V. Grimaldi, L.S.R. Simon, G. Courtois and H. Lissek,
"Effects of Head-Tracking Artefacts on Externalization and
Localization in Azimuth With Binaural Wearable Devices"
J. Audio Eng. Soc., vol. 71, no. 10, pp. 650–663, (2023 October).
DOI: https://doi.org/10.17743/jaes.2022.0097

Effects of Head-Tracking Artefacts on Externalization and Localization in Azimuth With Binaural Wearable Devices

VINCENT GRIMALDI,[1] LAURENT S.R. SIMON,[2] GILLES COURTOIS,[2] AND
(vincent.grimaldi@epfl.ch) (laurent.simon@sonova.com) (gilles.courtois@sonova.com)
HERVÉ LISSEK,[1] *AES Member*
(herve.lissek@epfl.ch)

[1] *LTS2 - Groupe Acoustique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[2] *Sonova AG, Stäfä, Switzerland*

Head tracking combined with head movements have been shown to improve auditory externalization of a virtual sound source and contribute to the performance in localization. With certain technically constrained head-tracking algorithms, as can be found in wearable devices, artefacts can be encountered. Typical artefacts could consist of an estimation mismatch or a tracking latency. The experiments reported in this article aim to evaluate the effect of such artefacts on the spatial perception of a non-individualized binaural synthesis algorithm. The first experiment focused on auditory externalization of a frontal source while the listener was performing a large head movement. The results showed that a degraded head tracking combined with head movement yields a higher degree of externalization compared to head movements with no head tracking. This suggests that the listeners could still take advantage of spatial cues provided by the head movement. The second experiment consisted of a localization task in azimuth with the same simulated head-tracking artefacts. The results showed that a large latency (400 ms) did not affect the ability of the listeners to locate virtual sound sources compared to a reference headtracking. However, the estimation mismatch artefact reduced the localization performance in azimuth.

0 INTRODUCTION

Head tracking combined with head movements enables dynamic binaural synthesis of virtual sounds. By retrieving the listener's head movements and position with a head-tracking device, binaural room impulse responses (BRIRs) can be selected accordingly while achieving real time convolution. This is necessary to render a realistic and plausible auditory experience of sound sources for which the position remains valid regardless of the listener's orientation. This dynamic sound reproduction can help resolve front-back confusions [1] and thus improves the localization performance. Moreover several studies have shown that head tracking combined with head movements can improve the perception of externalization [2–4].

A head-tracking algorithm based on two three-axis accelerometers was developed in [5]. The algorithm design was constrained by the limitations of the used wearable hearing devices. Thus, it could not rely on gyroscopes or magnetometers that are conventionally used to ensure a robust estimation of the head orientation in azimuth. Consequently, tracking artefacts were reported with this algorithm. The goal of this article is to evaluate the effects of such artefacts on the perception of a non-individualized binaural synthesis algorithm designed for wearable binaural communication devices (e.g., hearing devices or hearables). The present study focused on auditory externalization and localization in azimuth.

0.1 Head Tracking and Auditory Externalization

In the second experiment reported in [2], listeners had to either keep their head stationary or rotate their head between –15° and 15°, while head tracking could be active or not. The stimuli were generated using individualized BRIRs and "head-absent" IRs measured with two microphones on a stereo bar. Additional stimuli were obtained using linear interpolation of the two types of IRs to generate hybrid impulse responses. The results revealed that head tracking combined with head movements could improve externaliza-

tion in the case of "head-absent" IRs, suggesting that head tracking brings potentially more improvement in the case of non-individualized binaural synthesis. In the case of individualized BRIRs, head movement without head tracking drastically reduced the perceived externalization. However, the combination of head tracking with head movements did not significantly improve externalization compared to conditions with no head movements. Indeed, externalization is already high with individualized BRIRs, leaving little room for improvement from head movement.

Hendrickx et al. [3] found that head movements combined with head tracking could significantly improve externalization with non-individualized binaural synthesis, for frontal and rear sources in particular. The protocol required participants to make controlled head movements for every stimulus, and allowed more time to complete the movement compared to previous studies, i.e., 6.5 s vs. less than 3 s. The participants rated externalization while they remained static after completing the movement. This was to ensure that potential improvements provided by the head movement was not only because of the lateralization, and remained after completion of the movement. It was found that head tracking in combination with head movements substantially increased externalization compared to the conditions for which the listeners did not move their head. Head tracking combined with head movements also enhanced externalization for lateral and frontal sources compared to conditions with head movements but no head tracking, as found in [2].

A similar protocol was used in [4] with frontal sources, but the experiment additionally included a condition with virtual source movements with various trajectories while the head remained stationary. The main conclusion was that for both source and head movements, only large movements increased the perception of externalization, whereas small movements did not affect it.

In [6], non-individual BRIRs were measured in a listening room and truncated to different lengths (from 2.5 to 120 ms). Speech and music signals were convolved with those BRIRs, and the resulting binaural signals were presented over headphones to eight participants, who did or did not perform a large head motion. The results suggest that the improvement in perceived externalization with head movements was smaller for longer BRIR lengths. The study concluded that head movements combined with head tracking can significantly improve the perceived externalization for virtual sound sources synthesized with short BRIRs, for frontal sound sources in particular. In their study, this corresponded to BRIRs that were shorter than 10 ms. For longer BRIRs, they found that head movements had no influence on the perceived externalization of virtual sound sources. Such stimuli may indeed already be well externalized.

A few earlier studies suggested that head tracking combined with head movements might have a weak effect or no effect at all on the perception of externalization [7, 8]. Nevertheless, in [8] lateral and frontal azimuths were mixed in the analysis; hence, the potential improvement brought by head tracking for frontal sources could not be observed. Moreover, the head movement duration was probably too short to allow the listener to take advantage of them, as suggested by the authors. The same limitation can be mentioned in [7] where only a small effect of head-movement on externalization was found.

0.2 Head Tracking and Localization

Wallach [1] suggested that dynamic cues associated with head movements, such as ITDs and ILDs, were necessary to resolve front-back confusions, in the auditory localization of a sound source. Fixing the head of a listener has been shown to yield a large increase in front-back confusions [9]. Multiple studies showed that when listeners are free to move their head, they are more accurate at localizing a sound source than when their head is fixed or constrained, mostly owing to the front-back resolution [10–13]. In [8], head movements combined with head tracking improved localization performance of virtual sources compared to static rendering. A larger improvement was found with non-individualized head-related transfer functions (HRTFs).

0.3 Impact of Head-Tracking Artefacts on Spatial Perception

In the literature, the studies assessing the effect of head-tracking artefacts on the perception of binaural spatialization mainly focused on the effect of tracking latency. In [14], the impact of head-tracking latency on the localization of broadband sounds was first investigated. An increase of localization errors was found for brief sounds with latencies larger than 70 ms. An increase in the time required to locate a continuous sound source was found with latencies larger than 90 ms. The results suggest that head-tracking latencies lower than 60 ms might be acceptable for most virtual audio applications. In a subsequent study, they found that some listeners were able to detect latencies of 60–70 ms for isolated sounds. In a second part of this experiment, the delayed target sounds were presented in conjunction with a reference tone with minimal possible latency, that was co-located with the virtual sound source. In this case, their detection thresholds were approximately 25 ms lower. Hence, the results suggest that a latency of 30 ms or less should be difficult to detect even in complex virtual auditory scenarios.

In [15], only a single source was used, and the values for detection of latency had a mean and standard deviation of about 100 ms and 30 ms respectively (pooled threshold values), and the minimum detected latency was around 50 ms. The nature of the stimulus, as well as the reverberant vs. anechoic condition did not affect those results. In [16], the stimuli consisted of either a single frontal sound source or a complex sound scene including five sources. The study aimed at investigating to which extent the spatial stability of sound sources in binaural reproduction was influenced by head-tracking latency. The results suggest that with an increase in latency, the source instability was more audible with the single source. In this case, the threshold was 10 ms lower compared to the complex sound scene.

The effect of head-tracking artefacts on localization performance was investigated in [17] with an anechoic binaural spatialization. It was found that the average localization

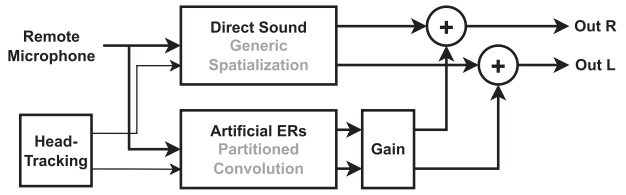


Fig. 1. Simplified block diagram of the non-individualized spatialization algorithm used in this experiment.

accuracy was not significantly degraded until the system latency reached a threshold of 96 ms and the tracking update rate was decreased down to a threshold of 10 Hz. The evaluation of the effect of head-tracking latency on externalization was addressed in [18]. The author found that latency up to 500 ms did not affect externalization. However, the study was conducted with only five participants.

0.4 Goal of the Study

The present study evaluates to what extent head-tracking artefacts potentially happening in scaled-down head-tracking algorithms, might affect spatial perception. In particular, this study describes two experiments, which aim to assess the effect of those artefacts on auditory externalization and performance in localization respectively. The binaural rendering used in this experiment is also adapted to the constraints of wearable hearing devices, with the use of non-individualized and low-computational cost spatialization algorithms. The results were expected to assess the advantages provided by the head-tracking algorithm described in [5] for binaural synthesis in the context of remote microphone systems for wearable devices.

1 SPATIALIZATION ALGORITHM AND SIMULATED ARTEFACTS

1.1 Binaural Synthesis Algorithm

The algorithm used in this experiment is designed to spatialize a remote microphone signal in the context of wearable communication devices. The algorithm, described in [19], aims at improving auditory externalization by introducing early reflections (ERs) in the remote microphone signal. It consists in superimposing synthesized non-individual ERs to a direct sound generated with anechoic generic HRTFs as depicted in Fig. 1.

The spatialized direct sound is generated using non-individual HRTFs approximated by a minimum-phase filter and pure time delays for the ITD as described in more detail in [20]. This method enables to use linear interpolation for intermediate azimuths. The HRTF database was measured with a KEMAR manikin in the anechoic room of Ecole Polytechnique Fédérale de Lausanne, with a resolution of 10°. The ERs are synthesized in real-time using a uniform partitioned convolution algorithm [21], as implemented in [22]. The pairs of BRIRs used for this experiment were measured with a step of 10° and a distance of 2 m. The BRIRs were measured in the room of the experiment (RT_{60} = 0.17 *s*), using a KEMAR manikin. The head-tracking device enables to retrieve the yaw orientation of the head. The angle is used to select the correct minimum-phase filter and pure delay values for the direct sound, and the correct pair of BRIRs for the artificial ERs. The independent gain between the direct sound and the ERs allows to achieve any desired direct-to-reverberant ratio (DRR).

1.2 Head-Tracking Conditions

Three main head-tracking conditions were used in the experiments: a reference and two different types of artefacts. These artefacts were simulated by artificially degrading a reference tracking in real-time, in order to obtain replicable artefacts across stimuli and subjects.

1.2.1 Reference

When all the necessary sensors, i.e., gyroscopes, accelerometers and magnetometers are available, a reliable and accurate estimation of the head yaw position can be achieved. For this, a low-latency attitude and heading reference system (AHRS) device (NGIMU[1]) was used to obtain a reference estimation. The total tracking latency was estimated to an average of 42 ms ($\sigma = 7$ ms).

1.2.2 Latency

The first type of artefact consists of a simple delay compared to the reference. When listening to a sound source in real life, there is no perceivable latency between the movements of the listener and the consequent changes in the sound reaching their eardrums. However, most virtual display systems introduce a certain amount of delay due to the inherent latency of the tracking device itself, the communication delay between that device and the audio display, the selection of the appropriate HRTFs, and the subsequent audio processing. In this study, the latency was simulated by applying a delay on the reference estimation. After informal pre-test sessions, it was decided to test a single and large latency value of 400 ms in the experiments. An example measurement of the reference and the related latency simulation is depicted for a simple motion in Fig. 2(a).

1.2.3 Yaw Estimation Mismatch

A head-tracking algorithm relying solely on the use of two three-axis accelerometers was developed in [5]. One of the main limitations of this algorithm comes from the difficulty in tracking slower parts of the motion, as the low values of accelerations have amplitudes too small to be distinguished from the noise of the accelerometers. Hence, the algorithm relies on freezing the computation when such values of accelerations are reached to avoid drifting associated with the integration of noise. This results in parts of the motion being missed when reaching final positions, and an underestimation of the absolute values of the yaw estimation.

Using the corresponding prototype with the embedded accelerometers and the developed algorithm would lead to unpredictable variations in the artefacts experienced by the

[1] https://x-io.co.uk/ngimu/.

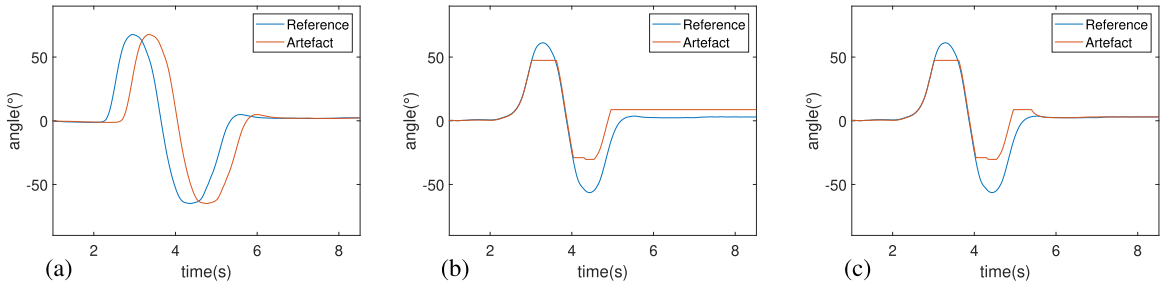


Fig. 2. (a) Example of the simulation of the latency artefact with a delay of 400 ms. (b) Example of the simulation of the yaw estimation mismatch. (c) Example of the simulation of the yaw estimation mismatch adapted to the externalization experiment.

listeners. Thus, it was decided to simulate a degraded estimation yaw_e mimicking the behavior of the developed algorithm by alteration of the reference tracking yaw_r obtained with the AHRS device. A method yielding similar results was implemented using the computation described in APPENDIX A. The value of γ_v was set empirically to 0.8, which leads to similar range of errors as the algorithm in [5]. An example of the resulting artefact estimation is depicted in Fig. 2(b).

2 PARTICIPANTS

Thirty naive listeners took part in both experiments (15 female, 15 male, average age = 23.9 years). All the listeners were self-reported as having normal hearing. For each participant, the two experiments were performed on different days, in order to limit the influence of listening fatigue. Half of the participants started with the externalization experiment, while the other half started with the localization experiment. A minimum of one week between the two experiments was ensured for each participant, to limit the effect of training with the non-individualized HRTFs and BRIRs used in the binaural synthesis. The gap between the two sessions was on average 2.1 weeks ($\sigma = 2.2$ weeks).

3 EXTERNALIZATION EXPERIMENT: PROTOCOL

3.1 Setup

The experimental setup was installed in a listening room (volume = 125 m^3, RT_{60} = 0.17 s). All stimuli were played through a pair of open headphones (Audeze LCD-2C) driven by a headphones amplifier (Lake People HPA RS 02). A low latency audio interface (RME Babyface Pro Fs) was used to play back the sounds from the Simulink implementation of the spatialization algorithm. The AHRS device was mounted on top of the headphones in order to track the head motions. The sampling rate was set to 50 Hz for the data acquisition from the device to a dedicated Simulink model. Data transmission was achieved with USB COM Port communication. The sound pressure level was adjusted to 65 dBA.

3.2 Preparation/Training

Prior to the actual experiment, the listeners used an interface on which they could play and listen to two versions of a pre-recorded audio file. The first version was 70 s of speech (male voice) moving around the listener, recorded using a pair of binaural microphones placed on the artificial ears of a KEMAR manikin. This recording was made in the same room as the one of the experiment. The second version was a binaural-to-monophonic reduction of this recording.

The first goal of this preparation was to ensure that they understood and could perceive auditory externalization. The second goal was to give them a reference of what could be a well externalized sound. Additionally, the recording aimed to highlight the subtle difference between auditory externalization and auditory distance. Because both percepts share at least partially the same continuum [23], and because DRR, which varied during our experiment, has an effect on auditory distance perception [24], it was important to make sure participants would be rating the correct percept. For this purpose, at some point of the audio file, the voice was moving from close of the ear of the manikin to the corner of the room. The goal was to showcase that a variation of the distance could still be perceived in the diotic recording, and the sound remained perceived as internalized. The experimenter therefore gave further explanations about the distinction between auditory externalization and distance, which appeared clear for all participants.

3.3 Stimuli and Conditions

The base stimulus consisted of a 8.5-s excerpt from an anechoic male speech recording in English. The same sentence was used for every run. The first samples of the BRIRs (corresponding to the direct sound) were set to zeros to avoid the superimposition with the direct sound spatialized using generic HRTFs. The BRIRs were truncated to 10 ms for every stimulus. This value was chosen as the constraints of wearable devices suggest to put limits on the memory usage, and thus it is of practical interest to investigate the possibility to provide externalization with short BRIRs. The value was determined during informal pre-tests in which the ER time was not found to be the most influential parameter. Moreover this value was mentioned in [6] as a threshold below which head tracking might be more beneficial for externalization. For this part of the experiment, a fourth condition with no head tracking was added. The different head-tracking conditions ("no head-tracking," "reference," "latency of 0.4 s," and "yaw estimation mismatch") are denoted respectively: No HT, Ref, Lat 400, and Yaw Mis.

The effect of the DRR was tested with the levels: 5, 8, or 30 dB (computed with the 10-ms truncation). The output was level-normalized in the real-time model with a gain at the output to obtain a consistent level across every DRR setting. Taking into account the hearing device application, a low-pass filtering with a cut-off frequency of either 7 or 10 kHz was applied. The laterality of the sound source has a substantial effect on the perception of externalization [25]. All stimuli evaluated in this experiment were simulated in front of the listener at an azimuth of 0°.

A full factorial design was used for this experiment, each subject rated every combination of the conditions described in this section, i.e., 4 head-tracking conditions × 3 DDR levels × 2 cut-off frequencies. Each of these 24 combinations was presented four times for every participant, and the order was randomized inside each of the four repetition blocks. Hence, with the 18 additional training stimuli, each participant had to evaluate a total of 114 stimuli. The experiment lasted about 60 min. After half of the stimuli had been evaluated, a break was imposed to mitigate the effect of listening fatigue and maintain focus of the participant.

3.4 Procedure

The subject was asked to perform the same motion for every stimulus. First they looked in front of them at a mark located on the wall at 0°. Then they pressed the "play" button on the interface. As soon as the speech was heard they were asked to turn their head, and look at a mark located on the wall at +80° first, then turn the head to the other side to look at a mark located at –80°, and finally point back at the initial position at 0°.

They had to perform this motion in synchronization with the words of the speech sample. They were asked to be back at the center position for a specific word in the sentence. This was to ensure that they heard the last 2 s of the speech stimulus as a source in front of them while they remained still. This protocol was inspired by the one used in [3]. It ensures that the potentially observed increase in externalization is not based on the lateralization of the sound sources while the listener is turning the head away from it. Instead, it aims at assessing if the cues provided by the movement of the listener yield a persistent impression of externalization after the movement.

Frontal sources are known to be perceived as less externalized, and were shown to potentially benefit more from the addition of head tracking in comparison to lateral sources [3]. After performing the instructed movement in this experiment, the yaw estimation mismatch artefact can lead to a final angle that is different from 0°. To prevent this from biasing the results, the artefact simulation was adapted for this part of the experiment. At a certain time t = 5.5 s, the tracking simulation was set to smoothly converge to fit the reference tracking. This was to ensure that, after completing the movement, the final angle was not larger for the Yaw Mis condition compared to the other conditions. A short silence was included in the recording from $t = 5.5$ s, so that this compensation did not affect the spatial processing. An example of the resulting tracking is depicted in Fig. 2(c) derived from the original artefact simulation pictured in Fig. 2(b).

Understanding how to perform this motion in synchronization with the speech sample took between three and eight runs for the participants. The first 18 runs of the experiment served only as training runs and were not included in the results. This number was defined during informal pre-test sessions. Trajectories were recorded, and the experimenter continuously monitored that correct synchronization of the motion was achieved.

Listeners were asked to rate the degree of externalization perceived at the end of the motion, for the last 2 s of the speech stimulus when they were back at the center and remained still. A diotic 2-s pink noise sample was played between every stimulus. This was intended to limit the effect of the order of presentation, so that the spatial attributes of the previous stimulus would not influence the perception of the next one.

To rate the externalization, the listeners had to use a continuous scale labeled at the extremities from "completely internalized" to "completely externalized," which corresponded to ratings of 0% and 100% respectively. The listeners had to move the cursor from an initial 50% position to give their rating and then validate it with a dedicated button. A similar scale was used to rate externalization in [26]. This method was preferred over categorical scales such as the ones used in [3, 4, 27], or scales with visual references used in auditory distance estimation experiments [28] as no auditory or visual reference was available to the listener. Moreover, the design of the experiment did not allow to use MUSHRA-style interface or paired comparisons, as it would require too much memory effort for the listener to remember accurately the perceived externalization of the previous stimuli after completing the movement of the next ones.

In [26], the listeners had to close their eyes. The purpose was to enable the listener to rate externalization without being constrained by auditory distance matching with visual references. For the present study, listeners kept their eyes open, as it was necessary to ensure that they were accurately at 0° after the movement. With available visual cues, it is possible that a stimulus for which the binaural cues for externalization are preserved, but the cues associated with perceived distance are distorted (e.g., the DRR) might never be given a 100% externalization rating, even if it is neatly perceived outside the head. This effect might have been partially limited in the present study as lights were dimmed and no potential visual reference to match was proposed to the listener.

4 EXTERNALIZATION EXPERIMENT: RESULTS

4.1 Auditory Externalization

The raw results from this experiment naturally exhibit significant standard deviation differences from one subject to another as each subject can use the scale in a different manner. In order to normalize the results, these ratings were transformed into z-scores [29]. The z-score was cal-

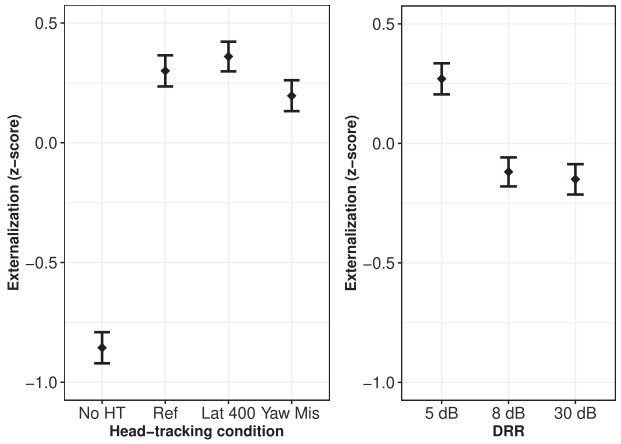


Fig. 3. Results of the externalization ratings, with z-score transformation: (a) as a function of the head-tracking condition (averaged across DRR, repetitions and cut-off frequency) and (b) as a function of the DRR level (averaged across head-tracking condition, repetitions and cut-off frequency). The diamond corresponds to the mean, and the bottom and top lines represent the associated 95% confidence intervals.

culated for each subject by subtracting the mean from each raw rating and then dividing the difference by the standard deviation of this subject's ratings. With this transform, no conclusion can be made on the absolute externalization ratings, but only on the relative ratings between conditions. To use this transformation, it is assumed that the externalization scale can be considered as interval-level. Indeed, the training aimed at ensuring that externalization was evaluated independently from auditory distance. Thus, the scale should not have been confused with one associated with auditory distance in rooms which is usually considered to be non-linear [30, 31]. The z-score ratings are displayed in Fig. 3.

A repeated-measures ANOVA was conducted with four within-subject factors: the head-tracking condition, DRR, low-pass filter cut-off frequency, and the repetition number. The assumption for sphericity was checked using Mauchly's test, and the Greenhouse-Geisser correction was applied when necessary. The normal distribution was visually checked by plotting the QQ-plots of the residuals for each of the independent variables. For large sample sizes as in this experiment, this method is recommended over evaluations such as the Shapiro-Wilk test [32].

Significant effects were found for the head-tracking condition [$F(3, 168) = 112.17$; $p < 0.001$] and the DRR [$F(2, 112) = 73.55$; $p < 0.001$]. Following Cohen's rule of thumb [33], a large effect size was found for both the head-tracking condition ($\eta^2_{partial} = 0.58$) and the DRR ($\eta^2_{partial} = 0.17$). No significant effect was found for the cut-off frequency [$F(1, 56) = 0.929$; $p = 0.339$] and the repetition number [$F(3, 168) = 2.83$; $p = 0.066$] (no significant interaction was found either), so the score for the two cutoff frequencies and all four repetitions were mixed for the rest of the analysis and on the plot in Fig. 3. No interaction was found between the head-tracking condition and the DRR value [$F(6, 336) = 0.206$; $p = 0.97$]. No three-way or four-way interaction was found either.

A post hoc Tukey's HSD analysis was conducted with a 95% confidence level for the two significant independent variables, i.e., the head-tracking condition and the DRR. It was found that the No HT condition was significantly less externalized than the Ref ($p < 0.001$), the Lat 400 condition ($p < 0.001$) and the Yaw Mis condition ($p < 0.001$). No significant difference was found between the Ref and Lat 400 conditions ($p = 0.54$) and the Ref and the Yaw Mis conditions ($p = 0.099$). Finally, with a small but significant difference, the Lat 400 was perceived as more externalized compared to the Yaw Mis ($p = 0.017$).

The stimuli with 5-dB DRR were perceived as more externalized compared to both the 8-dB DRR ($p < 0.001$) and 30-dB DRR ($p < 0.001$) stimuli. However, no significant difference was found between the 8-dB and 30-dB DRR stimuli ($p = 0.070$).

4.2 Head Movements

The head trajectories were recorded during the experiment for two purposes. The first was to verify that the participants could follow the instructions precisely enough, especially in terms of synchronization with the speech sample. The second was to assess if certain parameters of the head motion had an influence on the externalization perception. The following dependent variables were computed from the trajectory recordings: the maximum (right) and minimum (left) azimuth angles reached during the motion, the total amplitude of the motion, the final angle reached at the end of the motion (averaged over the last second) and the time spent at azimuth 0° at the end of the motion. The medians and interquartile ranges of the global data for those variables are summarized in Table 1.

These results suggest that the participants successfully managed to achieve the task as instructed by the experimenter. As can be seen in Table 1, no correlation was found between any of the dependent variables linked with the amplitude of the movement and the perceived externalization (negligible Spearman's ρ). This suggests that the variability and range in terms of amplitude of the movement in azimuth was small enough not to affect the externalization ratings. No correlation was found either with the final azimuth angle, which remained small.

Finally the time left after movement, i.e., the time when the participant was static and had to evaluate externalization, was not correlated with externalization ratings. The participants very rarely arrived too late (i.e., after the silence gap before the last part of the sentence), so no bias was created by participants failing to synchronize with the motion. The 0.5-s silence gap before the last part of the speech (on which externalization was evaluated) should be enough to compensate for the latency in the Lat 400 condition. However, this makes this condition potentially more sensitive to the ability of the participant to synchronize with the speech. Nevertheless, no correlation was found for the subgroup of the Lat 400 condition either (Spearman's ρ = 0.038; $p = 0.322$), suggesting that the silence gap was

Table 1. Medians, interquartile ranges and Spearman's ρ with p value in relation to the z-score of externalization for several dependent variables of the participants' movements.

	Median	Interquartile Range	Spearman's ρ (p value)
Maximum angle (right)	61.92°	12.74°	0.017 (0.375)
Minimum angle (left)	–62.18°	15.77°	0.019 (0.313)
Amplitude of the motion	124.24°	26.14°	0.021 (0.279)
Final azimuth angle	0.92°	2.52°	0.061 (0.001)
Time left after movement	2.23 s	0.28 s	0.017 (0.367)

probably enough to compensate for the latency and that it did not affect the perception of externalization for this condition.

The errors of tracking simulated in the Yaw Mis condition can vary from run to run, depending on the movement of the listeners. A score of distance to the related reference tracking was computed for every run in the Yaw Mis condition by summing differences between the two trajectories sample by sample. No correlation was found between the distance to the reference and the externalization ratings (Spearman's $\rho = 0.006$; $p = 0.891$).

5 EXTERNALIZATION EXPERIMENT: DISCUSSION

5.1 Effect of the Head-Tracking Condition

In this experiment, in which large movements were performed and only frontal azimuth sources were simulated, an increase in the perceived externalization was obtained for all conditions that included head-tracking. This is in agreement with [3] and [4], in which it has been shown that head-tracking combined with head movements provided an advantage for the perception of externalization if sources were frontal and movements were large.

In every condition where head-tracking was available, even with latency or yaw estimation mismatch, the additional cues provided to the listener by the dynamic reproduction always helped the listeners better externalize compared to the situation where no head-tracking was available. Additionally, in particular, in the Yaw Mis condition, the amount of error did not cause a reduction in the perception of externalization compared to the Ref tracking. This could be explained as the listeners still benefited from additional cues made available during motion, i.e., when listening with several pairs of HRTFs and BRIRs corresponding to various directions. As no correlation was found between externalization ratings and the amplitude and maximum values in azimuth reached, it is likely that the underestimation of the tracking usually happening in the Yaw Mis condition did not prevent listeners from externalizing either. Concerning latency in particular, this confirms the findings in [18], in which latency up to 500 ms did not have an influence on externalization.

It was clearly stated to the participant that the simulated speech source was in front of them at an azimuth of 0° (indicated by a mark on the wall). Additionally, they were informed that the sound source was not supposed to move. It is likely that this was not enough to affect plausibility, which can be degraded in the case of non-matching audio-visual presentations. Indeed, the perceived "slewing" of the sound source due to the tracking artefacts might have been perceived by the auditory system as if the source was moving, regardless of the instruction. It was shown in the literature that source movements can increase the perception of externalization [34, 6], which would support this hypothesis. In this case, there is no reason to think that the considered artefacts should have affected externalization significantly. Indeed, if the sound source was presented with a visual reference, plausibility and consequently externalization might be affected by tracking artefacts.

This could be tested, e.g., with a real person or a video of a speaker displayed on a screen placed in the direction the sound is supposed to be coming from. However, visual capture might also help to externalize. The same remarks can be made for the Lat 400 stimuli. This should be tested in further studies.

5.2 Effect of the ERs and Binaural Cues

ERs play an essential role in the perception of auditory externalization and the length of the impulse response affects externalization [35], with a larger influence up to approximately 30–40 ms [36]. A recent study suggested that head-tracking and head movements might not provide a substantial increase in externalization for longer BRIRs (already well externalized without tracking) compared to shorter BRIRs [6]. The present study investigated the possibility to increase externalization with 10-ms truncated non-individualized BRIRs measured in a listening room, superimposed to a generic direct sound. In [3], the spatialization was non-individualized, and the authors intentionally chose a room with "not too much reverberation" to record their stimuli. The full-length reverberation was used in the latter, for a room having a $RT_{60} = 0.24$ s, i.e., slightly larger but comparable to the listening room of the present study ($RT_{60} = 0.17\ s$). Conversely, no reverberation was available in [4], but generic HRTFs were also used.

A clear effect of the DRR on the perception of externalization was observed in the present study. The 5-dB DRR stimuli were perceived as significantly more externalized compared to both the 8-dB and 30-dB DRR stimuli. Improved externalization in the 5-dB DRR condition is expected to be caused by the higher level of early reflections than in the two other DRR case. However, because of intricate relation between externalization and auditory distance perception and because of the DRR being a known auditory distance perception cue [24], it is possible that participants were influenced by a farther perceived distance in this condition.

The results suggest that the cues derived during a large movement were sufficient to provide significantly more externalization, even though the spatialization was made using generic HRTFs for the direct sound and generic 10-ms truncated BRIRs for the ERs. The results reported in [2] are in agreement and additionally suggest that with non-individualized HRTFs, the benefit brought by head tracking can be larger than with individualized HRTFs. Externalization is indeed poorer in the case of non-individualized HRTFs, leaving more room for improvement. Nevertheless, individualized HRTFs were not tested in the experiment reported in this article.

Finally, the BRIRs used in this experiment were measured in the same listening room as the one in which the experiment took place. Room congruence has an important influence on auditory externalization [37]. As the BRIRs were truncated to 10 ms, they cannot be considered to be exactly congruent. It can be hypothesized that head tracking provides more improvement in externalization in such situation, i.e., when the initial conditions are more challenging for externalization. Further studies could investigate the potential improvement in externalization provided by head movements, depending on the divergence between the playback room and the room in which BRIRs are measured.

6 LOCALIZATION EXPERIMENT: PROTOCOL

6.1 Setup

The experiment took place in the same room and with the same setup as the externalization experiment. A subset of the stimuli (real sources) were played through eight loudspeakers (ELAC 301.2) located around the listener and amplified using an eight-channel amplifier (Allen & Heath GR8A). The loudspeakers were positioned on a rectangular frame present in the listening room, and hidden behind a black curtain. The transfer function of the curtains was measured in an anechoic room, and all sounds played through the loudspeakers were compensated for the subsequent attenuation in the high frequencies. Additionally, every stimulus played through the loudspeakers was compensated with a gain corresponding to the position of each loudspeaker, so that the level was constant and equal to 65 dBA at the listener's position. This served to minimize the effect of the level cue which might have biased the azimuth localization because of the visible rectangular shape of the frame. The listeners did not wear headphones in this first part.

The rest of the stimuli (virtual sources) were played through the same pair of open headphones and amplifier as in the first experiment. The target azimuths were the same as for the real sources. Moreover, the ERs of the virtual sources presented over headphones were simulated with BRIRs with a constant distance from the listener. The same complete AHRS device as in the first experiment was mounted on the top of the headphones. Additionally, a laser pointer was placed on top of the head-tracker in coincidence with its x axis and was activated during this experiment. The laser beam was used by the listeners who could aim at the perceived location of the sound sources by pointing the head toward it. A schematic representation of the setup is depicted in Fig. 4.

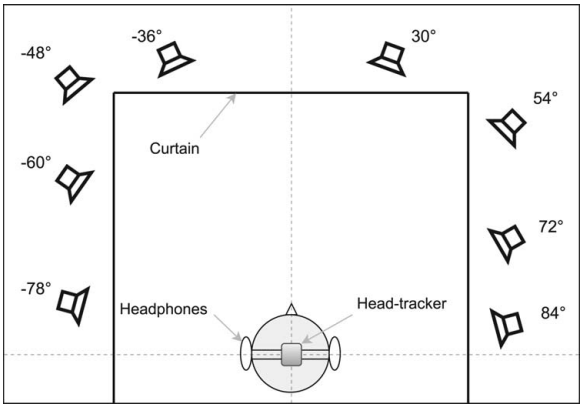


Fig. 4. Schematic representation of the setup used for the localization experiment, with eight loudspeakers hidden behind black acoustically transparent curtains, and the listener equipped with headphones and the AHRS head-tracking device.

6.2 Stimuli

The recording used to generate every stimulus consisted of a 25-s excerpt from an anechoic male speech recording in English. For the virtual sources, the BRIRs were still truncated to 10 ms. The cutoff of the low-pass filter was fixed to 10 kHz and the DRR was fixed to 8 dB.

The real sources were presented in a first randomized block. Then, for the virtual sources, the three following head-tracking conditions were considered and randomized: Ref, Lat 400, and Yaw Mis. Each combination of head-tracking condition and target azimuth was presented with four repetitions. Four initial stimuli were used as training runs at the start of each of both the real and virtual sources sub-parts of the experiment. Hence, the subject had to locate a total of 36 stimuli for the real sources part (8 target azimuths $\times$ 4 repetitions + 4 training stimuli) and 100 stimuli in the virtual sources part (3 head-tracking conditions $\times$ 8 target azimuths $\times$ 4 repetitions + 4 training stimuli). The total duration of the experiment was between 45 and 60 min.

6.3 Procedure

First, the listeners were asked to point toward a mark at azimuth 0°, in front of them. The laser helped them to achieve this precisely and easily. This served to initialize the head tracking between each stimulus. Then, they could press the space bar on the computer keyboard to play the stimulus. As soon as the stimulus could be heard, they were instructed to locate the sound source, and point the head toward the perceived azimuth using the laser placed on their head. It was suggested in [38] that this method is one of the most reliable for such a task. They could then validate their answer by pressing the space bar again. The instruction by the experimenter stated that accuracy was the priority in this task, but mentioned that the time was measured too, and that they should validate right away when they were sure of the answer.

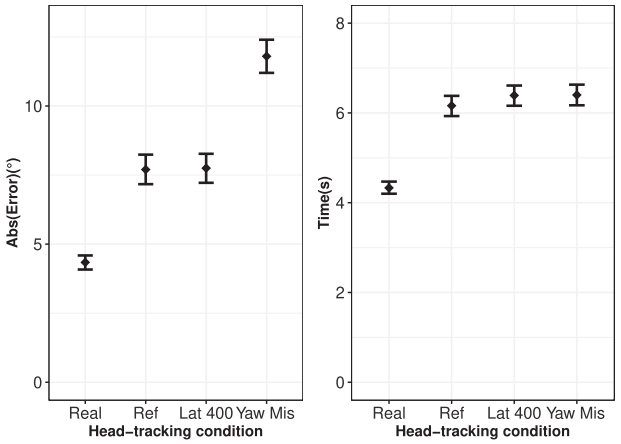


Fig. 5. (a) Absolute error of localization for each head-tracking condition. (b) Localization time for each head-tracking condition. The diamond corresponds to the mean, and the bottom and top lines represent the 95% confidence intervals. The repetitions are pooled together.

After validation, a message appeared on the screen to remind the listener to point the laser at the mark in front of them to initialize the head-tracker before playing the next stimulus. As in the externalization experiment, a diotic 2-s pink noise sample was played after each validation to limit the effect of the order of presentation for the virtual sources. Addition, the experimenter informed the participant that the stimuli could only come from frontal positions.

The listeners were sitting on a non-rotating chair with a fixed position. The azimuth positioning of the loudspeakers was measured from the point where the listeners' head was intended to be during the experiment. It is estimated that the position of the head of the listeners compared to its expected position may vary only in a range of about ± 10 cm maximum. This results in a potential uncertainty of $\pm 2^\circ$ in the answers collected in the case of the real sources.

7 LOCALIZATION EXPERIMENT: RESULTS

7.1 Localization Error

Absolute localization errors for each head-tracking condition are reported in Fig. 5(a). As the real sources were not randomized in the same presentation block, the results of the localization of real sources are only presented as an indicative value of the performance of the listeners. Thus, they were not mixed together in the statistical analysis.

The absolute error data was transformed using a cubic root transformation in order to obtain a normal distribution of the data, which was visually checked using QQ-plots of the residuals for each condition. All the statistical analysis was conducted on the transformed data.

7.1.1 Real Sources

The results indicate a mean absolute error of 4.34° with a standard deviation of 4.07°. The results were not included in the ANOVA analysis as this condition was tested in a single block before the virtual sources. This condition gives an idea of the baseline localization performance of the listeners.

7.1.2 Virtual Sources

The assumption for sphericity was checked using Mauchly's test, which showed that sphericity had not been violated for any of the independent variables. A repeated-measures ANOVA was conducted on the transformed data for the virtual sources, and a significant effect was found for the head-tracking condition [$F(2, 58) = 27.25$; $p < 0.001$]. Following Cohen's rule of thumb [33], a medium effect size was found for the head-tracking condition ($\eta^2_{partial} = 0.10$). No effect was found for the repetition number [$F(3, 87) = 0.10$; $p = 0.96$]. This suggests that participants did not improve their performance by learning HRTFs along the experiment.

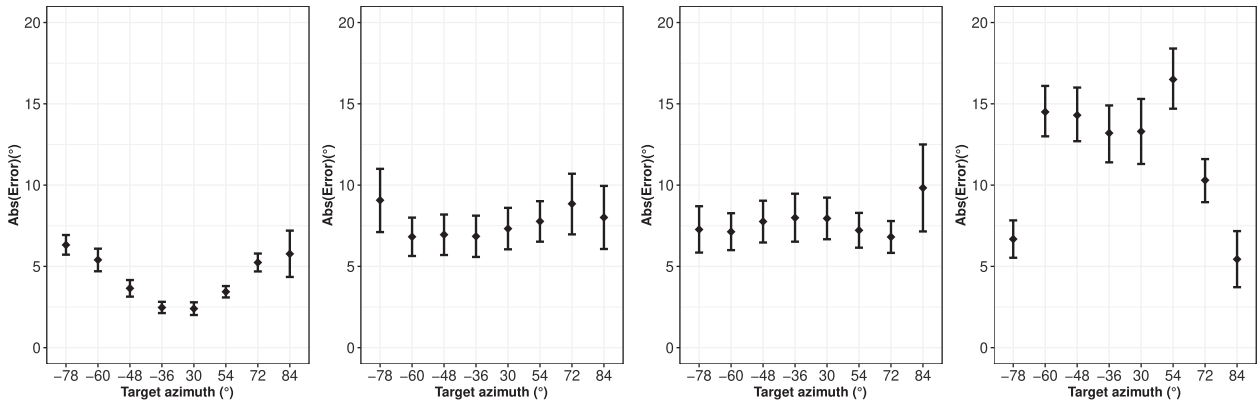
A post hoc Tukey's HSD analysis with a 95% confidence level was conducted on the head-tracking condition. It was found that the Yaw Mis condition ($\bar{x} = 11.8^\circ; \sigma = 9.83^\circ$) was located significantly less accurately compared to both the Ref condition ($p < 0.001$; $\bar{x} = 7.70^\circ; \sigma = 8.51^\circ$) and the Lat 400 condition ($p < 0.001$; $\bar{x} = 7.75^\circ; \sigma = 8.34^\circ$). On the contrary, the results suggest that the listeners did not perform differently between the Ref and Lat 400 conditions ($p = 0.88$).

The number of times the listener switched the sign of their head trajectories while they were trying to locate the sound source was computed out of the recorded trajectories. It could be expected that listeners might perform more movements when they have more difficulty to locate the sound. However, no correlation was found with the localization performance for the virtual sources (Spearman's $\rho = 0.033$; $p = 0.072$) nor for the Yaw Mis condition group (Spearman's $\rho = 0.011$; $p = 0.737$). Additionally, an error score compared to the reference tracking was computed from the trajectories for the Yaw Mis condition group, but the correlation with the localization performance was negligible (Spearman's $\rho = 0.14$; $p < 0.001$)

7.1.3 Effect of the Target Azimuth

When looking at the mixed data for the virtual sources, a significant effect was found for the target azimuth [$F(7, 203) = 3.65$; $p < 0.001$], with a small-medium effect size ($\eta^2_{partial} = 0.050$). However, as pictured in Fig. 6, results suggest it is more interesting to look at the effect of the azimuth for each head-tracking condition separately, as confirmed by the interaction between these variables [$F(14, 406) = 13.06$; $p < 0.001$].

In the case of the Real sources condition, a significant effect of the azimuth on the localization performance was found [$F(7, 203) = 14.41$; $p < 0.001$]. The p values of the post hoc Tukey's HSD analysis leads to too many combinations to be included here. A clear tendency can be observed as larger errors occur for more lateral sources compared to the more frontal sources. In the case of the Lat 400 condition, no significant effect of the azimuth was found [$F(7, 203) = 0.172$; $p = 0.991$]. No significant effect of



(a) Real sources (b) Reference (c) Latency - 400 ms (d) Yaw Mismatch

Fig. 6. Absolute error in localization, effect of the target azimuth for each head-tracking condition. The diamond corresponds to the mean, and the bottom and top lines represent the 95% confidence intervals. The repetitions are pooled together.

the azimuth was found either with the Ref head-tracking condition [$F(7, 203) = 0.829$; $p = 0.564$].

In the case of the Yaw Mis condition a significant effect of the target azimuth was found [$F(7, 203) = 19.11$; $p < 0.001$]. The post hoc Tukey's HSD analysis indicates that the performance for the very lateral angles, –78° and +84° was significantly better compared to the other more frontal azimuths. It can be hypothesized that this effect might be because of listeners being instructed that sources should only be coming from the frontal hemisphere, which may create a boundary effect, i.e., the distribution of errors close to the limit angles (±90°) is not symmetrical.

7.2 Localization Time

Localization time for each condition are reported in Fig. 5(b). The raw localization time data distribution is skewed. The data was transformed using a logarithmic transformation in order to obtain a normal distribution of the data, which was checked visually with QQ-plots of the residuals for each head-tracking condition and target azimuth. The assumption for sphericity was checked using Mauchly's test, which showed that sphericity had not been violated for any of the independent variables.

A repeated-measures ANOVA was conducted on the log-transformed time of answer, in the virtual sources subgroup. A significant effect was found for the head-tracking condition [$F(2, 58) = 6.01$; $p = 0.0043$], with a large effect size ($\eta^2_{partial} = 0.53$). The Ref condition was located faster than both the Lat 400 ($p = 0.0053$) and Yaw Mis ($p = 0.027$) conditions. No difference was found between the Lat 400 and Yaw Mis conditions ($p = 0.82$). Nevertheless, the general difference regarding the mean is modest.. This suggests that neither the latency nor the yaw estimation mismatch did affect extensively the time it took for the listeners to confidently locate the sound source. In the case of the Lat 400, the slightly longer time could be simply explained by the 400-ms delay.

No significant effect was found either regarding the target azimuth [$F(7, 203) = 2.28$; $p = 0.1967$]. This suggests that the difficulty was not increased for more lateral sources compared to frontal sources. Finally, a significant difference was found for the repetition number [$F(3, 87) = 7.70$; $p < 0.001$]. This suggests that the listeners performed the task slightly quicker over time during the session. Nevertheless a small effect size was associated with this observation ($\eta^2_{partial} = 0.011$). A post hoc Tukey's HSD analysis with a 95% confidence level was conducted and a significant difference was found between the first repetition and the three consecutive ones ($p = 0.033$, $p = 0.001$ and $p < 0.001$ compared to the second, third, and fourth repetition, respectively).

8 LOCALIZATION EXPERIMENT: DISCUSSION

The head-tracking latency did not decrease the performance in localization compared to a reference head tracking in this experiment.

It can be hypothesized that the subjects understood the nature of this type of artefact and were able to compensate for it. This is in agreement with the results in [18], in which it was found that localization was still accurate with a latency as large as 500-ms and 8-s stimuli. In the study conducted in [39], which used individualized HRTFs, accuracy was generally comparable for the shortest and longest latencies. The authors suggest that listeners might have been able to ignore latency while they were actively trying to locate the source, despite the spatial "slewing" of the source for large latencies. In the same study, the duration of the stimuli affected moderately the localization performance as a function of latency. The present study only used one setting of a large latency (400 ms) and a long speech stimulus. It is suggested in the literature that lower latency values might decrease the localization accuracy in the case of short stimuli. For example, this was the case for 1.5-s to 2.5-s long stimuli and 96-ms latency in [17].

Several studies have shown that performance in localization is poorer for lateral sources compared to frontal sources. For example results reported in [40] from two large scale studies (Preibisch-Effenberger 1966 and Haustein and Schirmer 1970), show that for a fixed head condition and a 100 ms white noise pulse, the localization uncertainty was ±3.6° for frontal sources (azimuth 0°), ±5.5° for rear

sources (azimuth 180°), and around ±10° for lateral sources (azimuth 90° and 270°).

The results in the present study suggest that the resort to a reference head tracking or even a delayed head tracking combined with head-movement yield a balanced performance for all azimuths. Indeed, by pointing the head toward the sound source, the source becomes frontal for the listener, which should thus compensate the larger error usually reported for more lateral azimuths. Nevertheless, for the real sources, an increase of error for lateral sources can still be observed. It is possible that for those angles, listeners deviated more than planned from the expected position of the head, resulting in a larger uncertainty than the ±2° previously mentioned for the real sources. However, errors are slightly smaller for lateral sources in this study compared to the results reported in [40], suggesting that the listeners still benefited from pointing toward the source in this case.

For the reference and latency head-tracking conditions, the accuracy of the listeners in this experiment was comparable to the performance of untrained listeners with non-individualized HRTFs found in [41] for frontal sources. The pointing method used in this article is a different task than locating a source when standing still. This may explain that the errors in this test are slightly larger than the errors for the frontal azimuth (0°) in [41]. The absolute error was smaller for those two conditions compared to the average error reported with non-individual HRTFs in [8]. This might be explained as the reported average error was mixed between azimuths in the latter. As expected, the listeners were less accurate in the case of the yaw estimation mismatch artefact. This is explained by the perceived angular shift of the sound source due to the tracking estimation mismatches. It is possible that the pointing method used in this experiment was disadvantageous for the yaw mismatch condition. Indeed this method requires the listeners to perform more movements that in a realistic situation.

It could be hypothesized that even a degraded head tracking with yaw estimation mismatch or latency still provides differential integration of the binaural cues which are enough to resolve most of the front-back confusions. However, this cannot be concluded from this experiment as the listeners knew that the source would not be coming from rear positions.

The results of the localization time suggest that the head-tracking artefacts did not affect the time it took for the listeners to locate the sound source. The small increase in localization time for the 400 ms latency artefact is in agreement with the results in [14], which found that for continuous sound stimuli, the response time was larger for latencies above about 90 ms.

9 CONSIDERATIONS FOR APPLICATIONS TO BINAURAL COMMUNICATION DEVICES

The present study suggests that, in the context of wearable binaural communication devices, a significant improvement in perceived externalization can be brought when the listener is performing head-tracked movements. Indeed, a significant enhancement in externalization was provided using a simplified dynamic binaural rendering algorithm designed to be implementable on wearable devices. All HRTFs and BRIRs were generic, and the length of the BRIRs was always truncated to 10 ms. A simulation of a yaw estimation mismatch, as can be obtained with a scaled-down head-tracking algorithm [5], did not affect the perception of externalization in the listening test reported in this study.

In realistic conditions, errors might accumulate with longer time of measure in the case of the yaw mismatch condition. This means that such an algorithm should include a re-initialization method to maintain errors in more acceptable ranges. In devices for which technical constraints allow it, this can be achieved with magnetometers and gyroscopes. A large latency (400 ms), did not affect auditory externalization either. Such latency is larger than any latency potentially occurring in the context of wearable binaural communication devices with remote microphone.

In application, when using partitioned convolution to generate ERs, the BRIRs would be fixed and stored in the wearable binaural communication devices. This means that the BRIRs may not be congruent with the room in which the user is. This raises the question of the dependence between the degree of room divergence and the potential improvement that head tracking combined with head motions could provide to the listener. Nevertheless, some of the methods described in [42] aim at extracting ERs from the signal retrieved from the microphones available on the devices themselves, which means that they would include the room information in every situation.

Externalization is known to be influenced by visual cues and phenomena such as the ventriloquist effect [43]. The experiment did not include a realistic visual reference for the sound source. Doing so could have decreased plausibility when the head-tracking artefacts yielded a perceptually moving auditory source while the visual reference was static. In a practical use of wearable binaural communication device, the presence of a visible real sound source might change how the tracking artefacts affect externalization, at least during movements. When the listener is not moving and the speaker is in the visual field, it is likely that visual capture could provide a certain amount of externalization. Hence, visual cues might increase the baseline externalization in every head-tracking condition.

In this study, the listeners were not given time to train with the generic HRTFs. At best they listened with those HRTFs in the first session and a minimum one week gap was respected between the two sessions. No visual feedback was given neither for them to learn. It is likely that a training could help improve the localization of sound sources in application. Indeed, with a wearable binaural communication device, the listener would constantly be stimulated by both audio and the corresponding visual feedback. This may help them learn quickly, and thus to perform better in localization [41]. It is also likely that visual capture could compensate small yaw mismatches when the target speaker is in the field of vision.

The results of this experiment suggest that the impact of latency on externalization and localization is small. Nevertheless, latencies might create some other disturbance on the long term. In the context of virtual reality, the issue of motion sickness, mostly thought to be visually induced, was shown to be potentially triggered by auditory cues as well [44]. Vection, which is the illusion of self-motion in the absence of real physical movement is another well-known artefact in this context. However, the influence of the visual cues might be larger than auditory cues for this type of issue [45].

10 CONCLUSION

In this study, two subjective listening experiments were performed to evaluate the effect of various head-tracking conditions on the perception of auditory externalization and the performance in localization. Artefacts potentially occurring in the context of wearable devices were used in particular. The binaural synthesis was achieved with 10-ms non-individualized BRIRs superimposed to a direct sound obtained with non-individualized HRTFs.

For each DRR setting, the condition with no head tracking always resulted in poorer externalization ratings compared to all the conditions including head tracking. The results suggest that the yaw estimation mismatch of the head tracking simulated in this study, as well as a large latency (400 ms), might not affect auditory externalization. This suggests that listeners could still benefit from the additional cues provided by the head motion to externalize the sound source, even when the head-tracking was substantially degraded. Further studies should investigate if this observation holds with a realistic visual reference, which might affect how the "slewing" of the sound source is interpreted by the auditory system.

A large head-tracking latency did not affect the performance in localization compared to the reference tracking. This suggests that the listeners might have understood spontaneously the nature of the artefact in this case and could take advantage of the length of the stimuli. The yaw estimation mismatch, naturally led to larger errors in localization. This is explained by the head-tracking errors, which result in a perceptual shift of the azimuth of the virtual sound source. Further works could investigate to which extent visual capture could compensate for those errors in a realistic scenario.

11 ACKNOWLEDGMENT

The work reported in this paper has been co-financed by Innosuisse under grant agreement 41312.1 IP-LS.

12 REFERENCES

[1] H. Wallach, "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization," *J. Exp. Psychol.*, vol. 27, no. 4, pp. 339–368 (1940 Oct.). https://doi.org/10.1037/h0054629.

[2] W. Brimijoin, A. Boyd, and M. Akeroyd, "The Contribution of Head Movement to the Externalization and Internalization of Sounds," *PLoS ONE*, vol. 8, paper 12 (2013 Dec.). https://doi.org/10.1371/journal.pone.0083068.

[3] E. Hendrickx, P. Stitt, J. Messonnier, J. Lyzwa, B. F. Katz, and C. Boishéraud de, "Influence of Head Tracking on the Externalization of Speech Stimuli for Non-Individualized Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 2011–2023 (2017 Mar.). https://doi.org/10.1121/1.4978612.

[4] S. Li, R. Schlieper J. E, and J. Peissig, "The Impact of Trajectories of Head and Source Movements on Perceived Externalization of a Frontal Sound Source," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 9988.

[5] V. Grimaldi, L. Simon, M. Sans, G. Courtois, and H. Lissek, "Human Head Yaw Estimation Based on Two 3-Axis Accelerometers," *IEEE Sens. J.*, vol. 22, no. 17, pp. 16963–16974 (2022 Sep.). https://doi.org/10.1109/JSEN.2022.3185712.

[6] S. Li, R. Schlieper, and J. Peissig, "The Impact of Head Movement on Perceived Externalization of a Virtual Sound Source With Different BRIR Lengths," in *Proceedingsof the AESConference on Immersive and Interactive Audio* (2019 Mar.), paper 40.

[7] E. Wenzel, "The Relative Contribution of Interaural Time and Magnitude Cues to Dynamic Sound Localization," in *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 80–83 (New Paltz, NY) (1995 Nov.). https://doi.org/10.1109/ASPAA.1995.482963.

[8] D. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio. Eng. Soc.*, vol. 49, no. 10, pp. 904–16 (2001 Feb.).

[9] F. L. Wightman and D. J. Kistler, "Resolution of Front–Back Ambiguity in Spatial Hearing by Listener and Source Movement," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2841–2853 (1999 May). https://doi.org/10.1121/1.426899.

[10] W. Thurlow and P. Runge, "Effect of Induced Head Movements on Localization of Direction of Sounds," *J. Acoust. Soc. Am.*, vol. 42, no. 2, pp. 480–488 (1967 Aug.). https://doi.org/10.1121/1.1910604.

[11] W. Noble, "Auditory Localization in the Vertical Plane: Accuracy and Constraint on Bodily Movement," *J. Acoust. Soc. Am.*, vol. 82, no. 5, pp. 1631–1636 (1987 Nov.). https://doi.org/10.1121/1.395154.

[12] S. Perrett and W. Noble, "The Contribution of Head Motion Cues to Localization of Low-Pass Noise," *Atten. Percept. Psychophys.*, vol. 59, pp. 1018–1026 (1997 Jan.). https://doi.org/10.3758/BF03205517.

[13] Y. Iwaya, Y. Suzuki, and D. Kimura, "Effects of Head Movement on Front-Back Error in Sound Localization," *Acoust. Sci. Technol.*, vol. 24, pp. 322–324 (2003 Sep.). https://doi.org/10.1250/ast.24.322.

[14] D. Brungart, A. Kordik, and B. Simpson, "Effects of Headtracker Latency in Virtual Audio Displays," *J. Audio. Eng. Soc.*, vol. 54, no. 1/2, pp. 32–44 (2006 Jan.).

[15] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," in *Proceedings of NAG/DAGA*, pp. 1063–1066 (Rotterdam, The Netherlands) (2009 Jan.).

[16] P. Stitt, E. Hendrickx, J.-C. Messonnier, and B. Katz, "The Influence of Head Tracking Latency on Binaural Rendering in Simple and Complex Sound Scenes," presented at the *140th AES Convention of the Audio Engineering Society* (2016 May), paper 9591.

[17] J. Sandvad, "Dynamic Aspects of Auditory Virtual Environments," presented at the *100th Convention of the Audio Engineering Society* (1996 May), paper 4226.

[18] E. Wenzel, "Effect of Increasing System Latency on Localization of Virtual Sounds," in *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction* (1999 Apr.), paper 16-004.

[19] V. Grimaldi, G. Courtois, L. Simon, and H. Lissek, "Externalization of Virtual Sounds Using Low Computational Cost Algorithms for Hearables," in *Proceedings of Forum Acusticum*, pp. 917–921 (Lyon, France) (2020 Dec.). https://doi.org/10.48465/fa.2020.0461.

[20] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande, "Binaural Hearing Aids With Wireless Microphone Systems Including Speaker Localization and Spatialization," presented at the *138th Convention of the Audio Engineering Society* (2015 May), paper 9242.

[21] W. Gardner, "Efficient Convolution Without Input-Output Delay," *J. Audio. Eng. Soc.*, vol. 43, no. 3, pp. 127–136 (1995 Mar.).

[22] A. Torger and A. Farina, "Real-Time Partitioned Convolution for Ambiophonics Surround Sound," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 195–198 (New Platz, NY) (2001 Oct.). https://doi.org/10.1109/ASPAA.2001.969576.

[23] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound Externalization: A Review of Recent Research," *Trends Hear.*, vol. 24, paper 2331216520948390 (2020 Sep.). https://doi.org/10.1177/2331216520948390.

[24] A. Kolarik, B. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory Distance Perception in Humans: A Review of Cues, Development, Neuronal Bases, and Effects of Sensory Loss," *Atten. Percept. Psychophys.*, vol. 78, pp. 373–395 (2015 Nov.). https://doi.org/10.3758/s13414-015-1015-1.

[25] J. Kates, K. Arehart, R. Muralimanohar, and K. Sommerfeldt, "Externalization of Remote Microphone Signals Using a Structural Binaural Model of the Head and Pinna," *J. Acoust. Soc. Am.*, vol. 143, no. 5, pp. 2666–2677 (2018 May). https://doi.org/10.1121/1.5032326.

[26] M. Lavandier T. Leclère, and F. Perrin, "On the Externalization of Sound Sources With Headphones Without Reference to a Real Source," *J. Acoust. Soc. Am.*, vol. 146, no. 4, pp. 2309–2320 (2019 Oct.). https://doi.org/10.1121/1.5128325.

[27] S. Li, R. Schlieper, A. Tobbala, and J. Peissig, "The Influence of Binaural Room Impulse Responses on Externalization in Virtual Reality Scenarios," *Appl. Sci.*, vol. 11, no. 21, paper 10198 (2021 Oct.). https://doi.org/10.3390/app112110198.

[28] G. Courtois, V. Grimaldi, H. Lissek, P. Estoppey, and E. Georganti, "Perception of Auditory Distance in Normal-Hearing and Moderate-to-Profound Hearing-Impaired Listeners," *Trends Hear.*, vol. 23, pp. 1–18 (2019 Feb.). https://doi.org/10.1177/2331216519887615.

[29] E. Kreyszig, *Advanced Engineering Mathematics* (Wiley, New York, NY, 1979), 4th ed.

[30] A. Bronkhorst and T. Houtgast, "Auditory Distance Perception in Rooms," *Nature*, vol. **397**, pp. 517–520 (1999 Feb.). https://doi.org/10.1038/17374.

[31] D. Mershon and L. King, "Intensity and Reverberation as Factors in the Auditory Perception of Egocentric Distance," *Percept. Psychophys.*, vol. 18, paper 409–415 (1975 Nov.). https://doi.org/10.3758/BF03204113.

[32] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R* (Sage London, UK, 2012).

[33] J. Miles and M. Shevlin, *Applying Regression and Correlation: A Guide for Students and Researchers* (Sage, London, UK, 2001).

[34] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. Katz, and C. Boishéraud, "Improvement of Externalization by Listener and Source Movement Using a 'Binauralized' Microphone Array," *J. Audio. Eng. Soc.*, vol. 65, pp. 589–599 (2017 Aug.). https://doi.org/10.17743/jaes.2017.0018.

[35] R. Crawford-Emery and H. Lee, "The Subjective Effect of BRIR Length on Perceived Headphone Sound Externalization and Tonal Colouration," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), 9044.

[36] J. Sinker and B. Shirley, "The Effect of Early Impulse Response Length and Visual Environment on Externalization of Binaural Virtual Sources," presented at the *140th Convention of the Audio Engineering Society* (2016 May), paper 9552.

[37] S. Werner, F. Klein, and K. Mayenfels, and T. Brandenburg, "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*, pp. 1–6 (Lisbon, Portugal) (2016 Jun.). https://doi.org/10.1109/QoMEX.2016.7498973.

[38] H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel, "Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments," *Acta. Acust. united Acust.*, vol. 102, no. 1, pp. 107–118 (2016 Jan.). https://doi.org/10.3813/AAA.918928.

[39] E. Wenzel, "Effect of Increasing System Latency on Localization of Virtual Sounds With Short and Long Duration," in *Proceedings of the International Conference on Auditory Display*, pp. 185–190 (Espoo, Finland) (2001 Jul.).

[40] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1999).

[41] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. Ferreira, and J. Santos, "On the Improvement of Localization Accuracy With Non-Individualized HRTF-Based Sounds,"

J. Audio. Eng. Soc., vol. 60, no. 10, pp. 821–830 (2012 Oct.).

[42] V. Grimaldi, *Auditory Externalization of a Remote Microphone Signal*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (2022 Mar.). https://doi.org/10.5075/epfl-thesis-8959.

[43] D. Warren, R. Welch, and T. McCarthy, "The Role of Visual-Auditory Compellingness in the Ventriloquism Effect: Implications for Transitivity Among the Spatial Senses," *Percept. Psychophys.*, vol. 30, no. 6, pp. 557–564 (1981 Nov.). https://doi.org/10.3758/BF03202010.

[44] B. Keshavarz, L. Hettinger, R. Kennedy, and J. Campos, "Demonstrating the Potential for Dynamic Auditory Stimulation to Contribute to Motion Sickness," *PLoS ONE*, vol. 9, no. 7, paper e101016 (2014 Jul.). https://doi.org/10.1371/journal.pone.0101016.

[45] B. Keshavarz, L. Hettinger, D. Vena, and J. Campos, "Combined Effects of Auditory and Visual Cues on the Perception of Vection," *Exp. Brain Res.*, vol. 232, pp. 827–836 (2013 Dec.). https://doi.org/10.1007/s00221-013-3793-9.

APPENDIX

Algorithm 1 Yaw estimation mismatch simulation

$v(n) = yaw_r(n) - yaw_r(n-1)$ & $v_{df}(n) = |v(n)| - |v(n-1)|$
if $sgn(v_{df}(n)) \neq sgn(v_{df}(n-1))$ & $v_{df}(n) < 0$ **then**
  $v_{max} = v(n)$
  if $|v_{max}| > Thr$ **then**
    $Thr = \gamma_v |v_{max}|$
if $|v(n)| \geq Thr$ **then**
  $yaw_e(n) = yaw_e(n-1) + v(n)$
else
  $yaw_e(n) = yaw_e(n-1)$

THE AUTHORS


Vincent Grimaldi


Laurent S.R. Simon


Gilles Courtois


Hervé Lissek

Vincent Grimaldi received an M.Sc. degree in Sciences and Technologies "Parcours ATIAM" from Sorbonne University and Institute for Research and Coordination in Acoustics/Music (IRCAM; Paris, France). He also has an M.Sc.Eng. degree with a major in Bioengineering from ENSAM ParisTech (Paris, France). In 2022, he graduated with a Ph.D. degree from Ecole Polytechnique Fédérale de Lausanne (Lausanne, Switzerland), in the Acoustic Group of the LTS2 Signal Processing Lab, working mainly on the enhancement of auditory spatial perception in the context of hearing instruments. His main research interests include spatial hearing and psychoacoustics in general.

•

Laurent S. R. Simon graduated from the Institute of Sound Recording, University of Surrey, with a Ph.D. on "perceptual cues-based sound reproduction on the horizontal plane" in 2011. His subsequent research work has been focusing on spatial auditory perception and processing, ranging from audio source separation to perceptual attributes for the evaluation of head-related transfer functions, applied either to consumer audio or to hearing research. Since 2021, he has been employed at Sonova AG, where he is part of the Research and Development team.

•

Gilles Courtois received an M.Sc./M.Eng. from the Institut National des Sciences Appliquées (INSA; Lyon, France) in 2012 and then a Ph.D. degree in hearing technologies from the EPFL in 2016. During his academic path, he worked on the topic of spatial hearing preservation/recreation for hearing aid applications. Since 2018, he has been with Sonova AG and has been working on multiple hearing aid and consumer audio features, including frequency compression, audio stream spatialization, and audio processing with variable latency. In 2022, he was given a hearing performance expert title, holding the topics of minimal-to-mild hearing losses and spatial hearing integrity.

•

Hervé Lissek received an M.Sc. degree in fundamental physics from Université Paris XI (Orsay, France) in 1998 and a Ph.D. degree in physics, with a specialization in acoustics from Université du Maine (Le Mans, France) in 2002. He was a post-doctoral researcher at EPFL between 2003 and 2005. Since 2006, he has been leading the acoustic group at EPFL (affiliated since 2015 with the Signal Processing Laboratory LTS2). His current research interests include electroacoustic absorbers, acoustic metamaterials, signal processing for microphone arrays, and binaural hearing aids. He has been elected Vice President of the Swiss Acoustical Society and French Acoustical Society as well as Swiss Representative of the European Acoustics Association, International Institute of Noise Control Engineering, International Commission for Acoustics, and Aeroacoustics Specialists Committee of the Council of European Aerospace Societies.