FIA 2020/22
XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA
XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC
Florianópolis, SC, Brasil

3-D Audio Synthesis: A DIY Approach for HRIR Database Acquisition

dos Santos, A. N.[1]; Fernandes, D.[2]; Masiero, B. S.[3]

[1,3] Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas (Unicamp), Campinas, SP, Brasil, a264372@dac.unicamp.br, masiero@unicamp.br

[2] Escola Politécnica, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brasil, denisf@pucrs.br

Abstract

3-D audio synthesis is a computational method for generating and converting monaural sound sources into binaural sound, e.g., virtual spatial audio for headphones playback. Binaural synthesis requires the use of head-related impulse responses (HRIR), which can be measured or modeled, and are responsible for providing listeners with the sensation of sound spatialization. It is known that HRIRs are dependent on the anthropometric characteristics of individuals, thus, for a more realistic experience, it is best if the HRIRs are fitted to the listener's own physiology. This paper aims to study a technique for capturing HRIRs using binaural in-ear microphones, and a computational approach for filtering out unwanted measurement artififacts, thus creating a personal HRIR database. By generating excitation signals and reproducing it through an audio monitor in an office-like environment, a raw database was first obtained, containing pairs of impulse responses (IR) for a few different points in space, varying the monitor's horizontal and vertical angles with respect to the subject's head, but with fixed head-position and distance radius. Since these measurements capture not only the subject's HRIRs, but also the influence of the loudspeaker, the room and the microphones response, a linear predictive coding (LPC) filter was implemented, modeling approximate parametric IRs for these undesired components, thus enabling to filter them out. As a result, a small HRIR database was acquired, and used as a convolutional filter to virtualize monaural sound sources into 3-D audio. This measured database was then objectively compared with its filtered version, showing a superior signal-to-noise ratio (SNR). Subjectively, the acquired database was also compared with a public domain one, and a commercial virtual studio technology (VST) plugin, showing a fair level of satisfaction. Finally, a discussion about complexity of implementation versus satisfaction of results is presented, and further possibility of headphone equalization and head-tracking sensor integration as future work is proposed.

Keywords: binaural synthesis, convolutional filter, head-related impulse response, linear predictive coding.

PACS: 43.66.-x, 43.66.Pn, 43.72.Qr.

1. INTRODUCTION

Perhaps, sounds would be of less interest to humans if it could not be heard. A subject's perception of sounds through the auditory system is called hearing. The human auditory system (HAS) comprises a set of organs that work together to allow sounds to reach an individual's consciousness, enabling the appreciation of sounds' characteristics. Physiologically, the HAS is subdivided into outer, middle, and inner ear. The outer ear works as an acoustic shell, gathering sounds' energy. The middle ear assures that the energy transfer between outer and inner ear is efficient. Then, the inner ear converts sounds' mechanical energy into electrochemical impulses which are passed on to the brain via the auditory nerve. Finally, the auditory pathway ends in the upper areas of the temporal lobe cortex, where the auditory sensations are interpreted [1–4].

Figure 1 illustrates a simplified cross-section of the HAS, identifying the outer, middle and inner parts of the ear.

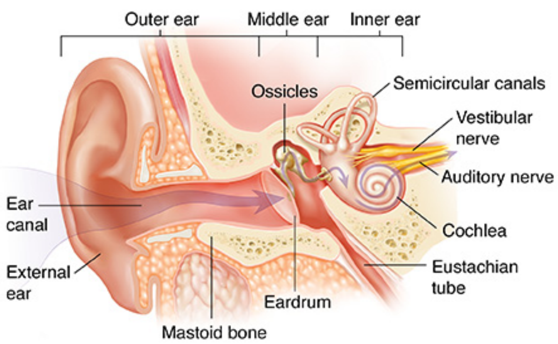


Figure 1: Simplified cross-section of the HAS [5].

Still, the study of human hearing must also account for the interactions between the physical properties of sounds and the physiological and psychological aspects of the HAS. Due to these aspects, humans cannot only perceive the level, pitch and timbre of sounds, but also the distance and direction of sound sources. A signal that represents an incoming sound received through the ears is called an auditory cue, which can be monaural or binaural. The distinction is that a monaural cue does not rely on any differences between sounds reaching the left and right ears to constrain the set of possible sound sources, whereas if it does, then it results in a binaural cue for localization. The perception of azimuthal angles between a sound source and the head (ϕ) is related to the inter-aural time difference (ITD) and the inter-aural level difference (ILD). ITD occurs when sound waves reach an individual's left and right ears at different instants of time, and ILD occurs due to the acoustic shadowing effect of a subject's head, which results in an intensity difference between the signal levels that reach both left and right ears, both depending on the angle ϕ. The perception of elevation angles between sound sources and the head (θ) occurs mainly due to the physiology of the outer ear, with its shape full of resonant cavities, that can amplify or attenuate frequencies according to its geometry and the direction of arrival of sound waves. Hence, the sounds that reach a person's eardrums can be considered filtered versions of the sound waves emitted by the sound sources. The mathematical models that describe these filters in time-domain are called head-related impulse responses (HRIRs), which vary their characteristics according to a subject's anthropometry. Therefore, the occurrence of sound events in a person's surroundings results in the perception of independent auditory images, that correspond to the distance and direction of each sound source. The superposition of these events is called an auditory scene, which relates to the physical surroundings of the listener [6–10].

Figure 2 illustrates the left and right ear HRIR measurements of a KEMAR head and torso simulator for a sound source at a distance of 3m from the center of the dummy head and located at $\phi = 60°$ and $\theta = 0°$.

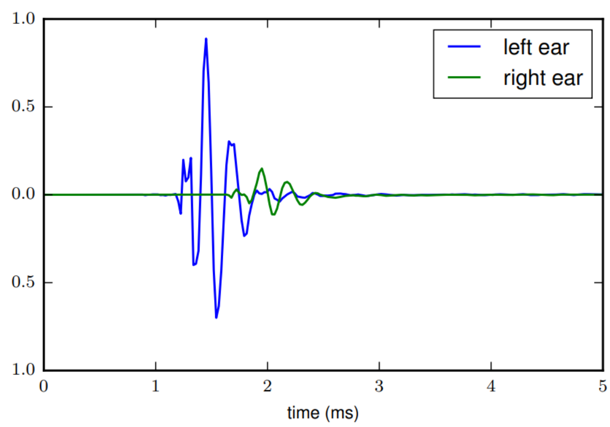


Figure 2: Example of an HRIR pair measurement [11].

Since the acquisition of personalized HRIR measurements is ideally required to be performed in anechoic spaces, to avoid interference from room impulse responses (RIR), e.g., echo, resonance, reverberation etc., this paper proposes a *do it yourself* (DIY) approach, to circumvent this idealism by modeling an inverse linear predictive coding (LPC) filter, to suppress unwanted components of HRIR measurements acquired in non-anechoic spaces.

The remainder of this paper is organized as follows: Section 2 describes the main theoretical foundations of this work, Section 3 details the development of this work, Section 4 presents results with some discussion, and, finally, Section 5 presents pertinent considerations to concludes this study.

2. FUNDAMENTALS

With the invention of the phonograph in the late 1870's, it became possible to record and playback sounds in physical media. Likewise, most early audio reproduction systems were monophonic, i.e., they used only one channel, which makes it difficult to distinguish sounds when mixed, due to low dimensionality. In the late 1930's, the stereo system was introduced, expanding phonographic systems to two channels, allowing for the creation of a panorama of sounds in a mix, thus enhancing the listening experience. In the late 1960's, the surround sound was introduced, first with the quadraphonic sound system, which never became as popular as its 1990's successor, the 5.1 standard. The reason for that was the home theater concept, which allowed consumers to have a cinematic experience at the comfort of their homes, merging sound and vision into an immersive experience. Subsequently, more channels were added to different loudspeaker array configurations in order to achieve even more immersive experiences, for both home and cinema theater applications (e.g., 7.1, 11.1, 22.2 etc.) [12].

However, all these systems have the common disadvantage of being channel-based audio (CBA) formats, which implies a dependency on the compatibility between audio production (recording, editing and mixing) and the playback system they are designed for, not to mention their portability issues. With the immense popularity of mobile devices, more recent applications, such as augmented and virtual reality (XR), are usually intended for headphones playback, which limits the representation of auditory scenes to only two channels. On top (of that), the usage of headphones often spoils stereophonic sound, by restringing the hearing experience to the impression that sounds are confined inside the listener's head [13]. To circumvent these issues, binaural sound production can be achieved by either natural recording or synthesis.

2.1 Binaural recording

Natural recording of binaural sounds can be accomplished with artificial head and torso simulators, e.g., Neumann's KU 100 (Figure 3a), GRAS' KEMAR (Figure 3b) etc., or binaural microphone sets, e.g., 3Dio's Free Space (Figure 3c), Hooke's Verse (Figure 3d) etc., all of which tries to mimic parts of the human physiology that contributes to the way we hear.

(a) KU 100.

(b) KEMAR.

(c) Free Space.

(d) Verse.

Figure 3: Examples of **(a)** an acoustic dummy head, **(b)** a head and torso simulator, **(c)** a binaural microphone set, and **(d)** an in-ear binaural microphone set [14–17].

Fundamentally, the technologies exemplified in Figure 3 incorporate matched microphone pairs, to capture the interactions between the physical properties of sounds and their artificial forms, e.g., pinna, head, torso etc. When the audio captured by these arrays is played back via headphones, listeners can often perceive the auditory images corresponding to the superposition of the recorded sound events. The greater the similarity between their forms and the physiology of the listener, the more realistic is the immersion. In the special case of in-ear binaural microphone sets, the very own physiology of the listener becomes part of the binaural audio production, i.e., if the person who wears them for the recording process is also the listener.

2.2 Binaural synthesis

In time-domain, binaural synthesis can be accomplished by convolving the time-domain representation of audio signals with HRIR pairs (as illustrated in Figure 4). An HRIR pair simply models how both left and right ears of an acoustic dummy, or a subject's head, interacts with sound sources in their surroundings. With fixed radius distance (r), for each possible ϕ and θ angles, a different HRIR pair must be either modeled or measured, which means that more pairs lead to a better representation of the physical surroundings of the listener [18, 19].

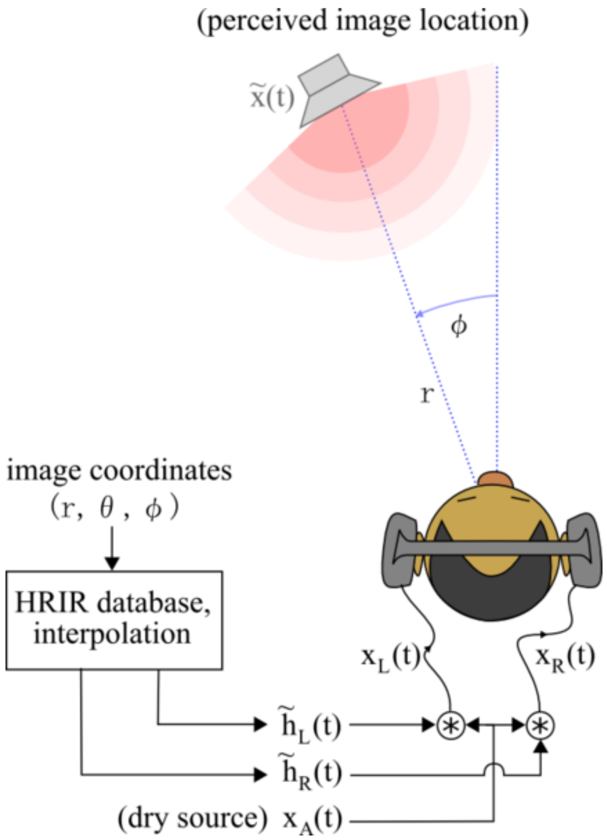


Figure 4: HRIR binaural synthesis scheme [20].

In the binaural synthesis scheme illustrated by Figure 4, r, θ and ϕ are the sound source coordinates, $x_A(t)$ is a dry audio signal, $\tilde{h}_L(t)$ and $\tilde{h}_R(t)$ are the left and right HRIRs, $(*)$ denotes the mathematical operation of convolution, $x_L(t)$ and $x_R(t)$ are the left and right channels of the synthesized binaural audio, and $\tilde{x}(t)$ is the perceived auditory image. Note that the scheme does not consider headphone equalization or head-tracking.

A large number of HRIR pairs (and head-related transfer functions (HRTFs), which are their frequency-domain counterparts) can be found in databases, e.g., the CIPIC HRTF database [21], the Listen HRTF database [22], the 3D3A Lab HRTF database [23] etc., yet, the hypothesis that a database that is tailored to the physiology of the listener can provide a more realistic experience, still holds for binaural synthesis.

Moreover, although many methods for HRIR / HRTF acquisition have been formerly proposed, for both acoustic dummy heads and human subjects, e.g., as in [24] and [25], these methods usually require experimental settings that include (but are not limited to) an anechoic chamber, professional monitors for the playback of excitation signals, acoustic dummy heads, head and torso simulators, binaural in-ear microphone sets, and appropriate sound-cards, not to mention programming softwares for data processing. Hence, we herein offer a low-cost methodology for personalized HRIR database acquisition and binaural synthesis.

3. DEVELOPMENT

The development of this project was constrained to the acquisition of HRIR measurements from human subjects, which exempted the use of an acoustic dummy head, or a head and torso simulator.

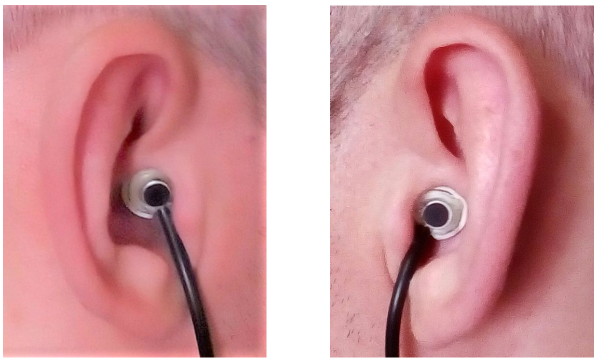
For generating the excitation signals (as well as data post-processing), programming and numeric computing platform MATLAB was chosen, running on a consumer level laptop without any special modifications, e.g., graphics card, sound card etc.

For the playback of excitation signals, a Genius' SP-HF800 Pro loudspeaker was used (directly connected to the laptop's line output), which is a $20W_{RMS}$, two-way loudspeaker, with a 2.75" driver unit [26], yet, only one of its loudspeakers was used, which is illustrated in Figure 7a.

For capturing the excitation signals, a binaural in-ear microphone set prototype was made, inspired by Hooke's Verse (Figure 3d) and Brüel & Kjær's 4101-B [27]. Figure 5 illustrates this prototype, which consisted of a pair of matched electret microphones, taken from consumer smartphone headsets, and mounted in an in-ear headphone, in the place of the loudspeakers.

Figure 5: Binaural in-ear microphone set prototype.

This way the ear canal was blocked by silicone ear-buds, and the microphones were facing outwards the ear, as illustrated in Figure 6.



(a) Left ear. **(b)** Right ear.

Figure 6: Prototype's ear-buds blocking the ear canal.

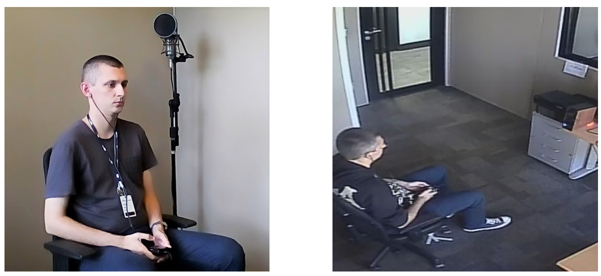
At the end of its cable, a 3.5mm TRS (tip, ring, and sleeve) stereo male plug was placed, with its tip connected the left-ear microphone, its ring connected to the right-ear microphone, and its sleeve connected to the signal ground.

To record the signals captured by the binaural in-ear microphone set prototype, a Marantz' PMD620MKII handheld solid state recorder was used. The reason for that, other than its availability, was that this equipment has a 3.5mm TRS female microphone input jack, which not only can record in stereo, but can also provide a phantom power of 5V, 1mA (maximum) [28], which is necessary to power the electret microphones. Figure 7b illustrates this equipment.

(a) SP-HF800 Pro. **(b)** PMD620MKII.

Figure 7: Playback and recording equipment.

The experiment was conducted with only one subject, in an office-like environment, with carpet floor, acoustic panels on the walls and ceiling, glass door and windows, some furniture, e.g., desk, chairs etc., and some equipment lying around, e.g., microphone stand, audio monitor etc., as illustrated in Figure 8.



(a) Subject. **(b)** Setting.

Figure 8: Experiment's subject and setting.

With this experimental setting, the excitation signals were produced by the laptop (running MATLAB), played back through the loudspeaker, convolved with the RIR and the subject's HRIR, captured by the binaural in-ear microphone set prototype, and stored into the recorder's memory card. These recordings were then passed on to the laptop again, where the post-processing was performed in MATLAB.

4. RESULTS AND DISCUSSIONS

Usually, an HRIR database is comprised of thousands of measurements for each ϕ and θ angles, but for this experiment, only five points in space were chosen, namely: $\phi = 0°$ and $\theta = 0°$ (*front*), $\phi = 180°$ and $\theta = 0°$ (*back*), $\phi = -45°$ and $\theta = 0°$ (*front-left*), $\phi = 45°$ and $\theta = 0°$ (*front-right*), and $\phi = 0°$ and $\theta = 45°$ (*top*).

For each ϕ and θ pair, nine measurements were performed, and then their mean values were computed. The subject was always sitting on a chair, 1m away from the loudspeaker, and with azimuthal plane roughly at the same height as the loudspeaker.

This procedure was then repeated, placing the microphones at the same spots as if they were inserted in the subject's ears. This was done in order to capture the response to the excitation signals without the interference of the subject's head. Figure 9 illustrates these results for the pair $\phi = 0°$ and $\theta = 0°$ (*front*).

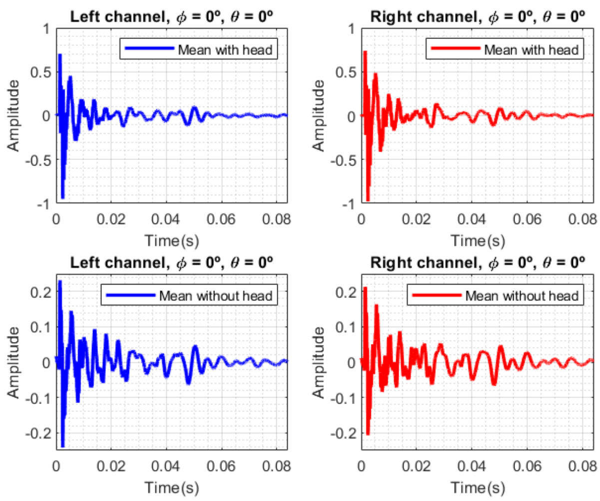


Figure 9: Measurement means with (top row) and without (bottom row) head for $\phi = 0°$ and $\theta = 0°$.

From Figure 9, it's evident that the impulse responses of the laptop, loudspeaker, room, microphones, and recorder, all interfere on the HRIR measurement.

To filter out these unwanted components, i.e., the IRs measured without head, an inverse filtering approach was adopted. Since the LPC analysis can be used to model signals, where a discrete-time sample can be estimated by a linear combination its past ones, signal modeling can be achieved in the form of an infinite impulse response (IIR) filter, according to

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{m=1}^{p} a_m z^{-m}}, \qquad (1)$$

where $H(z)$ is a transfer function in the z-domain, G is the filter gain, a_m is a set of linear coefficients, and p is the filter order. After analyzing the unwanted IRs, it becomes possible to synthesize them by means of their frequency responses. Since $H(z)$ is an all-pole filter, when inverting its numerator and denominator to perform the inverse filtering, $1/H(z)$ becomes a finite impulse response filter (FIR), which guarantees stability [29–32].

Of all methods available to compute G and $A(z)$, the Levinson-Durbin algorithm [33] is a popular choice, and was used in this experiment, with order p obtained empirically for each case. Figure 10 (top row) illustrates the resulting time-domain model for the left-channel of the pair $\phi = 0°$ and $\theta = 0°$, while Figure 10 (bottom row) is a zoomed-in version of the same signal, which illustrates the model's error.

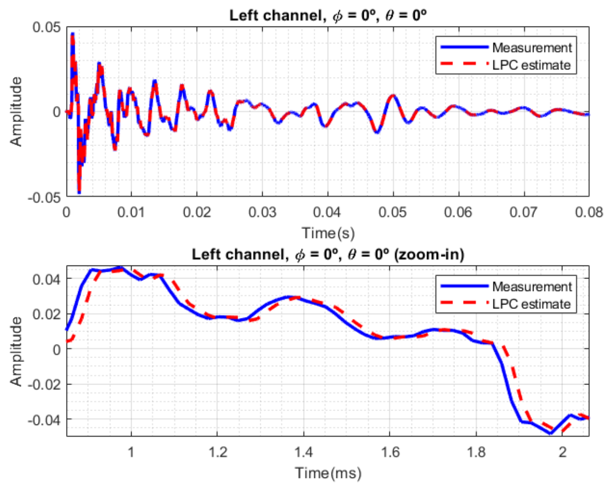


Figure 10: LPC resulting model for the left-channel of the pair $\phi = 0°$ and $\theta = 0°$ (top row), and zoom-in to show the model's error (bottom row).

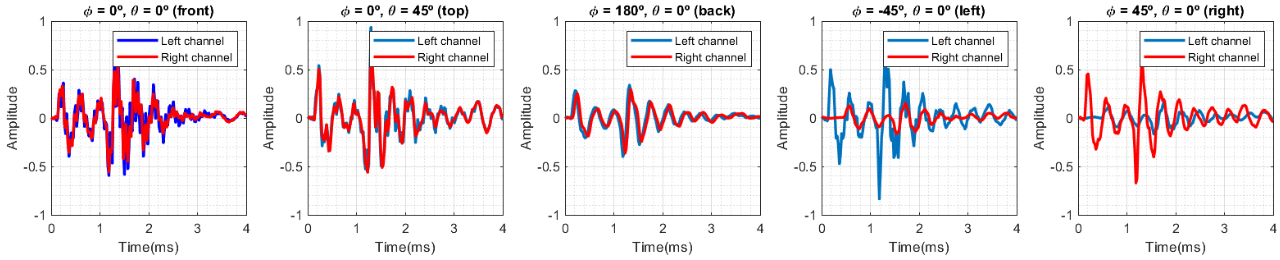


Figure 11: Resulting HRIR pairs, after inverse filtering, to suppress unwanted contributions.

Figure 11 illustrates the resulting database (truncated at 4ms for a better visualization), after filtering out the unwanted contributions to the HRIR pairs, using the aforementioned LPC analysis and synthesis.

Differences in the wave form of the the *front*, *top* and *back* pairs evince the modeling of monaural cues, whilst the inter-channel level and time differences (ICLD and ICTD) for the *front-left* and *front-right* pairs evince the modeling of binaural cues.

4.1 Objective evaluation

To objectively evaluate the inverse filtering approach, the signal-to-noise ratio (SNR) between the HRIR pairs *before* the inverse filtering, e.g., Figure 9 (top row), and the HRIR pairs measured without the head, e.g., Figure 9 (bottom row) was computed. This was achieved by computing the ratio of the signal's summed squared magnitude (A) to that of the noise

$$SNR\,[dB] = 10\log_{10}\left(\frac{A(signal)}{A(noise)}\right)^2, \qquad (2)$$

where the signals are the HRIR pairs *before* the inverse filtering, and noise are the HRIR pairs measured without the head.

Figure 12 (top row) illustrates the computed SNR values for the left and right channels on the horizontal plane, i.e., $-45^\circ \leq \phi \leq 180^\circ$ (left column), and on the vertical plane, i.e., $0^\circ \leq \theta \leq 45^\circ$ (right column).

Then, this procedure was repeated, for the SNR between the HRIR pairs *after* the inverse filtering, i.e., Figure 11, and the HRIR pairs measured without the head, e.g., Figure 9 (bottom row).

These values are similarly illustrated in Figure 12 (bottom row), and it evinces that the unwanted components of the measurements were suppressed at some level after the inverse filtering, however they still vary according to ϕ and θ.

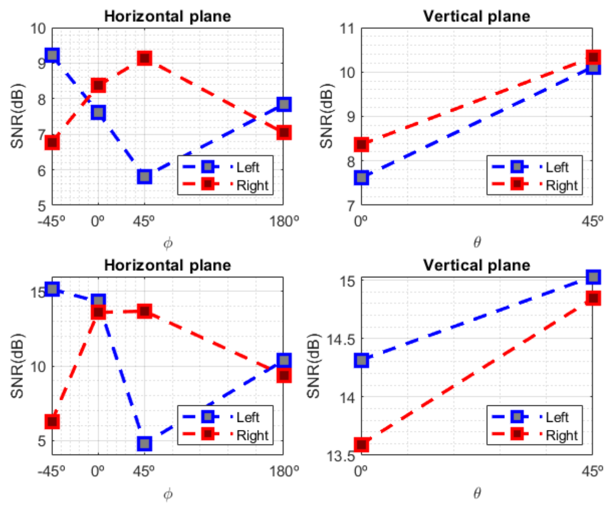


Figure 12: Horizontal and vertical planes SNRs, before (top row) and after (bottom row) inverse filtering, according to ϕ and θ.

4.2 Subjective evaluation

To access whether the acquired database is indeed useful for binaural synthesis, a subjective test was performed, in which 10 subjects anonymously answered a form[1].

The reference signal was a 5s long excerpt from a movie, where 3 sound events occur, one after the other, without overlaps. *Sound Event* 1 is a person speaking, *Sound Event* 2 is an old revolver's hammer click, and *Sound Event* 3 is another person responding to the first one.

[1] https://bit.ly/3NJCxPP (audio samples included).

This comprises a scene in which an impulsive sound (whose frequency response spreads all over the audible range) is sandwiched between 2 distinct speech signals, with different vocal qualities.

Subjects were asked to listen to this reference signal wearing headphones, an then answer 3 groups of questions.

For question group 1, each sound event was convolved with a different HRIR pair, so that *Sound Event* 1 would be perceived at the subject's *front-left*, *Sound Event* 2 would be perceived at the subject's *front*, and *Sound Event* 3 would be perceived at the subject's *front-right*. For this LFR (*left-front-right*) synthesis, 3 different HRIR pairs were used: the ones acquired in our experiment, the CIPIC HRTF database subject's 021 pairs, and Noise Maker's *Binauralizer* [34], a professional virtual studio technology (VST) audio plugin.

For question group 2, each sound event was convolved with a different HRIR pair, so that *Sound Event* 1 would be perceived at the subject's *front*, *Sound Event* 2 would be perceived at the subject's *top*, and *Sound Event* 3 would be perceived at the subject's *back*. For this FTB (*front-top-back*) synthesis, 2 different HRIR pairs were used: the ones acquired in our experiment, and the CIPIC HRTF database subject's 021 pairs. Binauralizer was not used here, because it can only synthesize azimuthal angles.

For question group 3, each sound event was convolved with a different HRIR pair, so that *Sound Event* 1 would be perceived at the subject's *front-left*, *Sound Event* 2 would be perceived at the subject's *back*, and *Sound Event* 3 would be perceived at the subject's *front-right*. For this LBR (*left-back-right*) synthesis, only Binauralizer was used.

After listening to each binaural sample, subject's were asked to evaluate whether their experience was excellent, good, fair, poor or bad, in terms of spatial hearing, and based on the previous knowledge of where each sound event was supposed to sound at.

Table 1 illustrates the results for the acquired database, in which it's evident that the LFR synthesis in mostly fair, and that the FTB synthesis is mostly poor.

Table 1: Acquired database's subjective evaluation

	Excell.	Good	Fair	Poor	Bad
LFR	10%	20%	40%	30%	0%
FTB	0%	10%	30%	40%	20%

Table 2 illustrates the results for the CIPIC HRTF database, in which it's evident that the LFR synthesis in mostly excellent, and that the FTB synthesis is mostly good.

Table 2: CIPIC HRTF database's subjective evaluation

	Excell.	Good	Fair	Poor	Bad
LFR	40%	30%	20%	10%	0%
LBR	30%	50%	10%	10%	0%

Table 3 illustrates the results for the *Binauralizer* VST audio pluguin, in which it's evident that both the LFR and LBR synthesis are mostly good.

Table 3: VST pluguin's subjective evaluation

	Excell.	Good	Fair	Poor	Bad
LFR	20%	50%	30%	0%	0%
FTB	10%	40%	10%	30%	10%

For each question group, subject's were also asked to descant over the following questions:

- Did you had any difficulty detecting a particular direction, e.g., "*confusion between front and back*", etc.?
- Did you suffer from internalization, i.e., the impression that sounds were confined to the space within your head?
- Do you have any considerations regarding the quality of the aforementioned samples, e.g., "*first one sounded muffled*", "*last one sounded bassy*", etc.?

For question group 1, some subjects reported that, for all sinthesis, they suffered from internalization regarding *Sound Event 2*, and that different coloration effects affected each synthesis. For the acquired database synthesis, sounds were described as "*sharp*", "*muffled*" and "*reverberated*". For the CIPIC HRTF database synthesis, sounds were described as "*clear*", "*dry*" and "*filtered*".

And for the VST pluguin synthesis, sounds were described as "*echoed*" and "*bassy*".

For question group 2, a few subjects reported that, for all sinthesis, there was confusion between all directions, that they suffered from internalization regarding all sound events, and that different coloration effects affected each synthesis. For the acquired database synthesis, sounds were described as "*lacking definition*", "*muffled*" and "*reverberant*". For the CIPIC HRTF database synthesis, no comments were made. And, for the VST pluguin synthesis, sounds were described as "*less reverberant*".

For question group 3, a few subjects reported that they suffered from internalization regarding *Sound Event 2*. No comments were made on confusion or coloration.

4.3 Discussions

It's important to emphasize that, giving that the experimental hardware used was either consumer level or prototyped, audio *definition* was at stake from the beginning, mainly because of the electret microphones used, with unknown frequency response, and the choice of excitation signals. Since, according to classical linear signals and systems theory, a system's transfer function can be acquired by measuring it's impulse response, a 1ms sound-step function was used, to approximately emulate a Dirac delta function. However, according to [35], a sine-sweep function would have been more appropriate.

Moreover, since the subjective test was remotely conducted due to COVID restrictions, there was no control over the experimental listening settings, i.e., in which environment did the subjects self-applied the test, or what equipment each subject had at hand (in-ear or over-ear headphones).

Also, according to [36–38], for a better experience, binaural synthesis should always match each subjects' HRIR measurements, which was never the case in our experiments.

Another important factor is that, apart from headphones' loudspeaker frequency response, acoustic coupling between headphones and *pinnae* can produce noticeable frequency notches, that may vary even due to headphones' repositioning, which causes coloration and the perception of artifacts [39–43].

Furthermore, confusion between angles on the *sagittal* plane are explained by the *cone of confusion*, in which different azimuthal and/or elevation angles share the same binaural cues, i.e., ILD and ITD [7].

On top of that, some studies claim that head-tracking can interfere on the internalization of auditory events, since in natural sound scenes, subjects can benefit from head-tilting movements to better estimate the direction of sound sources [44, 45].

5. CONCLUSION

In this paper, a DIY approach for HRIR database acquisition was proposed. Methodological complexity can somewhat be regarded as fair, compared to cost of implementation.

Moreover, results were subjectively evaluated as mostly fair, and considering that it enables the measurement of a personalized database, it evinces an equitable compromise.

Improvements are evidently on demand, primarily on hardware quality, e.g., the in-ear microphone set, secondly on the excitation signal's choice, thirdly on data post-processing, i.e., headphones equalization for suppressing the perception of coloration and undesired artifacts, and lastly on the addition of head-tracking to circumvent internalization issues.

Ultimately, the proposed approach for HRIR database acquisition can evidently be used to virtualize spatial audio for headphones playback, yet it still struggles to deliver accurate representations of sound scenes, regarding fidelity.

6. ACKNOWLEDGMENTS

This work was partially supported by the São Paulo Research Foundation (FAPESP), grant #2019/22795-1. The opinions, hypothesis and conclusions or recommendations expressed in this material are the authors' responsibilities, and not necessarily reflect FAPESP's views.

REFERENCES

[1] Antônio Amorim. *Fonoaudiologia geral*. Livraria Pioneiro Editôra, 1972.

[2] Brian C.J. Moore. Handbook of perception and cognition: Vol. 6. hearing, 1995.

[3] Brian C.J. Moore. *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons, 2007.

[4] Richard F. Lyon. *Human and machine hearing*. Cambridge University Press, 2017.

[5] Anatomy of the ear. URL https://bit.ly/3yLKFuP.

[6] K. Blair Benson. *Audio engineering handbook*. McGraw-Hill Companies, 1988.

[7] Frank Filipanits. Design and implementation of an auralization system with a spectrum-based temporal processing optimization. *Master's Research Project, University of Miami May*, 1994.

[8] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 99–102. IEEE, 2001.

[9] Ennio Cruz da Costa. *Acústica técnica*. Editora Blucher, 2003.

[10] Jeroen Breebaart and Christof Faller. *Spatial audio processing: MPEG surround and other applications*. John Wiley & Sons, 2008.

[11] Simulating room acoustics. URL https://bit.ly/3lqLzVK.

[12] Stanley R. Alten. *Manual del audio en los medios de comunicación*. Escuela de cine y vídeo, 1994.

[13] Ferdinando Olivieri, Nils Peters, and Deep Sen. Scene-based audio and higher order ambisonics: A technology overview and application to next-generation audio, vr and 360 video. *EBU Tech*, 2019.

[14] Neumann. Dummy head ku 100. URL https://bit.ly/38D8uKT.

[15] Gras 45bb kemar head & torso, non-configured. URL https://bit.ly/3yLSv80.

[16] Free space binaural microphone. URL https://bit.ly/3Nm0axD.

[17] Bluetooth in-ear binaural microphones. URL https://bit.ly/3wFd0Ai.

[18] Alan V. Oppenheim, Alan S. Willsky, Syed Hamid Nawab, Gloria Mata Hernández, et al. *Signals & systems*. Pearson Educación, 1997.

[19] Bhagwandas Pannalal Lathi and Roger A. Green. *Linear systems and signals*, volume 2. Oxford University Press New York, 2005.

[20] 3d audio effect. URL https://bit.ly/3FWvvVo.

[21] The cipic interface laboratory home page. URL https://bit.ly/3sNj60t.

[22] Listen hrtf database. URL https://bit.ly/3Lwcome.

[23] The 3d audio and applied acoustics (3d3a) laboratory at princeton university. URL https://bit.ly/3yRuooj.

[24] William G Gardner and Keith D Martin. Hrtf measurements of a kemar. *The Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995.

[25] Bruno Masiero, Martin Pollow, and Janina Fels. Design of a fast broadband individual head-related transfer function measurement system. In *Proc. Forum Acusticum*, page 136. Danish Acoustical Society, 2011.

[26] Elegant bookshelf speakers. URL https://bit.ly/3FZh9nc.

[27] Conjunto de microfone intra-auricular binaural. URL https://bit.ly/3MyAG03.

[28] Model pmd660 user guide. URL https://bit.ly/3MIB0JC.

[29] Lawrence R Rabiner. Digital processing of speech signal. *Digital Processing of Speech Signal*, 1978.

[30] Panos E Papamichalis. *Practical approaches to speech coding*. Prentice-Hall, Inc., 1987.

[31] Juan I Godino-Llorente, Santiago Aguilera-Navarro, and Pedro Gómez-Vilda. Lpc, lpcc and mfcc parameterisation applied to the detection of voice impairments. In *Sixth International Conference on Spoken Language Processing*, 2000.

[32] Usha Sharma, Sushila Maheshkar, and AN Mishra. Study of robust feature extraction techniques for speech recognition system. In *2015 International conference on futuristic trends on computational analysis and knowledge management (ABLAZE)*, pages 654–658. IEEE, 2015.

[33] Lpc analysis and synthesis of speech - matlab & simulink. URL https://bit.ly/3NswwaK.

[34] Binauralizer. URL https://bit.ly/3z6AGR6.

[35] Swen Müller et al. Transfer-function measurement with sweeps director's cut including previously unreleased material. 2001.

[36] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen. Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321, 1995.

[37] William M Hartmann. How we localize sound. *Physics today*, 52(11):24–29, 1999.

[38] Dorte Hammershøi and Henrik Møller. Binaural technique—basic methods for recording, synthesis, and reproduction. *Communication acoustics*, pages 223–254, 2005.

[39] Bruno Masiero and Janina Fels. Perceptually robust headphone equalization for binaural reproduction. In *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[40] Josefa Oberem, Bruno Masiero, and Janina Fels. Authenticity and naturalness of binaural reproduction via headphones regarding different equalization methods. In *Proceedings of AIA-DAGA 2013 Conference on Acoustics*, 2013.

[41] Javier Gómez Bolaños, Aki Mäkivirta, and Ville Pulkki. Automatic regularization parameter for headphone transfer function inversion. *Journal of the Audio Engineering Society*, 64(10):752–761, 2016.

[42] Isaac Engel, David Lou Alon, Philip W Robinson, and Ravish Mehra. The effect of generic headphone compensation on binaural renderings. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[43] Guangju Li, Chengshi Zheng, Xiaodong Li, Teng Yu, Stefan Bleeck, and Jinqiu Sang. Evaluation of headphone phase equalization on sound reproduction. *Applied Acoustics*, 156:208–216, 2019.

[44] Stephan Werner, Georg Götz, and Florian Klein. Influence of head tracking on the externalization of auditory events at divergence between synthesized and listening room using a binaural headphone system. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[45] Nolan Lem and Jens Ahrens. Head tracking binaural localization system for horizontal sound source detection.