Journal of the Audio Engineering Society
Vol. 65, No. 7/8, July/August 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2017.0018

Improvement of Externalization by Listener and Source Movement Using a "Binauralized" Microphone Array

ETIENNE HENDRICKX[1], **PETER STITT**,[2] *AES Associate Member*, **JEAN-CHRISTOPHE MESSONNIER**[3],
(etienne.hendrickx@univ-brest.fr) (peter.stitt@limsi.fr) (jcmessonnier@cnsmdp.fr)

JEAN-MARC LYZWA[3], **BRIAN F.G. KATZ**[4], **AND CATHERINE DE BOISHERAUD**[3]
(jmlyzwa@cnsmdp.fr) (brian.katz@upmc.fr) (cdeboisheraud@cnsmdp.fr)

[1] *University of Brest, CNRS, Lab-STICC UMR 6285, 6 avenue Victor Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France*
[2] *Audio Acoustics Group, LIMSI, CNRS, Université Paris-Saclay, Orsay, France*
[3] *Conservatoire National Supérieur de Musique et de Danse de Paris, 209, avenue Jean-Jaurès, 75019 Paris, France*
[4] *Sorbonne Universités, UPMC Univ. Paris 06, CNRS, Institut d'Alembert, Paris, France*

Several studies report a collapse of externalization when listening to binaural content using non-individualized HRTFs. In other words, sound sources tend to be perceived inside the head when they should be perceived outside the head, as in the real acoustic world. A previous experiment, conducted with experienced subjects, revealed that large head movements coupled with a head tracking device could substantially improve the externalization of a speech stimulus, recorded in slightly reverberant conditions with a six-channel microphone array and then "binauralized" for headphones as six virtual loudspeakers around the subject (one loudspeaker per microphone signal). In the present study a similar experiment was conducted with subjects having no previous experience with binaural audio. Similar improvements were found. In an additional condition the roles were reversed: the subjects' heads remained stationary while the sound sources were automatically moved around subjects. Results showed that source movements without tracking can also enhance externalization but to a lesser extent than head-tracked movements.

0 INTRODUCTION

Sound scenes can be recreated over headphones using binaural processing. The principle is to recreate at the eardrums of the listener the pressure signals corresponding to an intended sound source or scene. Binaural recordings can be achieved in two different ways—either naturally or synthetically. In natural recordings real sound sources are captured with microphones placed in the ears of a dummy head or of a listener. In binaural synthesis the Head-Related Transfer Functions (HRTFs) of a listener are anechoically measured at many source positions and incorporated as digital filters that are then used to synthesize stimuli. If the measurement takes place in a reverberant room, then the directional reverberation is also captured. Such a measurement is known as binaural room impulse responses (BRIRs).

One advantage of binaural synthesis is that the rendering system can be coupled with a head tracking device, thus enabling the virtual sources to move appropriately to the listener's head movements. Moreover, if the measurements of HRTFs (or BRIRs) have been conducted with several listeners, one can choose the HRTF that will be used to filter each individual sound source in the scene. Ideally, the HRTF should be personalized to the listener, however, measuring individualized HRTFs can often be a complex and expensive process [1].

An easier solution is to listen "through the ears" of another listener whose HRTF is already available. However, several studies have shown that the use of non-individualized HRTFs can cause problems such as increased numbers of front-back confusions [2] and "in-head" localization [3, 4]. In other words, stimuli tend to be *internalized*, that is perceived inside the head, when they should be *externalized*, or perceived outside the head, as in the real acoustic world.

The issue of front-back confusions can be alleviated by training to the HRTF [5], selecting the perceptually best-rated HRTF from a set [6], or adding a head tracking device to the binaural rendering system and moving the head [7–9]. However, it is still unclear how the collapse of externalization can be compensated efficiently.

In [10], training over a month to the HRTF improved externalization levels in the vertical plane. In the horizontal plane some improvements were also observed, but it was very slight. The addition of reverberation to the sound source has also been found to increase the perceived externalization of the sound source [7, 11–13]. Other studies have been performed to investigate the benefit of head tracking, however results have been conflicting.

Several studies claimed that head movements coupled with head tracking could enhance externalization substantially; however these studies were either informal [14], lacked sufficient subjects (three subjects only) [15], or used more or less degraded individualized HRTFs, but not non-individualized HRTFs [16]. Other studies suggested that the effect of head movements coupled with head tracking on externalization was small [17] or even null [7].

In [7], nine naïve subjects listened to brief speech stimuli (3 s long) reproduced at different azimuth positions (0°, ±45°, ±135°, 180°) with three different levels of reverberation: anechoic, early reflections only, and full reverberation (early reflections + late diffuse reverberation response, with a mid-band reverberation time of 1.5 s).

Results showed that freestyle head movements coupled with head tracking did not increase externalization, whether individualized or non-individualized HRTFs were used for the binaural rendering. However, the study acknowledged that the short duration of the stimuli may have limited the ability of the subjects to take advantage of cues derived from head movements. The fact that results were averaged across all positions before analysis may also explain why the effect of head tracking was not significant. As lateral sources are already well externalized without head tracking [3, 4], it is rather for frontal and rear sources that head tracking can be expected to have a substantial impact. Thus, any small improvements occurring for lateral sources may statistically mask larger improvements for frontal and rear sources.

In [17], six subjects listened to a 3 s broadband Gaussian noise presented from 40 different locations: eight azimuths every 45° for five different elevations (−36° to +36°), using non-individualized HRTFs. An improvement of the externalization rate (defined as the percentage of time a stimulus was perceived outside the head) could be observed with non-individualized HRTFs, but it was quite slight (from 74.5% to 83.5%). As in [7], stimuli were quite brief (3 s) and results were averaged across all positions.

In a recent study by the authors using speech stimuli synthesized in the horizontal plane with non-individualized HRTFs, large head movements (±90°) coupled with head tracking improved externalization substantially at some azimuth positions [18]. Although differences in methodology make comparisons of results problematic, it can be hypothesized that improvements were more substantial in [18] than in [7] and [17] because of the longer stimulus (8 s instead of 2–3 s) and larger head movements. Another important difference is that data in [18] were analyzed by position. Results confirmed that externalization of lateral stimuli was high even without head tracking, leading to small or null improvements when head tracking was added, whereas externalization of frontal and rear stimuli was very low without head tracking, leading to substantial improvements when head tracking was added.

The substantial improvements observed in [18] could also be explained by the chosen recording and reproduction method: a six-channel microphone array, binauralized by convolving each microphone signal with publicly available HRTFs (equivalent to a simulation over headphones of a six-channel virtual loudspeaker system in an anechoic room). Using a microphone array is an artistic choice and does not result in a natural reproduction of a sound source in a room, no matter how reproduced. Convolving an anechoic recording of a sound source with binaural room impulse responses, as in [7], could have led to a more natural reproduction, and it can be assumed that resulting externalization could have been higher when subjects were not head-tracked, thus minimizing the beneficial impact of head tracking. Investigating the case of binauralized microphone arrays is, however, essential for several reasons:

- Microphone arrays have been a major category of recording approaches for multichannel sound reproduction, as they are the natural extension of the principles inherited from traditional stereophonic recording techniques. Thus, sound engineers might not be willing, at least in the near future, to change their habits so drastically.
- A microphone array recording is also compatible with diffusion over real loudspeakers. This versatility enables substantial time saving during recording and post-production.
- In the context of a commercial recording, if the producers or artists want to record in a specific room, BRIRs of this room might not be available or it might not be possible to measure them. Such measurements are expensive in time and resources, requiring arrays of speakers, accurate speaker positioning systems, and specialized software and technicians [1]. In order to be compatible with head tracking, a large number of positions have to be measured, and the whole process can be especially time-consuming if several sets of HRTFs are to be recorded (so that the listener can choose between several HRTF sets). On the other hand, a microphone array is an easy way to capture both direct sound and reverberation of the room, and a preexisting HRTF database can be used for the binauralization.
- Such virtual loudspeaker systems are already used by major broadcasting institutions such as Radio France [19].

In [18] subjects were professional sound engineers used to listening to binaural content. The question now is to what extent the improvement of externalization due to large head movements coupled with head tracking observed previously can be generalized to the general population of listeners who have little experience with binaural audio. In a comparative study of binauralized recording setups [20] with non-individualized HRTFs and no head tracking, significant differences between expert and naïve subjects could

be observed: for example, naïve subjects localized sound sources less accurately and they reported less often that the virtual sources seemed to come from real sources out in the world. However, it can be hypothesized that the lack of head tracking substantially slowed down their adaptation to binaural spatialization.

As an alternative to using an actual head tracking system, which requires additional software and hardware equipment, head movements can be emulated by moving the virtual sound sources. In [8], source movements were as effective as head movements to resolve front-back ambiguities, yet only if the sound sources were moved by the subject himself (using the "arrow" keys on a keyboard to change the position of the sources). In [21], small head movements (±4°) were emulated using a random process that generated fluctuations statistically similar to the actual head micro-movements that subjects tend to make involuntarily even when they are requested to keep the head still. These fluctuations were then applied to the binaural rendering system. A slight increase in subjects' feeling of spaciousness and front/back extent was observed with movie soundtrack excerpts. In [22], emulations of smaller head movements (±2°) increased externalization for only some subjects (21%).

Improvements brought by micro-movements of sources in [21] and [22] seem rather moderate. More substantial effects might be observed if sources are moved along larger angles around subjects, in the same way as [18] suggests that head tracking cannot enhance externalization substantially unless subjects perform large head movements.

The present study investigates externalization of speech stimulus using non-individualized HRTFs and a binauralized microphone array with two aims: (1) to determine whether or not large head movements coupled with head tracking can enhance externalization for inexperienced subjects as were observed for experienced subjects in [18], (2) to determine whether or not similar improvements of externalization can be observed if large movements are performed by the sound sources for static head orientations.

1 EXPERIMENTAL SETUP

The protocol was very similar to the one described in [18]. Subjects listened to a binaural stimulus, consisting of a male voice either at 0° or at 180° (at 0° elevation), with reverberation from all around the subject. In one condition, head tracking was inactive and subjects were asked to keep their heads stationary. In another condition, head tracking was active and subjects were asked to make large head movements. In a third condition, the roles were reversed: subjects' heads were static while sound sources were automatically rotated around the subjects using binaural rendering. After each presentation, subjects reported to what extent the stimulus was externalized.

1.1 Stimulus

The stimulus consisted of an 8 s extract from the French poem "L'Albatros" by Charles Baudelaire, read by a male talker ($f_0 = 107$ Hz). The recording was made in a recording studio at the Conservatory of Paris (area of ≈30 m^2, midband reverberation time ≈0.24 s, direct-to-reverberant ratio ≈ 16 dB).

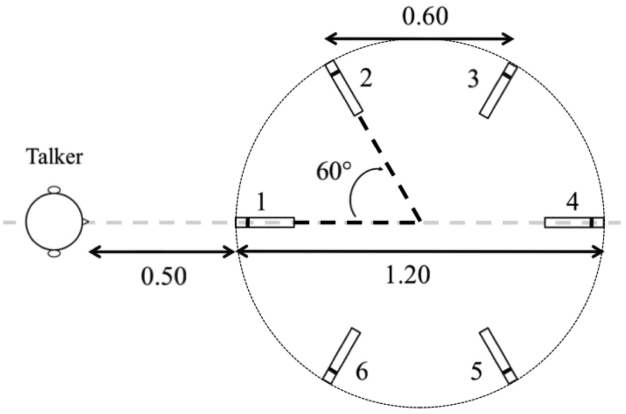

Fig. 1. Microphone array configuration used for the recording of the stimulus, consisting of one front microphone (microphone 1), principally capturing direct sound, and five other microphones (microphones 2–6), enabling to capture directional reverberation. The microphones were DPA 4021 (cardioid directivity). The array height was 1.65 m (height of the mouth of the talker). Dimensions in meters.

The stimulus was recorded with a six-channel equal-segment microphone array, presented in Fig. 1 and described in [23]. This configuration was selected on the basis of an informal comparative study of several microphone arrays with 10 subjects that suggested that this configuration provided the most natural audio scene when binauralized and listened over headphones.

1.2 Reproduction Setup

The listening test took place in a double-walled soundproof booth at the Conservatory of Paris (background noise level ≈25 dB A). The lights were turned off in order to minimize the influence of any visual stimuli. Subjects sat in the center of the room.

Stimuli were presented over headphones (Sennheiser HD 600). The sound pressure level was adjusted to ≈65 dB A (SLM, slow response) by placing the headphones on a dummy head (Neumann KU 100). Playback, interface, and data capture were controlled by software implemented in Max/MSP on a MacBook Pro computer connected to a RME Fireface 800 soundcard.

1.3 Binaural Rendering and Head Tracking Device

The binauralization of the resulting six-channel recording was made so as to give the impression of being at the center of the microphone array. The signals were processed in two different ways, corresponding to two different orientations of the sound scene. It was decided to focus on the following two orientations as they correspond to the azimuths at which the externalization rates reported in previous studies are the lowest [3, 4].

1. Orientation 0°: the signal from the front microphone (microphone 1) was processed using the HRTF for 0°. The "reverberation" microphones were rotated accordingly: the signal from microphone 2 was processed using the HRTF for 60°, the microphone 3 using the HRTF for 120°, etc. The six resulting processed signals (one for each microphone) were then added to generate the left and right ear signals.
2. Orientation 180°: the signal from the front microphone was processed using the HRTF for 180°. The "reverberation" microphones were rotated accordingly: the signal from microphone 2 was processed using the HRTF for −120°, the microphone 3 using the HRTF for −60°, etc.

When head tracking was active, head movements cause the position of the six signals to rotate in the opposite direction, keeping the signals fixed with respect to the external world.

The rendering was carried out using the binaural engine *Bipan* [24] that uses anechoic measured HRTFs at either 15° or 5° azimuthal spacings, with a diffuse-field equalization to compensate for the transfer functions of the measuring equipments (microphones and loudspeakers). HRTFs are decomposed into minimum phase (for spectral cues) and pure delay (for ITD cues). Minimum phase transfer functions are modeled by infinite impulse response filters that are linearly interpolated every 1°. Thus, filters change every 1°, with a 1 ms cross-fade to smooth transitions between filters. ITD delays vary continuously as the subject moves his/her head using linear interpolations between the ITD of two consecutive known positions.

The head tracking was carried out using the open-source hardware/software solution *Hedrot* [25]. The head tracking device, attached at all times to the subjects' headphones, was connected to the computer via a Teensy 3 USB board, updating the stimuli in response to head movements at a rate of 300 Hz (3 ms). The total tracking system latency averaged 48.1 ms (SD = 5.3 ms). This is below the threshold of detectability, which has been found to be 60 ms for the best listeners [26]. Furthermore, [27] found that latencies as high as 500 ms did not change the perceived externalization.

Three different non-individualized HRTF sets, selected from the publicly available LISTEN database [28], were intermixed in order to investigate whether or not the impact of head tracking could change depending on the employed HRTF and also in order to minimize any HRTF learning effect (if only one HRTF set was used, it could be difficult to separate head movements effects or source movement effects from those due to learning processes).

1.4 Subjects and Protocol

Nine naïve subjects took part in the experiment (four women and five men, aged 18–50 years). None reported any known hearing loss, and none had experience with binaural audio nor with laboratory listening tests.

Table 1. Six-point scale used to report externalization.

Grade	Reported externalization
0	The source is at the center of my head.
1	The source is not at the center of my head but still in my head.
2	The source is at my ear or on my skull.
3	The source is externalized but near the head.
4	The source is externalized and within my reach.
5	The source is externalized and remote.

They evaluated three different conditions:

- **NM**: no head movement, no head tracking, no source movement.
- **HM**: large head movements with head tracking, no source movement.
- **SM**: no head movement, no head tracking, large source movements.

Subjects were requested to hold their heads in a natural upright position when listening to a stimulus.

For condition **HM**, the presentation of the 8 s stimulus was divided into 3 parts:

1. 5.5 s of speech stimulus, during which subjects turned their head in one full cycle first to the left (−90°) and then to the right (+90°) before returning to forward-facing (0°). All subjects were asked to make the same movements as this ensured that they all received similar cues. The controlled requested movements were similar to those proposed in [29].
2. 1 s silence. By the end of this silence all head movements should be completed and subjects should be forward-facing again (0°), heads still.
3. 2.5 s of stimulus with subject's head stationary.

After each presentation subjects reported to what extent the sound source was externalized using a six-point scale displayed on a computer screen (see Table 1). It was specified to subjects that scores ≥ 3 (i.e., externalized source) were associated to when the voice of the talker appeared to be emanating from a source in the world, outside the head, and scores < 3 when the voice of the talker appeared to be emanating from somewhere inside the head (i.e., internalized source) or from a source located at the frontier between the skull and the exterior. Once subjects responded, the next stimulus was automatically played.

Note that subjects were to report to what extent the sound source had been externalized *during the last 2.5 s* of the presentation, that is from the moment they were forward-facing and stationary again. In other words, subjects were to report to what extent a sound source was externalized *after* they had moved their heads, in contrast with previous studies where subjects were to report to what extent a sound source had been externalized *while* they were moving their heads. This choice of procedure enabled the investigation of whether the externalization provided by head

movements persisted even though the subject had stopped moving his/her head.

For condition **NM**, the procedure was the same except that subjects were instructed to keep their heads still, looking straight ahead for the whole stimulus presentation.

For condition **SM**, subjects were also requested to keep their heads still and to look straight ahead during the whole stimulus presentation, however, the sound scene moved around subjects for the first 5.5 s of the stimulus:

- Orientation 0°: instead of remaining still at azimuth 0°, the signal obtained from the front microphone (microphone 1) was automatically rotated around the subject from 0° to the right (+90°) and then to the left (−90°) before returning to 0°. The rendered positions of the "reverberation" microphones (microphones 2–6 in Fig. 1) were rotated accordingly, so that the whole sound scene moved coherently. The rotation was computer-generated, with a constant angular speed and a duration of 5.5 s. Thus, the movement of the sound scene *relative* to the subject's head was very similar to that for condition **HM** at orientation 0°.
- Orientation 180°: instead of remaining still at azimuth 180°, the signal obtained from the front microphone was rotated from 180° to the left (−90°) and then to the right (+90°) before returning to 0°. The rendered positions of the "reverberation" microphones were rotated accordingly. The rotation of the sound scene was computer-generated, with a constant angular speed and a duration of 5.5 s. Thus, the movement of the sound scene *relative* to the subject's head was very similar to that for condition **HM** at orientation 180°.

In the real acoustic world, when a subject keeps his/her head fixed, there is little chance for an object to move around him/her with the same center of rotation as that of the head, and objects and directional reverberation never rotate accordingly as in condition **SM**. In other words, the goal of this additional condition was not to reproduce objects moving around the head in a realistic way, but to investigate whether continuous changes of coherent binaural and spectral cues similar to those experienced when rotating the head were sufficient to provide a substantial increase of externalization.

For each condition, there were 3 (HRTFs) × 2 (Orientations), repeated 10 times. Each condition took about 20 minutes to complete, and all subjects conducted the three conditions on three different days. The order of conditions was randomized for each subject. Within a condition, azimuth positions were presented in a randomized order that was different for each subject and for each repetition. The HRTF set always changed from one trial to another, thus minimizing potential HRTF learning effects.

2 HEAD MOVEMENTS

Head movements were recorded in order to verify how well the instructions were followed by subjects.

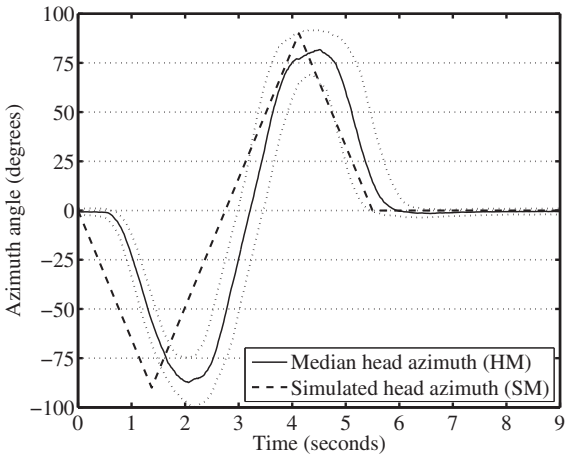

Fig. 2. Median head azimuth over the stimulus duration, across all subjects and all trials of condition **HM** (solid line), with first and third quartiles (dotted lines). Simulated head azimuth during condition **SM** (dashed line).

During conditions **NM** and **SM** subjects were asked to keep their heads as still as possible, and data shows that the median amplitudes of movement (defined as the difference between the maximum and minimum angles over the course of a given trial) were very small: 1.1° (IQR[1] =1.2°) for condition **NM** and 1.3° (IQR =1.6°) for condition **SM**.

Similar to the prescribed head movement protocol in [29], subjects in condition **HM** were asked to turn their heads in one full cycle first to the left (−90°) and then to the right (+90°) before returning to forward-facing. Data shows that the median minimum and maximum head angles were −92° and 94° with IQR of 18° and 14° respectively. The minimum angle occurred at a median value of 2.1 s (IQR = 0.36 s) and the maximum angle occurred at a median value of 4.4 s (IQR = 0.62 s). The median duration of the movement was 5.6 s (IQR = 0.8 s) and the median speed of the head motion for points at which the subjects were moving their heads (defined as faster than 10°/s) was 66°/s (IQR = 13°/s). A very weak correlation was found between the amplitudes of movement and the externalization scores ($\rho = -0.278$, Spearman's rho), which means that the variability of amplitudes of head movements was not large enough to have a substantial impact on externalization results. Similarly, the correlation between the speeds of motion and the externalization scores was very weak ($\rho = -0.122$, Spearman's rho).

Examination of data therefore suggests that subjects were reasonably compliant with the head movement instructions. Fig. 2 provides a comparison of subjects' head rotations during condition **HM** (solid line) with the computer-generated rotation of the sound scene of condition **SM** (dashed line). To make comparisons easier, the rotation of the sound scene in condition **SM** is converted into the corresponding head rotation that would give similar changes of binaural and spectral cues. The figure shows that subjects were globally

[1] Inter-Quartile Range

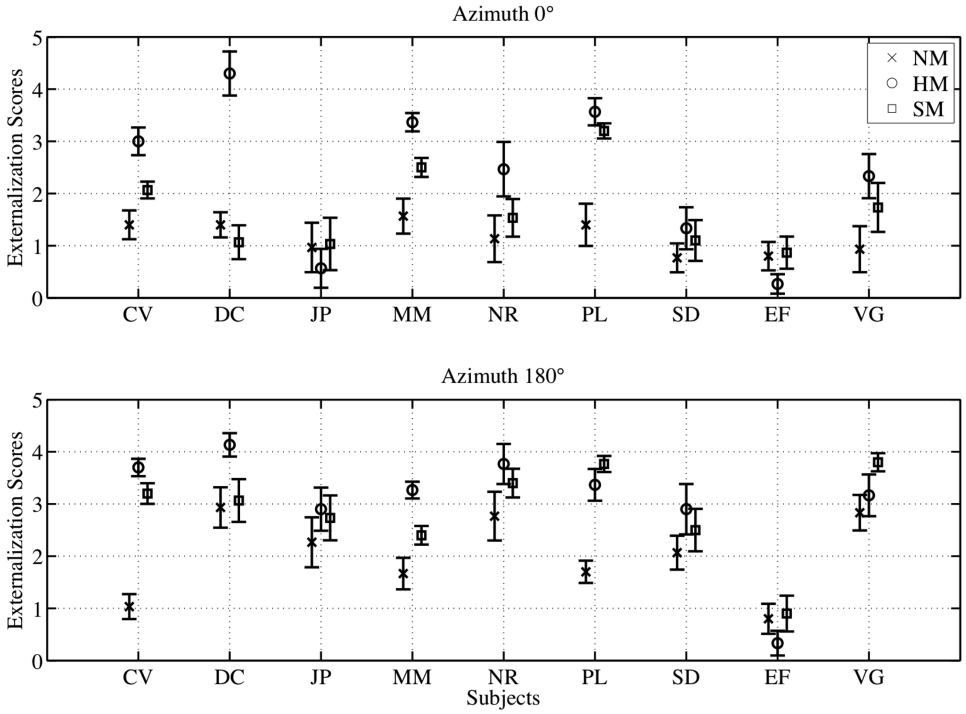

Fig. 3. Mean externalization scores and associated 95% confidence intervals of each subject by condition over all repetitions and HRTFs, for orientations 0° (top) and 180° (bottom). **NM**: no head movement, no head tracking, no source movement. **HM**: large head movements with head tracking, no source movement. **SM**: no head movement, no head tracking, large source movements.

a bit later compared to the computer-generated rotation of condition **SM**. Another important difference is that the angular speed was constant for condition **SM** whereas it alternated between acceleration and deceleration phases for condition **HM**. Note that the minimum and maximum head angles are slightly less for condition **HM**. This is because of the differences in the instants subjects changing direction during the stimulus, resulting in a smoothed median value.

3 EXTERNALIZATION RESULTS

A Friedman test conducted across all data reveals that there was no significant difference between the externalization scores of the three HRTF sets ($\chi^2(2) = 1.250$, $p = 0.535$). An in-depth examination of the data found that externalization scores were indeed very similar from one HRTF set to another, independent of condition and subject. Subsequent results are therefore presented across the three HRTF sets.

For orientation 0°, a Friedman test comparing externalization scores obtained for conditions **NM**, **HM**, and **SM** across all subjects, HRTFs, and repetitions, shows that the condition had a significant influence on externalization scores ($\chi^2(2) = 108.1$, $p < 0.001$). To follow up this finding, a series of Wilcoxon tests (non-parametric equivalent for *post hoc* procedures) was conducted between pairs of conditions, with p-values adjusted using the Bonferroni correction. Wilcoxon tests reveals that all conditions were significantly different one from another ($p < 0.001$ for all pair wise comparisons). Globally, condition **HM** provided the highest externalization, condition **NM** provided the lowest externalization, and condition **SM** was between the two.

Similar trends were observed for orientation 180°. A Friedman test shows that the condition had a significant influence on externalization scores ($\chi^2(2) = 120.9$, $p < 0.001$), and a series of Wilcoxon tests (with Bonferroni correction) reveals that all conditions were significantly different one from another much beyond the 0.05 level, except when comparing conditions **HM** and **SM**, where the difference was barely significant ($p = 0.048$).

Fig. 3 details mean externalization scores with associated 95% confidence intervals[2] obtained for each subject by condition at orientations 0° (top) and 180° (bottom). However, most previous studies rather used the externalization rate (defined as the percentage of time a stimulus was perceived outside the head, i.e., score ≥ 3 in the present study), which allows for a clearer presentation of results [16, 7, 17, 4]. Subjects' externalization rates obtained for each condition at orientations 0° (left) and 180° (right) are presented in Fig. 4. Even though results varied greatly between subjects and between the two orientations, some general trends can still be highlighted.

[2] 95% confidence intervals were obtained using the formulae: $CI(95\%)_{s,k,i} = \left[\bar{x}_{s,k,i} - 1.96\frac{\sigma_{s,k,i}}{\sqrt{n_{s,k,i}}}; \bar{x}_{s,k,i} + 1.96\frac{\sigma_{s,k,i}}{\sqrt{n_{s,k,i}}}\right]$ where : $\bar{x}_{s,k,i}$: mean score for subject s at orientation k in condition i

$\sigma_{s,k,i}$: standard deviation for subject s at orientation k in condition i
$n_{s,k,i}$: the number of trials conducted by subject s for orientation k during condition i.

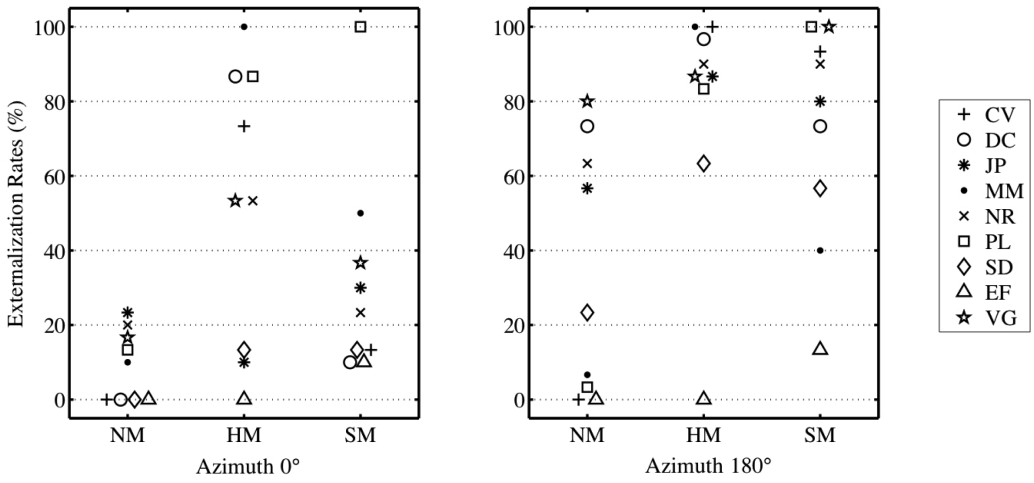

Fig. 4. Externalization rates of each subject by condition over all repetitions and HRTFs, for orientations 0° (left) and 180° (right). See Fig. 3 caption for condition notations. Externalization rate defined as the percentage of time a stimulus was perceived outside the head, i.e., externalization score ≥ 3 in the present study.

For orientation 0°, the externalization rates were always low during the static condition (**NM**), ranging from 0% to 23%, and with four subjects reporting no externalization at all (subjects CV, DC, EF, and SD). Compared to the static condition (**NM**), head movements (**HM**) improved externalization rates for all subjects except subjects EF and JP. The improvements could be quite substantial: $\geq +33\%$[3] for six subjects, and even $\geq +73\%$ for four subjects. Source movements (**SM**) also led to a higher externalization rate compared to the static condition (**NM**), yet improvements were much less pronounced: +20% for seven out of nine subjects. Substantial improvements were observed with subjects MM (+40%) and PL (+86%) only.

For orientation 180°, the externalization was globally higher compared to 0° for all conditions. Externalization rates were especially high for condition **HM**: $\geq 83\%$ for seven subjects, with subjects CV and MM reporting 100% externalization. However, it can be noted in Fig. 3 that the mean externalization scores were never greater than 4 (corresponding to a source "externalized and within reach"), except for subject DC. An in-depth examination of the data revealed that the proportion of stimuli perceived "remote" (score 5) was actually very low for all conditions and all orientations (from 0% to 7% of the trials). Except for subject EF, head movements (**HM**) always improved the externalization rate compared to the static condition (**NM**), with substantial improvements for most subjects: $\geq +23\%$ for seven subjects, $\geq +80\%$ for three subjects. The advantage of head movements (**HM**) over source movements (**SM**) was much less pronounced than at 0°, with three subjects (subjects EF, PL, and VG) reporting slightly more externalization during condition **SM** than during condition **HM**.

[3] +33% does not mean that the externalization rate increased by a third of its original value, i.e., from 10% to 13, 3%. It means that the externalization rate increased from 10% to 43%. All subsequent increases of externalization rates are presented this way.

4 DISCUSSION

4.1 Large Head Movements Improve Externalization Substantially

During condition **HM**, large head movements coupled with head tracking improved externalization substantially for most subjects. At azimuth 180°, large head movements enabled most subjects to obtain a very high externalization rate ($\geq +83\%$ for seven out of nine subjects). At azimuth 0°, even though the externalization rates provided by large head movements could be moderate ($\leq 53\%$ for five out of nine subjects), it was still a substantial improvement compared to the dramatically low externalization rates of condition **NM** (from 0% to 23% depending on the subject).

For condition **HM**, subjects were asked to report whether or not the stimulus was externalized *after* they had moved their heads, in contrast with previous studies, where subjects were asked to report whether or not the stimulus was externalized *while* they were moving their heads. The fact that more externalization was obtained during condition **HM** therefore shows that a stimulus, externalized by head movements and head tracking, can remain externalized even though the subject has stopped moving his/her head. This persistence of externalization highlights even more the practical interest of head tracking for everyday listening: listeners do not have to move their heads continuously to listen to binaural content with better externalization. A simple $\pm 90°$ movement is sufficient to observe persistent and substantial improvements, and informal tests suggest that improvements can be even more pronounced if head movements are larger, or repeated several times.

These findings, obtained with naïve subjects who had no experience at all with binaural audio nor with laboratory listening tests, are in accordance with a previous experiment conducted with professional sound engineers, who were used to listening to binaural content [18]. Results suggest that head tracking is effective in increasing externalization for most subjects, independent of experience. Moreover,

there was no significant difference between the three HRTF sets used in the experiment, which suggests that head tracking can improve externalization, independent of the HRTF set chosen by the listener.

Results of the present study are however at odds with [7] and [17]. In [7], no significant effect of head tracking on externalization could be observed, and subjects externalized judgments at a mean rate of 79% under reverberant conditions and of 40% under anechoic condition, whether head tracking was active or not. In [17] there was a significant increase of externalization rates due to head tracking but it was quite slight: from 74.5% to 83.5%.

Several hypothesis have already been mentioned in Sec. 0 to explain these conflicting results: substantial improvements may have been obtained in the present study because of the longer stimulus (8 s instead of 2–3 s), which gave subjects more time to "take advantage" of cues derived from head movements and enabled them to make larger movements. Another important difference is that the present study used orientations 0° and 180° only, which, according to [3, 4] are the orientations where stimuli are most likely to be internalized when the subject is not head-tracked, in contrast with lateral orientations, where stimuli are almost always externalized. In other words, the present study may have obtained substantial differences because it only used azimuths where there was much room for improvement when head tracking was added, whereas previous studies included azimuths at which sources were already well externalized even without head tracking, thus providing little or even no room for improvement when head tracking was added, minimizing the global effect of head tracking.

The fact that substantial improvements were obtained in the present study could also be explained by the chosen recording and reproduction methods: a speech stimulus recorded with a microphone array in a small slightly reverberant room, then convolved with anechoic measured HRTFs (corresponding to a simulation of loudspeakers in an anechoic listening room). If ideal binaural reproduction had been used (such as convolving an anechoic sound source with head-tracked binaural room responses, as in [7]), if the stimulus had been recorded in a more reverberant environment, or if the simulated listening room had included some reverberation (even in the form of a few early reflections), externalization could probably have been higher for condition **NM**, thus minimizing the beneficial impact of head tracking.

Head tracking might also be less essential in the case of audiovisual content such as movies and gaming as the presence of consistent visual cues might enhance externalization substantially. Indeed, it has been shown that the ventriloquist effect[4] can be effective in depth [30, 31]. On the other hand, it has been observed that we are sensitive to spatial congruence of audio and visual objects in 3D movies that is affect by seating position [32]. It has also been observed in the case of 2D and 3D movies that large disparities in depth between a sound source and its related visual source can affect the perceived suitability of a soundtrack for the image [33]. Further investigation is thus required for the case of audiovisual content.

4.2 Large Source Movements Also Improve Externalization Substantially, but to a Lesser Extent than Head Movements

Results obtained for condition **SM** show that large source movements can also improve externalization. At azimuth 180°, source movements were nearly as effective as head movements to improve externalization. At orientation 0°, improvements provided by source movements (**SM**) were much less pronounced compared to those provided by head movements (**HM**), but they were still significant.

The current general consensus is that the "realism" of the acoustics of signals entering the ear canal is critical for externalization [16]. Thus, head movements may increase externalization when the head is tracked because head tracking enables sound sources to move appropriately, or "realistically," to the listener's head movements. According to [16], whether a subject perceives a sound source inside or outside his/her head could be the result of a combination of the acoustic features of the auditory signal with an ongoing internal comparison between the subject's movement and the apparent movement of the sound source. However, during condition **SM**, source movements provoked continuous changes of binaural and spectral cues that were very similar to those experienced during condition **HM**, with the difference that movements during condition **SM** were unrealistic and not caused by the listener's own movements. Still it did enhance externalization, sometimes in a substantial way. This suggests that improvements due to head movements cannot only be explained by the listener and the sound source moving coherently, and that some improvements may be intrinsically linked to the continuous changes of binaural and spectral cues.

Fig. 2 also suggests that source movements in condition **SM** could be generated in a more "natural" way, for example with accelerating and reduction phases instead of a constant angular speed. This is a potentially interesting future study, as more natural continuous changes of binaural and spectral cues might enhance externalization even more.

Again, subjects were asked during condition **SM** to report whether or not the stimulus was externalized *after* the source movements. The fact that more externalization was obtained during condition **SM** compared to condition **NM** therefore shows that a sound scene, externalized by source movements, can remain externalized even though the sources have stopped moving. In practice, it suggests that a simple two-step solution can be proposed to listeners willing to listen to binaural content with better externalization, yet having no head tracking system: (1) they listen to a sample of the sound scene that moves around them, (2) the subjects then continue to listen to the content normally (that is without rotation), with enhanced externalization provided by the preliminary source movements.

[4] When presented with a spatially discordant auditory-visual stimulus, subjects sometimes perceive the sound stimulus as coming from the location of the related visual stimulus. Such a phenomenon is referred to as *ventriloquism*.

5 CONCLUSION

In the present study a male voice was recorded in slightly reverberant conditions with a six-channel microphone array and then "binauralized" over headphones by simulating six virtual loudspeakers around the subject using non-individualized HRTFs. The resulting binaural stimulus was presented to inexperienced subjects with two different orientations (0° and 180°).

In one condition, head tracking was inactive and subjects were asked to keep their heads stationary. In another condition, head tracking was active and subjects were asked to make large head movements. Results showed that head tracking can be effective in improving externalization for naïve subjects (as previously found with experienced subjects) with persistent effects.

In a third condition, subjects' heads were stationary while the sound sources were automatically rotated around them mimicking head movements. Results showed that, for the direct source at 180°, source movements were nearly as effective as head movements to enhance externalization. At 0°, source movements also enhanced externalization, but to a lesser extent than head movements. These results still suggest that rotating the sound scene before listening to binaural content can be a relevant alternative if no head tracking is available.

6 ACKNOWLEDGMENT

The authors would like to thank Alexis Baskind, Thibaut Carpentier, Vincent Koehl, Julian Palacino, Mathieu Paquier, Claire Voirin, Olivier Warusfel, and all the subjects. This work was funded in part by the French FUI project BiLi ("Binaural Listening," www.bili-project.org, FUI-AAP14).

7 REFERENCES

[1] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos, "On the Improvement of Localization Accuracy with Non-Individualized HRTF-Based Sounds," *J. Audio Eng. Soc.*, vol. 60, pp. 821–830 (2012 Oct.).

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123 (1993). https://doi.org/10.1121/1.407089.

[3] S. M. Kim and W. Choi, "On the Externalization of Virtual Sound Images in Headphone Reproduction: A Wiener Filter Approach," *J. Acoust. Soc. Am.*, vol. 117, pp. 3657–3665 (2005). https://doi.org/10.1121/1.1921548.

[4] D. R. Begault and E. M. Wenzel, "Headphone Localization of Speech," *Hum. Fac. Erg. Soc.*, vol. 35, pp. 361–376 (1993). https://doi.org/10.1177/001872089303500210.

[5] G. Parseihian and B. F. G. Katz, "Rapid Head-Related Transfer Function Adaptation Using a Virtual Auditory Environment," *J. Acoust. Soc. Am.*, vol. 131, pp. 2948–2957 (2012). https://doi.org/10.1121/1.3687448.

[6] B. F. G. Katz and G. Parseihian, "Perceptually Based Head-Related Transfer Function Database Optimization," *J. Acoust. Soc. Am.*, vol. 131, pp. 99–105 (2012). https://doi.org/10.1121/1.3672641.

[7] D. R. 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, pp. 904–916 (2001 Oct.).

[8] 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, pp. 2841–2853 (1999). https://doi.org/10.1121/1.426899.

[9] R. L. Martin, K. I. McAnally, and M. A. Senova, "Free-Field Equivalent Localization of Virtual Audio," *J. Audio Eng. Soc.*, vol. 49, pp. 14–22 (2001 Jan./Feb.).

[10] C. Mendonça, G. Campos, P. Dias, and J. A. Santos, "Learning Auditory Space: Generalization and Long-Term Effects," *PloS One*, vol. 8, 2013. e77900. https://doi.org/10.1371/journal.pone.0077900.

[11] D. R. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *J. Audio Eng. Soc.*, vol. 40, pp. 895–904 (1992 Nov.).

[12] G. Plenge, "On the Differences between Localization and Lateralization," *J. Acoust. Soc. Am.*, vol. 56, pp. 944–951 (1974). https://doi.org/10.1121/1.1903353.

[13] N. Sakamoto, T. Gotoh, and Y. Kimura, "On 'Out-of-Head Localization' in Headphone Listening," *J. Audio Eng. Soc.*, vol. 24, pp. 710–716 (1976 Nov.).

[14] J. M. Loomis, C. Hebert, and J. G. Cicinelli, "Active Localization of Virtual Sounds," *J. Acoust. Soc. Am.*, vol. 88, pp. 1757–1764 (1990). https://doi.org/10.1121/1.400250.

[15] J. I. Kawaura, Y. Suzuki, F. Asano, and T. Sone, "Sound Localization in Headphone Reproduction by Simulating Transfer Functions from the Sound Source to the External Ear," *J. Acoust. Soc. Jpn.*, vol. 12, pp. 203–216 (1991). https://doi.org/10.1250/ast.12.203.

[16] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The Contribution of Head Movement to the Externalization and Internalization of Sounds," *PloS One*, vol. 8 (2013). e83068. https://doi.org/10.1371/journal.pone.0083068.

[17] E. M. Wenzel, "The Relative Contribution of Interaural Time and Magnitude Cues to Dynamic Sound Localization," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 80–83 (1995). https://doi.org/10.1109/ASPAA.1995.482963.

[18] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz, and C. de Boishéraud, "Influence of Head Tracking on the Externalization of Speech Stimuli for Non-Individualized Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 141, pp. 3678–3688 (2017). https://doi.org/10.1121/1.4978612.

[19] H. Dejardin and E. Ronciere, "nouvOson Website: How a Public Radio Broadcaster Makes Immersive Audio Accessible to the General Public," presented at the *AES 57th International Conference: The Future of Audio Entertainment Technology – Cinema, Television and the Internet* (2015 Mar.), conference paper 6-2.

[20] R. Nicol, L. Gros, C. Colomes, and J.-C. Messonnier, "Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation [Comparative Study of Several Spatial Audio Recording Setups after Binauralization]," *Proc. Acoustics 2016 Conf.,* Le Mans, France (2016).

[21] C. Faller, F. Menzer, and C. Tournery, "Binaural Audio with Relative and Pseudo Head Tracking," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9223.

[22] G. Wersényi, "Effect of Emulated Head-Tracking for Reducing Localization Errors in Virtual Audio Simulation," IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, pp. 247–252 (2009). https://doi.org/10.1109/TASL.2008.2006720.

[23] M. Williams, "Microphone Arrays for Natural Multiphony," presented at the *91st Convention of the Audio Engineering Society* (1991 Oct.), convention paper 3157.

[24] A. Baskind, T. Carpentier, M. Noisternig, O. Warusfel, and J.-M. Lyzwa, "Binaural and Transaural Spatialization Techniques in Multichannel 5.1 Production," *27th Tonmeistertagung, VDT Intl. Conv.* (2012).

[25] A. Baskind, "Hedrot, an Open-Source Head Tracker" (2016). https://abaskind.github.io/hedrot/.

[26] D. S. Brungart, A. J. Kordik, and B. D. Simpson, "Effects of Headtracker Latency in Virtual Aaudio Displays," *J. Audio Eng. Soc.*, vol. 54, pp. 32–44 (2006 Jan./Feb.).

[27] E. M. Wenzel, "Effect of Increasing System Latency on Localization of Virtual Sounds with Short and Long Duration," *Proc. 7th Intl. Conf. on Auditory Display (ICAD2001)*, pp. 185–190 (2001).

[28] O. Warusfel, "Listen HRTF Database" (2003). http://recherche.ircam.fr/equipes/salles/listen/.

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

[30] M. B. Gardner, "Proximity Image Effect in Sound Localization," *J. Acoust. Soc. Am.*, vol. 43, p. 163 (1968). https://doi.org/10.1121/1.1910747.

[31] D. H. Mershon, D. H. Desaulniers, T. L. Amerson, and S. A. Kiefer, "Visual Capture in Auditory Distance Perception: Proximity Image Effect Reconsidered," *J. Aud. Res.*, vol. 20, pp. 129–136 (1980).

[32] C. André, E. Corteel, J.-J. Embrechts, J. Verly, and B. F. G. Katz, "Subjective Evaluation of the Audiovisual Spatial Congruence in the Case of Stereoscopic-3D Video and Wave Field Synthesis," *Intl. J. Human-Computer Studies*, vol. 72, pp. 23–32 (2014 Jan.). https://doi.org/10.1016/j.ijhcs.2013.09.004.

[33] E. Hendrickx, M. Paquier, and V. Koehl, "Audiovisual Spatial Coherence for 2D and Stereoscopic-3D Movies," *J. Audio Eng. Soc.*, vol. 63, pp. 889–899 (2015 Nov.). https://doi.org/10.17743/jaes.2015.77.

THE AUTHORS

Etienne Hendrickx

Peter Stitt

Jean-Christophe Messonnier

Jean-Marc Lyzwa

Brian F.G. Katz

Catherine de Boishćraud

After a Ph.D. degree in psychoacoustics from the University of Brest, Etienne Hendrickx worked on binaural sound at the Conservatory of Paris and is now a teaching and research associate at the University of Brest. In 2013 he won a Lumiere Award from the International 3D Society for his studies on sound related to 3D movies. He is also a member of the French Acoustical Society, elected to the Board of the Sound Perception Group.

•

Peter Stitt received his Ph.D. in 2015 from Queen's University Belfast. His Ph.D. focused on the localization of phantom sources for listeners outside the sweet-spot with Ambisonics. He has since worked on the Binaural Listening (BiLi) research project at LIMSI-CNRS in Orsay, France. His research interests are in spatial audio reproduction, spatial perception, and perceptual modelling.

•

Jean-Christophe Messonnier studied sound engineering at ENS Louis-Lumière and acoustics at CNAM. He is now a sound engineer and a teacher at the Conservatory of Paris. His main research interests are objet-based audio recording methods and spatial audio.

•

Jean-Marc Lyzwa studied musicology and sound engineering at the University of Strasbourg. He is now a sound engineer and a teacher at the Conservatory of Paris. His main research interests are multichannel recording techniques and binaural/transaural sound.

•

Brian F.G. Katz is a CNRS Research Director at the Institute ∂'Alembert, UPMC/CNRS, and coordinator of the Sound & Space research theme. His fields of interest include spatial 3D audio rendering and perception and room acoustics. With a background in physics and philosophy, he obtained his Ph.D. in acoustics from Penn State in 1998 and his HDR in engineering sciences from UPMC in 2011. Before joining CNRS he worked for various acoustic consulting firms including Artec Consultants Inc., ARUP & Partners, and Kahle Acoustics. He has also worked at LIMSI-CNRS and IRCAM.

•

Catherine de Boishéraud has for ten years participated in the production of albums at Radio France, working with classical record labels such as Erato, EMI, Harmonia Mundi or Ocora. In 1989 she joined the recently created audiovisual department of the Conservatory of Paris, which engages actively in research into recording and broadcast techniques. She is now director of this department.