x *CONTENT* showed up as ratings of the speech stimulus were – when averaged over all interpolation algorithms – sensitive above average to the number of available microphones. Speech ratings showed the widest spread, in the case of 8 microphones it was on average rated even worse than the musical stimulus. This pronounced sensitivity could be explained with speech being a highly familiar signal.

Additionally we found trends (not significant) to second order interactions, for instance when – depending on type of interpolation *and* stimulus – plausibility increased either linearly *or* quadratic with the number of microphones. For the noise stimulus plausibility always increased linearly, and the same was found for the speech stimulus with algorithm 3 (FR-LI). Thus, for noise and both algorithms 3 (full range cross fading) and 4 (high frequency switching) the perceptual optimum number of microphones would be equal or even larger than 32 microphones. In contrast, for the speech stimulus and algorithm 4 (TB-NM) the relation seems quadratic with an optimum of 24 microphones. For the string quartet

[1] The reason for the content dependency is addressed below.

the relation appears quadratic for algorithm 3 (optimum at 24 microphones), and cubic for algorithm 4 (local optimum at 16 microphones). Thus, for modulated stimuli the situation appears less clear than for steady state noise. Optima in the number of microphones potentially depend in a complicated way on the interplay of type of modulation in the signal, the algorithms artifacts (fading or switching, frequency content) and the average velocity of head movements. Anyway, as the best algorithm TB-SI were rated in a much simpler manner, these (insignificant) dependencies appear to complicate as that they could be exploited in a profitable way in practical applications. Yet maybe, as TB-NM shows a distinct perceptual optimum for 24 microphones and as it is computationally simpler than TB-SI, it might serve as high-quality low-effort alternative for speech applications. Another trend to second order interaction was found as, for noise, high frequency switching combined with low frequency cross fading (TB-NM) gained from increasing the number of microphones unlike as for full range switching (FR-NM). This indicates that for steady state sounds with FR-NM the perception of plausibility was dominated by hard switching of the low frequency range. Cross fading in TB-NM probably suppressed a 'sampled' impression of the environment which remained with FR-NM independent of how fast sectors were switched. Another higher order trend was found when – only for the two natural stimuli and for 8-microphone setting – algorithms 3 and 4 were rated worse than algorithm 2 (FR-NM). For speech being a strongly amplitude modulated signal intermixed with random and distinct signal pauses, artifacts of full range switching appear less perceivable than coloration-like comb filter artifacts introduced by linear interpolation (see conclusions).

6. MTB Applications

In the beginning of this paper we motivated our studies by considering MTB an interesting means for recording ambient scenes. Additionally, we sometimes encounter listeners criticizing an 'aseptic feel' of binaurally synthesized music performances which is ascribed to the missing impression of a natural social and technical acoustic environment or 'soundscape' (audience noise, background noise of technical origin). As said before, since time-variant effects are on principle eliminated by using impulse responses, and since the synthesis of a surrounding audience seems unreasonably expensive to prepare using impulse response based scene representations, we found MTB to be a valuable system for live recording of ambient scenes in a dynamic pseudo-binaural format. Therefore, and in accordance with the findings of our study a real MTB array was built. Instead of discussing construction in details here too much, we merely want to present our practical solution and introduce some of planned and accomplished applications.

As it is a crucial parameter special effort was put on the determination of the sphere's diameter. We used the regression formula for an optimum head radius from [29], feeding it with anthropometric data from the German standard DIN33402-2 E [30] averaged over both sexes and all age groups. This way we specified the diameter to be 176 millimetres. The construction was done entirely in CAD with the help of a professional mechanical engineer (cf. Fig. 8, left & middle). The shell of the MTB array was fabricated from a thermoplastic material in a rapid prototyping process (SLS, selective laser sintering, cf. Fig. 8, right).

Flush with the surface 24 miniature microphones (Panasonic WM-61A, Ø = 6 mm) are mounted. Audio signals are transmitted using a professional 3-wire 24 channel multicore cable. Via voltage dividers, soldered into the XLR connectors at the breakout end of the multicore, standard phantom power sources can be employed as power supply. For that purpose we use three 8-channel microphone preamps with analogue (XLR) and digital

(ADAT optical) interfaces. Using a 24 channel optical interface with a PCMCIA card the MTB array can be connected to a standard laptop for recording. From measurement results and from informal listening tests we decided to abandon additional frequency response compensation of the microphones. Though, as the WM-61A exhibits a noticeable spread in sensitivity gain correction factors are applied individually for each channel during playback.

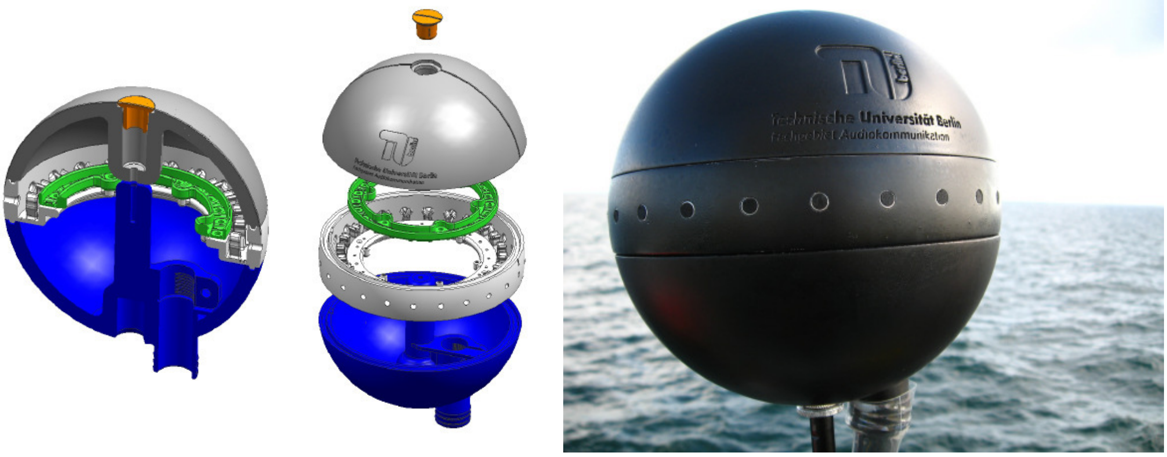


Fig. 8: Assembled (left) and exploded (middle) view of CAD prototype at final design stage. Right: TU's 24-channel MTB array at high-sea operation.

The horizon for potential applications of MTB is wide. A first classification into the three categories 1) remote listening, 2) recording, and 3) immersive interactive multimedia applications was given already in [11]. Our main focus lies on musical and artistic applications though. Here, promising applications are for instance:

- immersive documentation and archiving of sound art installations,
- recording of natural soundscapes,
- transmission of interactive spatial concert sound, and
- interactive multidirectional transmission in distributed concerts.

In a first practical application our MTB array was used to record all of the performances presented at the 2010 sound art festival "Inventionen" in Berlin. Furthermore, in a current research and development project the applicability of MTB for acoustic high-sea surveillance is assessed. As mentioned before, we see a major application in the combination of data based binaural synthesis with naturalistic MTB ambience recordings. Therefore, our binaural audiovisual concert hall simulator is currently extended to incorporate such environmental cues.

7. Summary and conclusions

In 2004, Algazi et al. [11] introduced motion-tracked binaural (MTB) sound as a new method for capturing, recording, and reproducing spatial sound. Their study was accompanied with thorough quantitative analyses of potential approaches to interpolation and discretization. Melick et al. [17] discussed perceptual shortcomings of MTB and presented possible remedies. Nevertheless, both studies lacked a formal perceptual assessment of MTB sound quality. Starting from here, we introduced plausibility as a suitable criterion for assessing interpolation and discretization of MTB sound. We described a method to synthesize MTB stimuli systematically varying in 1) audio content, 2) interpolation algorithm, and 3) number of microphones and presented results from a listening test conducted with 26 subjects. In a thorough analysis of results we showed the plausibility of MTB reproduction to be highly interdependent on all three parameters.

Finally, resuming the hypothesized main effects, it can be stated that differences in audio CONTENT ratings seemed to a lesser degree be due to differences in bandwidths than due to amount of modulation in the stimuli. Moreover, the speech stimulus formed a special case as its ratings were most sensitive for the type of interpolation algorithm used. INTERPOLATION algorithms' plausibility indeed increased as expected, though, the highly content-dependent ratings of algorithm 2 (full range switching) form a special case. For modulated natural signals only it was even found to be second best. The benefit of the amount of DISCRETIZATION has – though plausibility on average rose with the number of microphones – also to be judged regarding the specific interpolation algorithm. Probably the most surprising result was that with a superior signal processing (TB-SI) the relevance of DISCRETIZATION became nearly negligible.

Satisfactory, a perceptually clearly superior configuration could be found. If there do exist no constraints regarding processing power, the Two Band Spectral-Interpolation Restoration (TB-SI) should always be preferred as interpolation algorithm, especially as its performance is nearly independent from number of microphones (from 8 up to 32) and type of content. If audio quality is of prime importance, and audio signals are most critical there is no equivalent alternative solution. If processing power is limited though, but quality of critical signals is still important, algorithms 3 and 4 (Full Range Linear Interpolation, Two Band Nearest Microphone Selection) perform nearly equally well but signals should be recorded with the highest possible number of microphones (from our results: 32 at least). Algorithm 1 and 2 (Two Band Fixed-Microphone Interpolation, Full Range Nearest Microphone Selection) should never be used if high quality transmission of critical audio content is aimed at. However, if the application is limited to speech transmission, algorithms 2, 3 and 4 (Full Range Nearest Microphone Selection, Full Range Linear Interpolation, Two Band Nearest Microphone Selection) can be used though the latter both should not be applied using less than 16 microphones. Instead, if bandwidth and processing power are limited, for speech transmission algorithm 2 (Full Range Nearest Microphone Selection) applied with 8 microphones appears most recommendable.

Finally we discussed the convenient design of a 24 channel MTB prototype and presented some of its applications. For the future it would be interesting to assess MTB recordings' authenticity following the proposed listening test scheme. As Two Band Spectral-Interpolation Restoration was found to be the superior interpolation algorithm such an assessment could be limited to this algorithm only.

8. References

[1] Lindau, Alexander (2010): "Zu den Dimensionen des Unterschied live aufgeführter und reproduzierter Musik: Ergebnisse einer qualitativ/quantitativen Umfragestudie." In: *Fortschritte der Akustik: Tagungsband d. 36. DAGA*. Berlin, pp. 609-610

[2] Spors, Sascha; Ahrens, Jens (2008): "A Comparison of Wave Field Synthesis and Higher-Order Ambisonics with Respect to Physical Properties and Spatial Sampling." In: *Proc. of the 125th AES Convention*. San Francisco, preprint no. 7556

[3] Wagner, Andreas et al. (2004): "Generation of highly immersive atmospheres for Wave Field Synthesis reproduction." In: *Proc. of the 116th AES Convention*. Berlin, preprint no. 6118

[4] Karamustafaoglu, Attila et. al. (1999): "Design and applications of a data-based auralization system for surround sound." In: *Proc. of the 106th AES Convention*. München, preprint no. 4976

[5] Lindau, Alexander; Hohn, Torben; Weinzierl, Stefan (2007): "Binaural resynthesis for comparative studies of acoustical environments." In: *Proc. of the 122nd AES Convention*. Vienna, preprint no. 7032

[6] Mackensen, Philip (2004): *Auditive Localization. Head Movements, an additional cue in Localization*. Doct. diss. Technische Universität Berlin

[7] Gardner, William G. (1997): *3-D Audio Using Loudspeakers*. MIT Media Laboratories. Cambridge

[8] Schärer, Zora; Lindau, Alexander (2009): "Evaluation of Equalization Methods for Binaural Signals." In: *Proc. of the 126th AES Convention.* Munich, preprint no. 7721

[9] Lindau, Alexander; Brinkmann, Fabian (2010): "Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings." In: *Proc. of the 3rd Third International Workshop on Perceptual Quality of Systems*. Dresden, pp. 137-142

[10] Christensen, Flemming et al. (2005): "A Listening Test System for Automotive Audio - Part 1: System Description." In: *Proc. of the 118th AES Convention*. Barcelona, preprint no. 6358

[11] Algazi, V. Ralph; Duda, Richard O.; Thompson, Dennis M. (2004): "Motion-Tracked Binaural Sound." In: *J. Audio Eng. Soc.*, Vol. 52, No. 11, pp. 1142-1156

[12] Theile, Günther (1986): "Das Kugelflächenmikrofon." In: Bildungswerk des VDT (Hrsg.): *Bericht der 14. Tonmeistertagung*. München, pp. 277-293

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

[14] Strutt, J. W. (1907): "On Our Perception of Sound Direction." In: *Philosophical Magazine*, Vol. 13, pp. 214–232

[15] Wightman, Fred; Kistler, Doris J. (1992): "The dominant role of low-frequency interaural time differences in sound localization." In: *J. Acoust. Soc. Am.*, Vol. 91, No. 3, pp. 1648-1661

[16] Griffin, Daniel W.; Lim, Jae S. (1984): "Signal Estimation from Modified Short-Time Fourier Transform." In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No. 2, pp. 236-243

[17] Melick, Joshua 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

[18] Lindau, Alexander; Estrella, Jorgos; Weinzierl, Stefan (2010): "Individualization of dynamic binaural synthesis by real time manipulation of the ITD." In: *Proc. of the 128th AES Convention*. London, preprint no. 8088

[19] Wright, Matthew; Freed, Adrian; Momeni, Ali (2003): "Open Sound Control: State of the art 2003." In: *Proc. of the 2003 Conference on New Interfaces for Musical Expression (NIME-03).* Montreal

[20] Lindau, Alexander; Weinzierl, Stefan (2006): "FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom." In: Bildungswerk des VDT (Hrsg.): *Bericht der 24. Tonmeistertagung*. Leipzig

[21] The Mathworks Inc. (2009): *Matlab Vs. 7.8.0 (R2009a). The language of technical computing.* Natick, MA, USA

[22] Blauert, Jens (1997): *Spatial Hearing. The Psychophysics of Human Sound Localization*. 2. Aufl., Cambridge, MA.: MIT Press

[23] Smyth, Stephen M. (2006): *Personalized headphone Virtualization* (US Patent Application Publication). US 2006/0045294 A1

[24] ITU (2003): *ITU-R Rec. BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems*, Geneva: International Telecommunication Union

[25] Faul, Franz et al. (2007): "G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences." In: *Behavior Research Methods*, Vol. 39, No. 2, pp. 175-191

[26] Bortz, Jürgen; Döring, Nicola (2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 4. Aufl., Heidelberg: Springer

[27] Shrout, Patrick E.; Fleiss, Joseph L. (1979): "Intraclass Correlations: Uses in Assessing Rater Reliability." In: *Psychological Bulletin*, Vol. 86, No. 2, pp. 420-428

[28] Bech, Søren; Zacharov, Nick (2006): *Perceptual Audio Evaluation: Theory, Method and Application.* Chichester: Wiley

[29] Algazi, V. Ralph; Avendano C.; Duda, Richard O. (2001): "Estimation of a Spherical-Head Model from Anthropometry." In: *J. Audio Eng. Soc.*, Vol. 49, No. 6, pp. 472-479

[30] DIN (2005): *DIN 33402-2 E: Ergonomie - Körpermaße des Menschen - Teil 2: Werte. Entwurf*, Berlin: Beuth