Audio Engineering Society

Convention Express Paper 156

Presented at the 155th Convention

2023 October 25-27, New York, USA

Evaluation of binaural rendering quality for professional audio.

Simone Fontana[1], Paolo Martignon[1], Mario di Cola[1], Alessandro Arturi[1], Alberto Bianco[1], and Angelo Farina[2]

[1] *Contralto Audio, Via Frentana 17, 66043 Casoli (CH) ITALY*

[2] *Università degli studi di Parma, Via delle Scienze 181/A Parma, 43100 ITALY*

Correspondence should be addressed to Simone Fontana (s.fontana@contralto-audio.com)

ABSTRACT

Electroacoustic simulation tools can be enhanced by realistic binaural audio rendering of the modelled audio system. Even if binaural technology is known since decades, it is still struggling to have a strong impact as working tool for acoustic consultants or audio system engineers, mainly because it is difficult to assess the perceptual equivalence between a real sound scene and its binaural rendering by headphones. In the last years, research efforts have been devoted to investigating binaural authenticity. In this brief we address binaural authenticity for a practical use case, in the context of binaural rendering of modelled electroacoustic systems. We ran a perceptual test in which audio system engineers compared stereo loudspeaker playback of a given audio content with binaural rendering via headphones of the same playback, to verify whether the electroacoustic modelling and binaural synthesis processes could have introduced critical artefacts to the original loudspeaker-based sound scene perception.

1 Introduction

Professional sound system engineers are called to design sound reinforcement systems that deliver an optimal sound experience to the audience. When the event is intended to be attended by thousands of people, the audio systems can become cumbersome and preliminary design has to be carried out using dedicated predictive software that informs in advance about how the system will sound.

Currently, software tools exist that allow to model the acoustic scene and visualize sound maps indicating the space-frequency distribution of SPL in a modelized venue, given a modelized sound reinforcement system. In some cases, *binauralization*[1] of the system in the venue is also featured.

In the context of electroacoustic simulation tools, binauralization quality depends on the model of the acoustic space (especially for indoor modelling), on the model of the electroacoustic system and on the design and implementation of the binaural rendering engine.

The main goal of this paper is to compare (by the means of listening tests) a real sound scene, consisting in a simple stereo playback via two loudspeakers placed in an acoustically dry lab, with a headphones binaural rendering of the same setup. Binauralization was based on a model of the used loudspeakers and binaural synthesis of the acoustic

[1] (Bin)auralization is defined ([1]) as the process of rendering audible the sound field of a source in a space, by physical or mathematical modelling, in such a way as to simulate the binaural listening experience at a given position in the modelled space.

path between the loudspeakers and the listener eardrums.

1.1 Research goal

The idea behind binaural audio is that, as hearing is based on two signals (the sound pressures at each of the eardrums), if these are recorded at the eardrums of a listener and reproduced at the same eardrums exactly as they were, then the complete auditive experience is assumed to be reproduced, including timbre and spatial aspects ([2]).

Binaural technology has been developed and studied for many decades and it has found applications in many fields. However, a recent survey ([3]) shows how it is struggling to have an impact as a working tool for acoustic consultants or audio system engineers.

Binaural technology could reach more critical areas of applications only if it was recognized as *authentic*, meaning providing perceptual identity with an explicitly presented real sound scene ([4]). In recent years, several researchers (see [4] for a survey, [5] and [6]) have tried to understand whether binaural rendering can be considered authentic, but results show that perceptual identity (i.e. authenticity) cannot be easily generalized.

In our study we rather focused on *plausibility*, similarly as defined in [4]. The plausibility of binauralization refers to its agreement with the listener's expectation toward a corresponding real event (agreement to an internal reference). In particular, we focused on plausibility of binaural rendering for system audio engineers. System audio engineers, on one hand, have high listening skills and high-quality expectations, on the other hand, they have the experience to identify what binauralization should preserve of a real playback for guaranteeing future installation success.

The research question could then be: *can binaural synthesis be plausible enough for a system audio engineer to be used as a working tool*[2]?

1.2 Research method

To provide a response to the research question, first we asked system audio engineers to help us define some key requirements that binauralization should fulfill in order to be plausible and useful for their audio system optimization duties.

Then, we ran a perceptual test in which audio system engineers could compare music playback delivered by professional audio equipment and binaural synthesis of the previous playback, rendered on headphones. The test has been designed to evaluate metrics associated to the previously defined key requirements.

Binauralization was based on both electroacoustic modelling and HRTF-based synthesis. Concerning audio system equipment modeling, a self-developed acoustic simulation platform has been used. This software tools allows (among other features) for simulating the impulse response of a complex audio system in a given point (in free field), based on an acoustic model obtained from balloon and electric measurements. In the paper we will not describe this process in detail, and focus on the HRTF synthesis processing chain instead. The results of the test allowed to implicitly check if both the electroacoustic modelling and binaural synthesis processes could have introduced critical artefacts to the original loudspeaker-based sound scene perception.

This paper describes the audio setup (section 2) and audio processing chain (section 3), highlighting issues related to audio transmission via Bluetooth channel and to loudness mismatch between loudspeaker and headphones. In section 4 we present the listening tests, followed by a discussion of the results in section 5.

2 Audio system

The lab in which the experiment took place has a volume of 1700 m^3 appx (5,4 m x 12,5 m x 25 m); it is fully acoustically treated, and has an average reverberation time (RT30) of 290 ms. We considered the lab as acoustically dry and thus avoided, in first approximation, to use room acoustic modelling in the context of this study.

Two self-produced two-way loudspeakers (1 x 8" woofer, 1 x 1" tweeter with a H60° / V40° coverage horn) have been installed in front of the listener in a narrow stereo layout (see figure 4). The distance

[2] For the sake of precision, we have to note that in this study there was indeed an external reference (the loudspeaker stereo setup), but we asked the system engineers to rate binauralization according to their experience (internal reference), judging if, even if not authentic, binaural rendering can be still be useful for preliminary system design.

between the speakers (positioned at 1.9m height) is 3.8m and between speakers and listener (ears at 1.7 m from the floor) is 6.1 m. The elevation of the system with respect to the listener is 2° appx, while the azimuth of one speaker is 20° appx (narrower than standard stereo).

Neumann manikin KU100 was used for all binaural responses acquisition. The soundcard was a Roland Octamic connected to a Mac mini M2 running Monterey. Sampling frequency was 48 kHz. Impulse responses were measured using 5s long exponential sweeps as implemented in MAX/MSP/SPAT suite.

As monitoring system for binaural rendering, we used Apple Max wireless headphones. This choice has been made after considering relevant issues related to Bluetooth channel, low frequency headphones response, limited response variations for headphones repositioning, and transparency mode feature. We will shortly address these aspects in the next subsections.

- Bluetooth channel characterization

The synchronization of the soundcard used for recording and the Apple Max Bluetooth headphones used for playback is critical, because clock mismatch would result in audio artefacts, such as glitching. In order to avoid synchronization issues, we used Apple aggregate device tool. We created an aggregate device using Apple Max as master clock (and Roland Octa Capture in slave mode) and flagging the drift correction boxes (on Monterey OS).

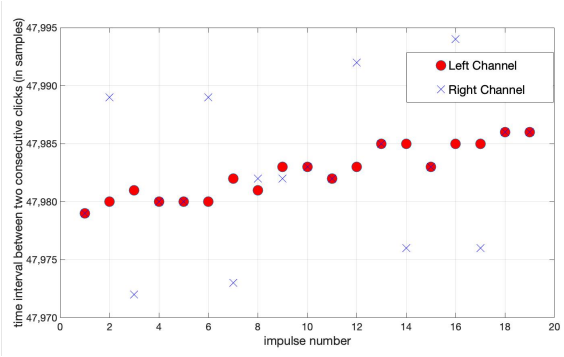

Figure 1: Bluetooth channel characterization

In order to characterize the Bluetooth channel, we reproduced via headphones a stereo 20s long pulse train with 1s period, mixed with a low-level white noise (used in order to keep the Bluetooth session open between impulses). We recorded the 2 signals at the Neumann manikin ears and measured for both channels the delay between consecutive pulses, (which is expected to be 1s, in sync for left and right channels).

In figure 1 we show the results of the test. The global latency of the system appears to be slowly drifting during the observation interval: for example, the delay between two impulses grows of about 10 samples (0.2ms @48kHz) for the left channel. The synchronization between channels is inside a 20 samples interval (0.4ms @48kHz) and may also include different left/right headphones cup positioning. We also note the delay between impulses is always less than 1s.

Even if the time of arrival of the impulse may be prone to estimation errors up to some samples, due to the recorded pulse smearing (caused for example by D/A/D conversion) we can conclude the Bluetooth channel is time varying, which is common for wireless channels due to time variability of the physical channel and the transmission protocol ([16]).

In the context of time varying channels, we should use time varying impulse responses as described in [17]. However, we consider the impulse response of the system to be time invariant over the measurement interval of the sweep (5s), which results in a drift < 0.05ms, and accepted to consider standard time invariant IR as representative of the system response.

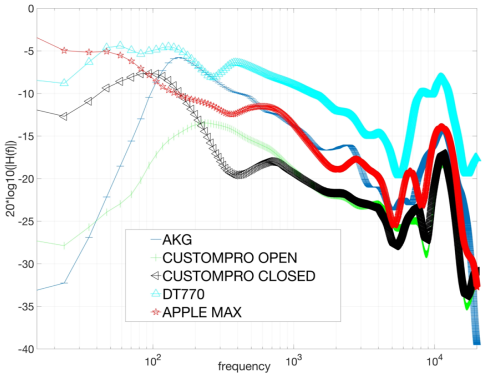

Figure 2: Headphones transfer function comparison

The lack of perfect synchronization of the left and right channels may result in critical spatial distortion. Based on trigonometry, a difference of propagation path of ±10 samples corresponds to a maximum rotation error of the order of 30°.

- Low frequency response

Apple Max frequency response presents more energy in the low end, and extends to lower frequencies (below 50 Hz), compared to other headphones taken into consideration (see figure 2, responses not normalized), without conveying excessive boosting. Apple Max can then properly reproduce the information delivered by the used loudspeakers (which has a -6dB LF cutoff at around 40 Hz).

- Response variations for repositioning.

Apple Max resulted in better response stability over repositioning, especially at low frequencies (see figure 3).

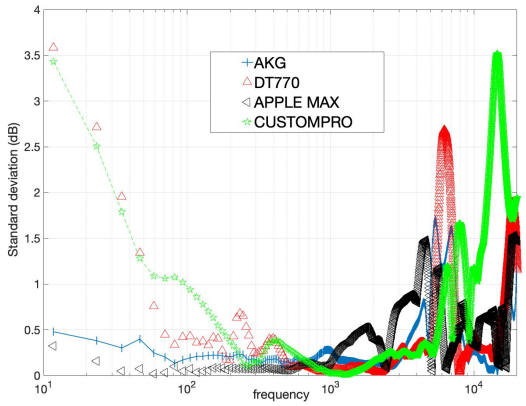

Figure 3: Robustness to positioning. Standard deviation of frequency response over frequency, for 5 repositioning.

- Transparency mode feature.

Apple Max features a transparency operation mode, that allows the user to hear the external sound environment electro-acoustically through the device. In order to run comparative tests without the bias of knowing to be wearing headphones (as done for example in [4], in which extra-aural headphones were worn also when listening to loudspeakers), we considered the option to use the transparency mode to listen to transducers without removing headphones.

As preliminarily performed tests and literature [7] show, external sounds suffer from artefacts due to coloration and comb filtering, and may also present binaural artefacts due to different delays of the Bluetooth stereo channels ([7]).

As the scope of the test is to compare real source perception and binaurally rendered signals, we could not accept the degradation of loudspeakers signals in hear-through mode and decided to keep a transparent listening of the original system (not wearing headphones) as an unprocessed reference.

3 Audio processing

As presented in figure 4, the variables are defined as follows: electrical source signals s_1 and s_2 are fed to speaker 1 and 2. The acoustic source signal output from the speakers are x_1 and x_2. Binaural Room Impulse Response (BRIR) are defined as b_{1L}, b_{1R}, b_{2L}, and b_{2R}; BRIRs contain the contribution of the loudspeaker, the medium and the manikin microphones; the recorded signals at the manikin microphones are y_L and y_R.

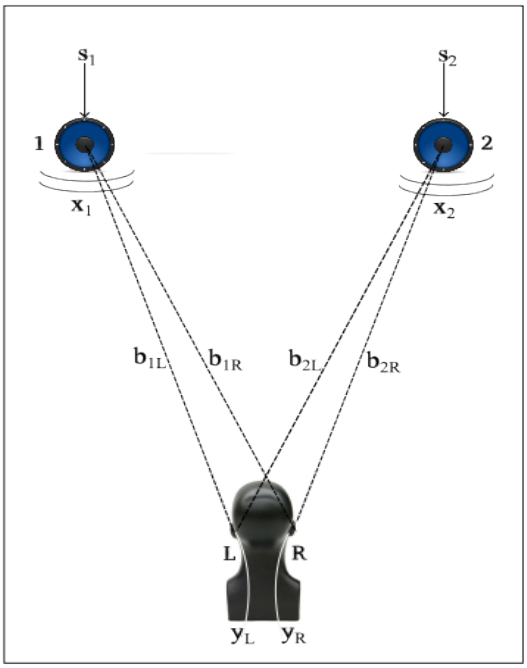

Figure 4: installation layout

Headphones (not present in the figure) responses to the ear are called hp_{1L} and hp_{2R}, where the first numerical index correspond to the first headphones channel (left) and the second alphabetical index corresponds to the ear (L for left). Note that the response of the manikin microphone is also included in the headphones response.

We note y'_L and y'_R the signal received by the microphones in case of headphones reproduction. The binaural principle is fulfilled if $y_L = y'_L$ and $y_R = y'_R$.

Assuming linear superposition of the loudspeaker signals, the signals received at the manikin

microphones could be also obtained from recorded BRIRs using the model presented in Figure 5, that is:

$$y_L(t) = s_1(t) * b_{1L}(t) + s_2(t) * b_{2L}(t) \quad \textbf{(Eq.1)}$$
$$y_R(t) = s_1(t) * b_{1R}(t) + s_2(t) * b_{2R}(t)$$

If we feed the headphones with signals y_L and y_R, the signals received by the ears with headphones reproduction would then be:

$$y'_L(t) = (s_1(t) * b_{1L}(t) + s_2(t) * b_{2L}(t)) * hp_{1L}(t) * eq_L(t) \quad \textbf{(Eq.2)}$$

$$y'_R(t) = (s_1(t) * b_{1R}(t) + s_2(t) * b_{2R}(t)) * hp_{2R}(t) * eq_R(t)$$

where $eq_L(t)$ and $eq_R(t)$ are the equalizers for left and right channel, used in order to remove headphones coloration and fulfil the binaural principle. For example, from Eq.1 and Eq.2 it results that the binaural principle is fulfilled if at the left ear we have: $hp_{1L}(t) * eq_L(t) = \delta(t)$.

In a predictive scenario, no binaural recordings or BRIRs are available: binaural synthesis can be used instead. The loudspeakers output signals x_1 and x_2 can be obtained using the synthetic free field response of the loudspeaker, (we call it $h(t)$) computed by our self- developed loudspeaker system modeling software platform. Binaural synthesis can then be carried out using HRTFs taken, for example, from BILI database ([8]). A common HRTF set was chosen in preliminary tests. The HRTF were convolved with the simulated impulse response of the loudspeaker. No room effect has been added, due to the previous consideration on the acoustic of the laboratory.

Let us call $y''_L(t)$ and $y''_R(t)$ the signals reproduced by headphones at the ears obtained with binaural synthesis.

$$y''_L(t) = (s_1(t) * h'(t) * hrtf_{1L}(t) + s_2(t) * h(t) * hrtf_{2L}(t)) * hp_{1L}(t) * eq_L(t) \quad \textbf{(Eq.3)}$$

$$y''_R(t) = (s_1(t) * h'(t) * hrtf_{1R}(t) + s_2(t) * h(t) * hrtf_{2R}(t)) * hp_{2R}(t) * eq_R(t)$$

Headphones equalization

Left and right channel equalizers were obtained as 4096 taps FIR filters, using a method similar to the one proposed in [9], based on frequency inversion of the smoothed average Headphones Transfer Function (HPTF) with frequency-dependent regularization. Apple Max HPTFs have been measured on KU100 with 5 repositioning. The average HPTF has been smoothed with a ¼ octave window. No phase average has been carried out. The target function was a delta function, filtered between 20 Hz and 20 kHz, in order to avoid excessive energy of the equalizer close to the hearing band borders, which has little perceptual effects, but would determine excessive equalizer length or pre-ringing. Regularization was frequency dependent: the frequency dependent regularization parameter was shaped in order to avoid residual notches (i.e. destructive interference inside the cavum conchae around 8 kHz, [10]) that would also result in excessive ringing in the equalizer.

The results on EQ performances are shown in figure 5, where we plot (as ideal case) the equalized HPTF in the optimal position, that is when the headphones are not moved between HPTF measurement and equalized HPTF measurement (and the equalizer is built on that single measured HPTF); we also plot (as realistic use case) the equalized HPTFs for 2 different positions, in case the EQ based on average HPTF is used.

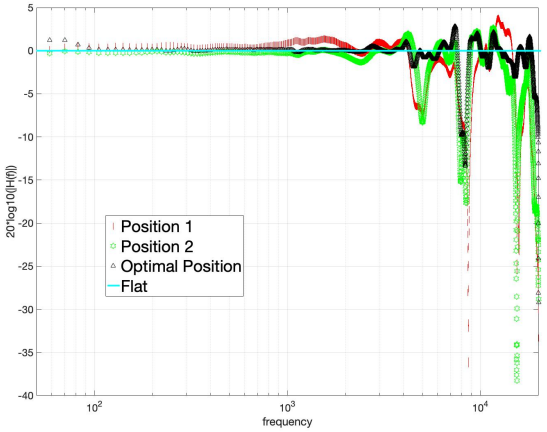

Figure 5: EQ performances

In the case of optimal equalizer, we can observe a flat response up to 4kHz; above this frequency the smoothing process used in the equalizer design results in a more moderate correction (the response is within 3 dB range from flat). The concha resonance is not corrected, as expected.

In the realistic case the equalized response is more deteriorated in high frequency, where some secondary dip notches are not corrected due to shift in frequency domain (because of repositioning); however, these uncorrected deep notches have little contribute to timbral perception in wideband music listening. Some narrowband high frequency boosting (< 4dB) may result instead in perceptual artifacts

determining ringing, depending on the reproduced content.

We note no wideband HF emphasis can be observed in the equalized response. This is in contrast to preliminary tests in which listeners reported too bright sound compared to real playback, in a systematic way across different musical audio tracks. Other authors report similar findings ([9]). However, due to the previous considerations on EQ performances, this effect cannot be related to EQ quality.

The emphasis cannot be due to the diffuse field microphones used in KU100 either (see [10]), because the response of the microphone from frontal direction of arrival (with respect to the capsule, so on the side of the head) is in any case corrected by the headphones equalizer.

HPTFs and HRTFs have not been individualized, so that spectral distortion may happen between listener and responses used for headphones equalization and binaural synthesis. However, such differences may determine variable distortion, but not a clear brightness emphasis trend. See for example various HRTF reported in [11].

Loudness mismatch correction EQ

The most reasonable explanation can then be related to loudness mismatch between headphones and loudspeaker reproduction, due the *disparity between headphones and loudspeaker presentation* ([12], [13]). Both studies report that the hearing system appears to be less sensitive to low frequencies in headphones listening compared to loudspeaker listening. Namely, the sensitivity to headphones playback decreases gradually towards low frequencies.

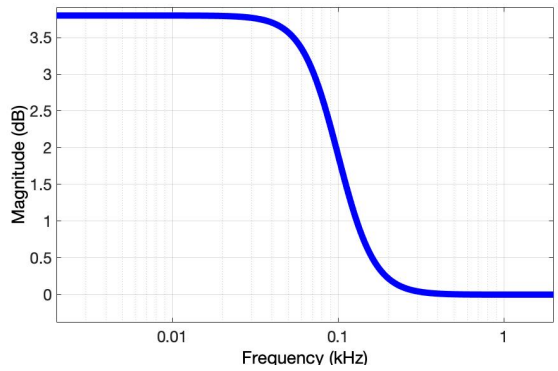

Figure 6: loudness mismatch correction EQ

The sensitivity curve is shaped as a low frequency shelving filter with center frequency in the mid-range, and gain of some dB (depending on conditions) in the low-range.

During the tuning of the system, we decided to introduce a *loudness mismatch correction EQ*. Preliminary listening tests resulted in an equalizer whose shape is depicted in figure 6: it corresponds to a low shelving filter with central frequency = 1 kHz, a gain of 3,8 dB and a Q factor of 1. This correction is in agreement with the findings in ([12], [13]).

4 Listening test design

Twelve experienced sound system and electroacoustic engineers have been invited to participate to the experiment. In a first step, they have been asked to indicate what characteristics of binaural rendering would be essential for supporting virtual system design. In table 1 we report the six characteristics engineers suggested, with associated rating 4-points scale:

	Ratings
Timbral fidelity (HF, MF, LF)	1. not good 2. sufficient 3. good 4. perfect
Presence of artefacts	1. presence of strong artifacts 2. presence of medium artifacts 3. presence of soft artifacts 4. no appreciable artifact
Stereo Image Fidelity	1. not good 2. sufficient 3. good 4. perfect
Externalization	1. not good 2. sufficient 3. good 4. perfect

Table 1: the six parameters for evaluating binauralization quality

A playlist of reference songs was prepared for the listening tests, choosing among the songs commonly used by the invited engineers for system tuning. The same song was reproduced by the loudspeakers and, after binauralization, via headphones. Listeners could choose to listen to one or more tracks and switch between loudspeaker and binaural reproduction simply removing or wearing the headphones. An

operator was switching the content, using a dedicated MAX/MSP patch. No time limitations were applied.

The patch main features were:

- the convolution engines (for headphones EQ and loudness mismatch correction filters)
- the delay compensation between loudspeaker playback and wireless headphones playback
- the relative gain between the two playback modes, which was adjusted perceptually in a preliminary stage.

5 Listening test results

The HF frequencies emphasis/lack of low frequencies that listeners reported in preliminary tests was not mentioned after application of the loudness mismatch correction EQ.

Some listeners that repeated the tests in different days reported slightly different perception. Whether this depended on personal disposition variation or on different Bluetooth condition could not be fully explained. Also, listeners reported the perception was slightly different from track to track. The ratings were then decided and recorded at the last listening, and represented a personal average impression after exhaustive listening of one of several tracks on one or several days.

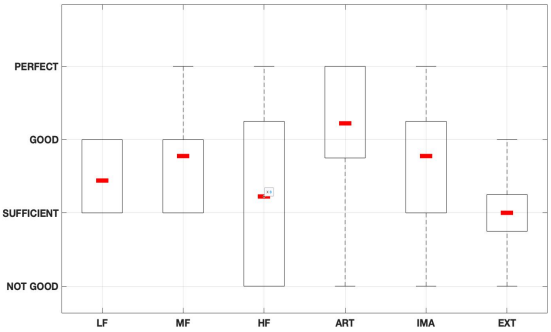

Figure 1: Listening test results

Test results are shown in figure 8. On each box, the central mark indicates the average ratings, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers. Given the small amount of data, exhaustive statistical analysis is not meaningful, so that a more qualitative analysis is made.

We can observe that

- all parameters are considered between sufficient and good on average, and no annoying artefacts were highlighted;
- high frequency timbral fidelity was judged quite differently from listener to listener. This seems to be coherent with the lack of individualization of HPTF and HRTF, which has strongest impact on high frequencies and is highly listener-dependent;
- stereo image fidelity was rated almost good on average, despite the (variable) lack of perfect synchronization between left/right channel mentioned in section 2;
- externalization scored the lowest. Externalization is a very complex phenomenon and is currently related to ([14]): acoustic cues in the direct sound (such as HRTF individualization), reverberation-related cues and multimodal factors such as head movements and vision. The lack of dynamic and individualized binauralization and room acoustic modelling (even though the room is quite dry itself) may explain the poor results.

  It is worth to notice that externalization was not considered by the experts as a critical feature of binauralization in the context of audio system tuning.

6 Conclusions and future work

We ran a listening test comparing loudspeaker and headphones playback in order to investigate whether or not *binaural synthesis is plausible enough for a system audio engineer to be used as a working tool*.

In a preliminary step, we decided to use Max Apple Bluetooth headphones mainly because of the extended frequency response and robustness to repositioning at low frequencies. We highlighted some critical issues of Bluetooth audio transmission, namely its time variability and the lack of perfect synchronization between channels. We also discussed presentation mode mismatch between headphones and loudspeaker listening and introduced a loudness mismatch equalization filter to avoid high frequency emphasis.

Then, twelve experienced audio system engineers and electroacoustic engineers defined relevant characteristics of binauralization and participated to the test, rating those characteristics in an AB test between loudspeaker and binaurally synthesized playback over headphones.

Test ratings show that binauralization performances are good enough in order to be used as working tool in virtual system design. Lowest scores for HF timbral fidelity and externalization could be enhanced by introducing individualization and/or a more conservative headphones equalization, dynamic binauralization and room modelling. These aspects will be investigated in future work.

References

[1] M. Kleiner, B. Dalenbäck, P. Svensson, "Auralization- an Overview", *JAES Volume 41 Issue 11 pp. 861-875; November 1993*

[2] H. Møller, "Fundamentals of binaural technology", *Appl. Acoust. 1992, 36, 171–218*

[3] D. Thery, V. Boccara, and B. F. G. Katz, "Auralization uses in acoustical design: A survey study of acoustical consultants", *The Journal of the Acoustical Society \of America 145, 3446 (2019)*

[4] F. Brinkmann; A. Lindau; S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis", *J Acoust Soc Am 142, 1784–1795 (2017)*

[5] M. Blau, A. Budnik, M. Fallahi, H. Steffens, S. D. Ewert, and S. van de Par, "Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario", *Acta Acustica 2021, 5, 8*

[6] F. Stärz, L. Kroczek, S. Roßkopf, A. Mühlberger, S. Van de Par, M. Blau, "Comparing Room acoustical ratings in an interactive virtual environment to those in the real room", *Forum Acusticum 2023, Turin, Italy*

[7] F. Denk, H. Schepker, S. Doclo, and B. Kollmeier, "Acoustic Transparency in Hearables - Technical Evaluation", *J. Audio Eng. Soc., vol. 68, no. 7/8, pp. 508–521, (2020 July/August.).*

[8] Carpentier T, Bahu H, Noisternig M, Warusfel O (2014), "Measurement of a head-related transfer function database with high spatial resolution." *Forum acousticum, pp 1–6. European Acoustics Association, Krakow.*

[9] Z. Schärer et and A.Lindau, "Evaluation of Equalization Methods for Binaural Signals", *AES 3126th Convention 2009 May 7–10 Munich, Germany*

[10] Brüel&Kjær, "Microphone Handbook", *https://www.bksv.com/media/doc/be1447.pdf (accessed 20 sett 2023)*

[11] V.R. Algazi; R.O. Duda; D.M. Thompson; C. Avendano, "The CIPIC HRTF database", *in Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WAS-PAA), 2001.*

[12] F. Denk, M. Kohnen, J. Llorca-Bofí, M. Vorländer, B. Kollmeier, "The "Missing 6 dB" Revisited: Influence of Room Acoustics and Binaural Parameters on the Loudness Mismatch Between Headphones and Loudspeakers", *Frontiers in Psychology, March 2021, Volume 12, Article 623670*

[13] Kohnen, M., Denk, F., Llorca-Bofí, J., Kollmeier, B., and Vorländer, M. "Cross-site investigation on head-related and headphone transfer function measurements: Implications on loudness balancing", *Acta Acustica 2021, 5, 58*

[14] V. Best, R. Baumgartner, M. Lavandier, P. Majdak and N. Kopco "Sound Externalization: A Review of Recent Research", *Trends in Hearing Volume 24, 2020*

[15] Werner, S.; Klein, F.; Mayenfels, T.; Brandenburg, K. "A summary on acoustic room divergence and its effect on externalization of auditory events", *In Proceedings of the 8th International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016*

[16] F. Hlawatsch, G Matz, "Wireless communications over rapidly time-varying channels", *San Francisco, CA, USA: Academic, 2011.*

[17] J. G. Proakis, "Digital Communications", *4th Ed, New York: McGraw-Hill, 2000.*