Audio Engineering Society

Conference Paper 71

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

Augmented Audio with Behind-the-Head Listening Device

Otto Puomio[1], Raimundo Gonzalez Diaz[1], and Tapio Lokki[1]

[1]*Aalto University, Department of Computer Science, POBox 13000, 00076 Aalto, Finland*

Correspondence should be addressed to Otto Puomio or Raimundo Gonzalez
(`otto.puomio@aalto.fi,raimundo.gonzalezdiaz@aalto.fi`)

ABSTRACT

This study case explores a novel personal listening device in the context of spatial audio which, in normal use, is located behind the listener's head. Several techniques were applied to binaural signals with the intention of rendering spatial sound in front of the listener. Implementations of binaural, ipsilateral compensation, crosstalk cancellation techniques and combinations displayed mixed results due to the device's haptic feedback and proximity to the user's head. Listening tests revealed that the implemented techniques worked to some extent opening new research questions for future research.

1 Introduction

Virtual and augmented reality audio has opened the possibility for innovative personal listening devices. So far, headphones have been a common solution for the rendering of reasonably convincing binaural audio while still presenting certain challenges, e.g., inside-the-head localization. One solution to achieve better externalization is the rendering of binaural audio through loudspeakers applying crosstalk cancellation which, as all spatial audio rendering techniques, it also has its pros and cons. In this case study, we investigate a novel audio reproduction product in order to see if it could be used as a spatial audio reproduction device for personal use. To explore this question, we implemented several common and uncommon spatial audio techniques with this device and evaluated them with a listening test.

1.1 Behind-the-head Listening Device

HUMU[1] Augmented Audio Cushion™ can be described as a cushion and personal audio listening device, which enables the user to perceive both audio and vibrations at the same time. Shaped similarly to a rectangular cuboid, it produces sound through two electromechanical actuators suspended inside the pillow by various layers of foam and soft material. The actuators are separated by 20 cm from their axis, each one 10 cm from the device's center. The cushion's different layers maintain the structural integrity of the device while also making it comfortable to use as a device. Though there is no mandatory physical configuration on where the user should place this device, the cushion is recommended to be used as a pillow with the user's head placed in the middle as shown in Fig. 1. This way, each actuator is located close to one ear providing a kind of stereo listening setting. Also tactile feedback is

---

[1]https://humu.fi

Fig. 1: Kemar dummy head and HUMU cushion arranged in the "Ideal Listening Position".

felt at low frequencies due to the mechanical coupling between the users head and the vibrations of the device. For the purpose of this case study, this way of using the listening device is defined as the *"Ideal Listening Position"* (ILP).

In order to find out the acoustic behaviour of this device, impulse responses (IR) were measured to characterize the frequency response at the ILP. IRs were acquired directly in front of the actuators, 3 cm away from the pillow's surface as shown in Fig. 2a. These positions were chosen as they correspond to the position of the users in the ILP. The measurement equipment used for these measurements were two $1/4$" GRAS pressure microphones connected through a MOTU UltraLite MK3 Hybrid audio interface to a laptop computer. Impulse responses were acquired using 10 second sine sweeps [1] in a standard classroom by windowing the direct sound from the whole response.

The magnitude of the frequency response acquired for these positions and their average is presented in Fig 2b. The response of the pillow is uneven containing large peaks at 2.2 kHz, 4.7 kHz, and 15 kHz as well as major notches around 200 Hz and 2 kHz. However, it was not possible to determine if these acoustic features are due to the electromechanical elements themselves or due to the soft enclosure which they are a part of. Slight high frequency boost might be a designed feature as the user's pinnae attenuates high frequencies in the ILP.

2   Spatial Reproduction Methods

In an intent to reproduce spatial audio through the device that is perceived in front of the listener, two-channel spatial audio signals were rendered in a similar

(a) HUMU cushion measurement setup. Two $1/4$" GRAS pressure microphones positioned 3 cm from from the surface of the pillow, directly in front of its drivers.
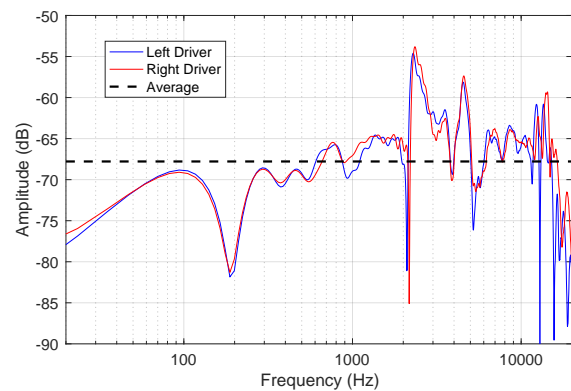
(b) Measured frequency responses from drivers.

Fig. 2: Measurement of the listening device.

fashion to headphone-based spatial audio. Audio signals were convolved with Binaural Room Impulse Responses (BRIR) [2] and then acoustically compensated to achieve similar reproduction as for loudspeaker-based binaural audio as described in Section 2.2.

2.1   Binaural Rendering Method (BIN)

A traditional way to render spatial audio is to apply BRIRs to the input signal. The BRIRs contain the spatial information for the direct sound and room reflections as well. In other words, BRIRs can be thought of as being a combination of Spatial Room Impulse Responses (SRIR) and Head-Related Transfer Functions (HRTF).

As shown in Fig. 3, the binaural rendering, including the room response, is performed in a different way. Instead of recording and applying BRIRs directly, we
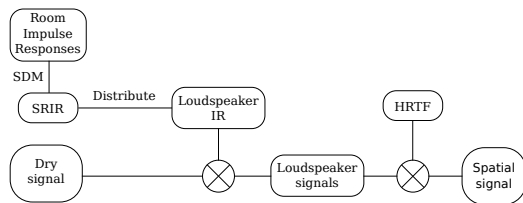
Fig. 3: Rendering process to create binaural signals.

Fig. 4: Acoustic transfer paths of the listening device cushion. X_L and X_R are the incoming binaural signals. P_L and P_R are the structure-borne transfer paths between the transducers and the ipsilateral surface of the cushion. BC indicates the mechanical transmissions of vibrations from the pillow to the user's head. SB refers to the structure-borne crosstalk from transducers to the contralateral surface of the pillow. H_{LL} and H_{RR} are the ipsilateral paths between actuators and ears while H_{RL} and H_{LR} are the contralateral paths. These four transfer paths form an acoustic transfer matrix H. Y_L and Y_R are the signals presented at the listener's ears.

present the acoustic space via a set of virtual loudspeakers. For each source-receiver pair, we measured a set of IRs of the room with a microphone array. The IRs were then analyzed with the Spatial Decomposition Method (SDM) [3] in order to generate Spatial Room Impulse Response (SRIR). This SRIR is further distributed to a position-optimized loudspeaker setup [2], which provides loudspeaker-wise IRs working as a convolution reverb. Finally, the loudspeaker signals are convolved with the dry input signals and later with HRTFs to generate the spatialized binaural output.

In this paper, the used HRTFs were acquired from the CIPIC public-domain database, which contains a comprehensive set of HRTF measurements for the KEMAR dummy head [4]. The SRIR in turn was acquired from real measurements from a medium sized recording studio using an array of six microphones [3].

2.2 Reproduction Compensation

To reproduce binaural sound properly, the output signal should arrive exclusively to one ear in similar way as with headphones. To achieve this, we implemented both Crosstalk Cancellation (CTC) and Ipsilateral Compensation (IPS) to compensate for the various acoustic transfer paths between the surface of the cushion and the ears of the user. Figure 4 depicts these paths that include reflections and scattering effects between the vibrating surface of the cushion and the torso, head, and ears of a listener. Sound pressure measurements inside the cushion revealed that structure-borne contralateral transfer paths, represented in the Fig. 4 by SB, had a 10 dB attenuation in comparison to the ipsilateral structure-borne transfer paths P_L and P_R. Therefore, the structural crosstalk was assumed to be minimal and the effect of SB to be negligible.

2.2.1 Transfer Paths & HRTFs

The transfer functions required for implementing the compensation filters were initially acquired from IR
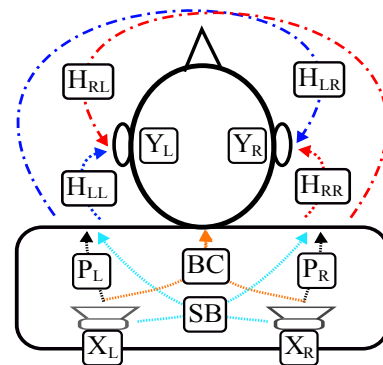
measurements of a KEMAR dummy head and a listening device in the ILP as shown in Fig. 1. The measurements revealed a 9 dB level difference between each ear and its contralateral speaker. They also displayed very unstable spectra, displaying several strong peaks and notches as well as near-field effects. Tikhonov regularization and spline interpolation methods were applied to stabilize these filters without success due to multiple drastic singularities of the spectra at high frequencies. Filters implemented with these near-field HRTFs displayed strong and unpleasant spectrum coloration. Potential causes for the instability of these measurements are the resonances produced by the narrow gaps between the dummy head and the surface of the pillow as well as the near-field acoustic behaviour of the cushion.

Thus, an alternative solution was implemented. Compensation filters were instead designed using far-field HRTFs with similar angle configurations between the drivers of the listening device and a head in the ILP. These HRTFs produced much more stable filters with

fewer and more manageable spectrum singularities. The justification for using these transfer functions was the assumption that far-field HRTFs would share similar acoustic head-shadowing and pinnae effects as their near-field counterparts. The database used for this implementation was CIPIC [4].

2.2.2 Crosstalk Cancellation (CTC)

Initially proposed by Bauer in 1961 [5] and implemented by Schroeder in 1970 [6], Crosstalk Cancellation (CTC) is a method for acoustic reproduction that uses two or more speakers to generate binaural signal at the listener's ear. For the purpose of this study case, CTC was implemented to produce ipsilateral compensation as well as to cancel the interference produced by the cushion's contralateral drivers on each ear.

The system in Fig. 4 can be expressed in matrix form:

$$\begin{bmatrix} X_L \\ X_R \end{bmatrix} \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix} = \begin{bmatrix} Y_L \\ Y_R \end{bmatrix} \quad (1)$$

In order to achieve listening conditions similar to headphone reproduction, both the input vector $\mathbf{X}$ and output vector $\mathbf{Y}$ must be equal. Therefore some filter matrix $\mathbf{C}$ must be applied to compensate for the acoustic transfer matrix $\mathbf{H}$, thus $\mathbf{XHC = Y}$. Hence, the product of the $\mathbf{C}$ and $\mathbf{H}$ must be unity and $\mathbf{C}$ must equal to the inverse of $\mathbf{H}$, i.e., $\mathbf{C = H^{-1}}$. Finally, $\mathbf{C}$ becomes 4x4 filter matrix with a determinant:

$$\mathbf{C} = \begin{bmatrix} H_{RR} & -H_{RL} \\ -H_{LR} & H_{LL} \end{bmatrix} \frac{1}{H_{RR}H_{LL} - H_{RL}H_{LR}} \quad (2)$$

2.2.3 Ipsilateral Compensation (IPS)

Assuming the acoustic shadow effect of the head is strong enough, the contralateral paths H_{RL} and H_{LR} can be considered negligible. Therefore the acoustic transfer matrix $\mathbf{H}$ reduces to the diagonal matrix presented in Eq. (4), which contains only the ipsilateral paths

$$\mathbf{H} = \begin{bmatrix} H_{LL} & 0 \\ 0 & H_{RR} \end{bmatrix} \quad (3)$$

When $\mathbf{H}$ is inverted, $\mathbf{C}$ simply becomes a matrix of two inverse filters

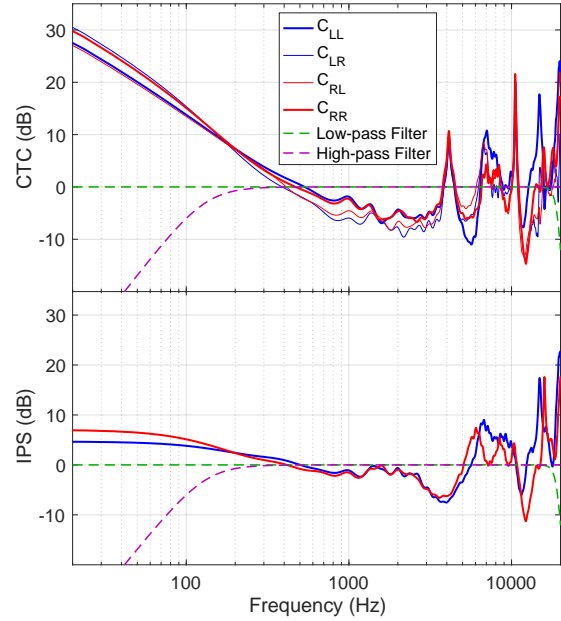$$\mathbf{C} = \begin{bmatrix} \frac{1}{H_{LL}} & 0 \\ 0 & \frac{1}{H_{RR}} \end{bmatrix} \quad (4)$$

Fig. 5: CTC and ipsilateral filter and equalizing filters.

2.2.4 Stability & Regularization

Inverse filters tend to be unstable due to numerical singularities and ill-conditioned determinants. For example, high frequency features of the HRTF, such as the pinna notches, become peaks when inverted, inducing strong coloration into the final signal. A similar scenario occurred with the implementation of the CTC and IPS compensation filters. Figure 5 illustrates the two unregularized filters displaying several peaks at high frequencies. In addition, there is strong low frequency amplification in the case of the CTC filter. Existing methods using Tikhonov and Minimum-Phase reconstruction to regularize inverse filters [7, 8] were implemented, but still strong coloration and distorting were displayed. For this instance, it was assumed that the high frequency region of the filter, 15 kHz and above, would not have a significant impact on the quality of sound localization and therefore an additional low-pass filter with a cut-off frequency at 17 kHz was applied. In addition, a high-pass filter with a 100 Hz cut-off was applied both to control the high gain from the CTC filter at low frequencies as well as to reduce the device's haptic response. These solutions still displayed some coloration for signals with strong high frequency energy but maintained the spatial perceptual properties.

2.3   Methods & Combinations

Initial spatial audio implementations on the listening device involved the direct use of binaural signals with no acoustic compensation. Though this did not shift the spatial audio image front, it did widen the perceived listening space in the back. Once CTC and IPS compensation were applied to the rendered signals, initial evaluations indicated that the perceived audio image shifted outside of the pillow to either the listener's sides (IPS) or inside his head (CTC).

In an intent to extend the wide behind-the-head audio image provided by the purely binaural rendering (BIN), the authors combined the techniques with the other more frontal ones, i.e. CTC and IPS. Audio signals rendered with the aforementioned methods were simply added together in hopes to combine their perceptual properties. Initial observations of these combinations indicated that the final spatial audio image included spatial properties from each method. The combinations generated were the following:

1. BIN + CTC

2. BIN + IPS

3. BIN + CTC + IPS

3   Test setup

In the listening test, we wanted to investigate the differences of the compensation techniques presented in Section 2. Preliminary listening indicated that there could be differences in perceived direction and width of the sound field. Therefore, the listening test focused on describing the perceived direction of the sound.

We decided to run a pilot test to refine the main test. The pilot test was conducted with 12 subjects (excluding the authors) that were audio experts and researchers from the lab. The subjects used a colored pencil to mark the direction they heard the sound to come from. The pilot test included all the technologies described in Section 2. After a short training session, all the samples were played in one go with no chance for replays. The test results indicated that the sound was most often located out the subject's line of sight, particularly into the device right behind the subjects' head. However, several subjects also reported hearing the sound coming frequently from the above. The subjects also told

that in some cases the sound coming from the back did not appear to come from the device but from further back (i.e. the chair) instead. In addition, the subjects reported the haptic response to collapse the perceived direction back to the device.

As stated above, few subjects reported that the haptic response kept the sound image in the back. We reasoned that in this case, it should be easier to achieve the goal to render frontal sounds by eliminating the device vibration. Therefore, we decided to filter out the haptic response with a 300 Hz high-pass filter in the actual listening test.

The listening test was completed by 15 expert listeners. From these subjects, the data of one subject was excluded due to incorrect answer technique. The rest of the results were processed as described later. The subjects received no compensation from completing the test.

The listening test was held in the same room where we conducted the initial measurements of Section 1.1. The tables in the room were moved to the sides, leaving an open space in the middle. A high-backrest chair was then placed in the middle of that space. The backrest supported the device against the listener's neck and head, ensuring the ILP during the experiment.

The test consisted of six technologies, four sound clips and three replications, totaling 72 sound samples, played in random order to subjects. The technologies were the ones presented in Section 2, namely stereo, BIN, CTC, IPS, BIN+IPS, and BIN+IPS+CTC. Stereo recording represented a low anchor and a baseline for all the other methods. BIN response was obtained by convolving this recording with a dry room SRIR as described in Section 2.1. IPS and CTC based their renderings on BIN, further applying their compensations as described in Section 2.2. The last three technologies were linear combinations of the three signals, that is BIN+IPS and BIN+CTC+IPS. The pilot test indicated that BIN+CTC did not differ much from the its base technologies, therefore left out from the test set.

The used sound signals were selected to have varying properties. The selected signals were speech (mono), drums (mono transient source), an orchestra (stereo music) and a robot (a moving stereo noise source). Each sound clip lasted 10–15 seconds. As described before, all sound clips were filtered with a 300 Hz high-pass filter in order to remove the haptic response.
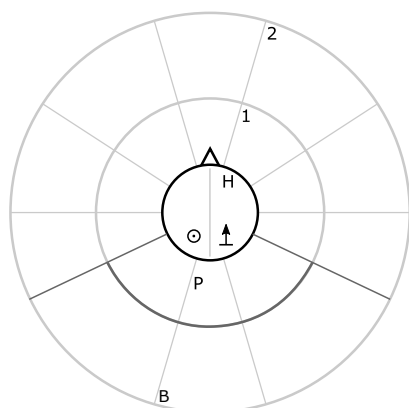
Fig. 6: The revised sound sample answer layout. The subjects were asked to mark the direction the sound was perceived to come from. The areas marked with letters were (P) inside the device (B) behind the device (H) inside (circle with a dot) or above (upright arrow) the head (1) sound within hand's reach (2) sound further than hand's reach.

The answers were collected with an image illustrated in Figure 6. The Image is radially divided into sectors, each box indicating a certain direction and distance from the listener. The distances are determined as (P) sound coming from the inside of the listening device; (B) sound coming behind the device; (H) sound inside/above the head; (1) sound within the subject's hand's reach; and (2) sound further than hand's reach. Moreover, the head region is divided to two boxes, ($\odot$) being sound inside and ($\uparrow$) above the head. In the listening test, the subjects' task was to tick all the boxes that they perceived the sound to come from.

The subjects gave their answers on paper that were scanned, aligned and de-noised after answering. This was followed by answer indexing and extraction. The extracted results were first averaged over replications and then over subjects. These "selection probabilities" were finally visualized over the original template image.

Quantified results allowed us to use statistics to analyze the results in detail. We assumed the data points to be independent of each other and that there was no learning or fatigue effects present. The first assumption together with the non-normality of the data lead us to use the Kruskal-Wallis test, and the latter assumption allowed us to average the per-subject replications. The averaged data was then divided sound sample-wise and analyzed with IBM SPSS statistics software.

4 Results

The results of the listening test are visualized in Fig. 7. Between the six compared technologies, Kruskal-Wallis test shows that there is a statistically significant difference in the P-mid sector in speech ($\chi^2(5) = 17.027$, $p = 0.004$) and drums ($\chi^2(5) = 16.126$, $p = 0.006$). These two cases were further analyzed with Kruskal-Wallis paired comparison with Bonferroni correction. In the speech case, significant differences were found between stereo–IPS ($p = 0.008$) and stereo–BIN ($p = 0.037$) pairs, whereas drums showed significant difference only for stereo–BIN+IPS ($p = 0.006$) pair. Moreover, the robot sound sample had significant differences between technologies in the overhead sector ($\chi^2(5) = 15.886$, $p = 0.007$). Further analysis reported significant differences in two cases, namely stereo–IPS ($p = 0.028$) and stereo–BIN+IPS ($p = 0.038$). The changes in all other sectors were not statistically significant.

The main listening test collected less verbal feedback than the pilot test. Only few subjects reported audible noise on some test samples.

5 Discussion

The listening test indicated that BIN, IPS and BIN+IPS technologies move the perceived sound away from the center of the device with speech and drums samples. The sound image were wider and for a few subjects even brought sound to the front of them, as was desired. Yet in the end, the ultimate goal was not achieved.

Reproducing binaural sound most probably failed because the attenuation for contralateral ears was not enough. Our system was measured to have 9 dB broadband attenuation, which was found helpful. In contrast, complete passive cancellation would require at least –15 dB channel separation for narrow-band and –20-25 dB for broad-band signals [9]. Moreover, suboptimal source separation is reported to increase degraded performance located behind the listener [10], also supporting the theory about insufficient amount of source separation.

Verbal feedback let us suspect that head movements collapse the created sound field back to the device.
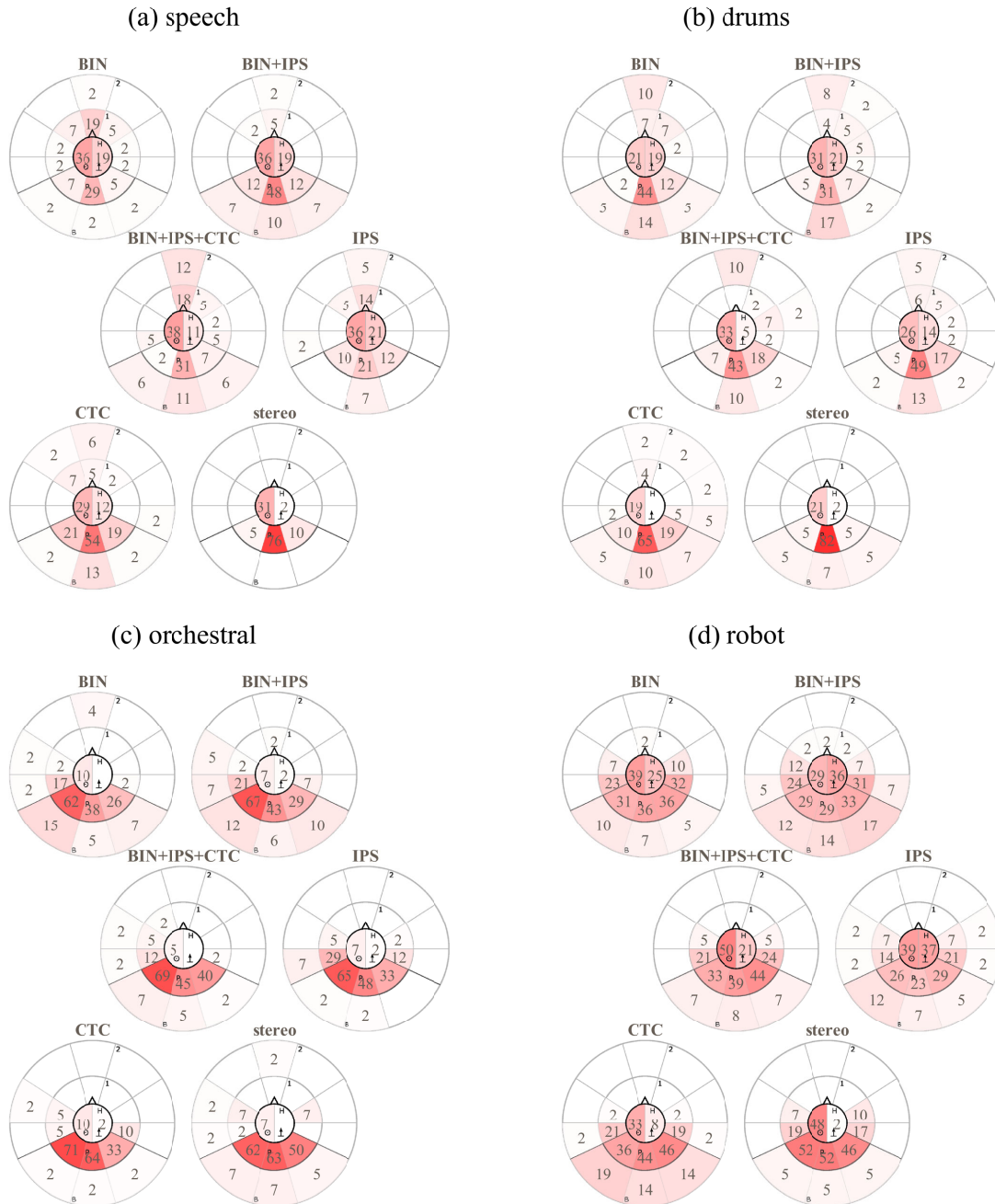
(a) speech

(b) drums

(c) orchestral

(d) robot

Fig. 7: Results of the listening test (N=14). Collective results of the subjects are grouped first by sound sample and then by technology. Sound samples are (a) speech, (b) drums, (c) orchestral and (d) robot. Technologies in turn are stereo, binaural (BIN), ipsilateral (IPS), crosstalk cancellation (CTC) and combinations of the latter three. The numbers and color intensities of the sectors indicate how probably the subject has marked the sector (in percent).

The phenomenon is also recognized in binaural headphone rendering where head movements tend to move the sound inside the listener's head. There, the phenomenon is fended off with head-tracking, which would most probably help in this case as well. Nevertheless, head-tracking was not implemented mainly due to time constraints, yet it should definitely be researched on this kind of device.

Interestingly, the CTC appeared to not work with the listening device. However, the implemented CTC was designed on far-field HRIRs instead of the measured BIRs due to instability issues. In addition, the listener position is located in the acoustic near-field of the device, which makes controlling the phase response of the CTC filters hard. There might also be other prominent transfer paths to the contralateral ear, as presented in Fig. 4. Literature hints that the reason could partly be the mismatched ITD and ILD cues. Akeroyd et al. [11] had simulated the effect of mismatched HRIRs of the crosstalk cancellation filters and of the listener. They reported that this mismatch reduces the channel separation drastically from –60-70 dB to approximately –25 dB. They reasoned that the reduction was caused by the differences in ITD and ILD cues. In our case, the HRIR mismatch was not limited to the filter and the listener only, because the system mismatched the distance of the HRIRs as well.

It is hard to tell whether removing the haptic response affected the listening test results or not. The exact comparison between the pilot and the actual test is not possible due to the changes in the measurement procedure. Nevertheless, to compare the two, we aligned the result image sets on top of each other in order to get hints of possible changes. This way no remarkable changes could be found between the sets. This implies that the haptic response had only a limited impact on the perceived direction, therefore requiring more research to prove its effect.

The inverse filtered signals (BIN and IPS) were definitely spectrally colored, yet we decided not to investigate the effect in detail. Nevertheless, it is highly probable that coloration caused the perceived elevation of the sound detected in samples involving IPS and especially with the robot sample. The perceived elevation is usually associated with a narrow-band peak at 7–9 kHz frequency band [12, 13], which lines up with the used ipsilateral filters shown earlier in Fig. 5. There, the right filter has a 8.4 kHz peak in the response.

As the robot sound started from the right side of the subject, the peak in question probably dominated the perceived elevation over the cues given by the left ear filter. In any case, even though the spectral peak affected the perceived elevation, it should have no effect on the azimuth perception.

6 Summary

We studied spatial audio rendering on a novel personal listening device, which includes also haptic feedback. The main objective was to find a rendering method that would enable to render sound sources around the user's head, in particular in front as the applied device is located behind the user. Three different rendering algorithms were tested; binaural rendering including a small room in the applied BRIRs, binaural audio with ipsilateral compensation and with basic crosstalk cancellation method. The capabilities of different methods and combinations of them were evaluated with a listening test, which incorporated various sound signals.

The results were promising, but we did not manage to produce sound to be perceived in front of the listener. We suggest two main factors for this outcome. First, the used far-field HRTFs did not create a sufficient channel separation due to near-field effects of the device. Second, the lack of head-tracking most probably collapsed the spatial image back to the device in some cases. There was also audible spectral coloration in the samples which affected the perceived elevation of the sound. However, the coloration was reasoned not to explain the missing frontal sound image.

This case study revealed that rendering the spatial sound with such a near-field personal device is indeed quite challenging. The list of future work items is long ranging from individualized HRTFs and head-tracking to more sophisticated crosstalk algorithms and isolated tests on haptics. To conclude, we wish that the presented case study paves the way for innovative research on novel personal listening devices to find out ways to use them in immersive spatial audio reproduction.

References

[1] Farina, A., "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," *110th AES Convention*, 2000.

[2] Puomio, O., Pätynen, J., and Lokki, T., "Optimization of Virtual Loudspeakers for Spatial Room Acoustics Reproduction with Headphones," *Applied Sciences*, 7(12), pp. 1–16, 2017.

[3] Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T., "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, 61(1/2), pp. 17–28, 2013.

[4] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., "The CIPIC HRTF Database," *Proc. 2001 IEEE WASPAA*, pp. 99–102, 2001.

[5] Bauer, B. B., "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc.*, 9(2), pp. 148–151, 1961.

[6] Schroeder, M. R., "Digital simulation of sound transmission in reverberant spaces," *J. Acoust. Soc. Am.*, 47(2), pp. 424–431, 1970.

[7] Kirkeby, O. and Nelson, P., "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc.*, 47(7/8), pp. 583–595, 1999.

[8] Norcross, S. G. and Bouchard, M., "Multichannel inverse filtering with minimal-phase regularization," *Proc. 123rd AES Conv.*, pp. 1–8, 2007.

[9] Parodi, Y. L., *A systematic study of binaural reproduction systems through loudspeakers: A multiple stereo-dipole approach*, Ph.D. thesis, Aalborg University, Denmark, 2010.

[10] Majdak, P., Masiero, B., and Fels, J., "Sound localization in individualized and non-individualized crosstalk cancellation systems," *J. Acoust. Soc. Am*, 133(4), pp. 2055–2068, 2013.

[11] Akeroyd, M. A., Chambers, J., Bullock, D., Palmer, A. R., Summerfield, A. Q., Nelson, P. A., and Gatehouse, S., "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics," *J. Acoust. Soc. Am*, 121(2), pp. 1056–1069, 2007.

[12] Blauert, J., "Sound localization in the median plane," *Acta Acust. united Ac.*, 22(4), pp. 205–213, 1969.

[13] Hebrank, J. and Wright, D., "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am*, 56(6), pp. 1829–1834, 1974.