Externalization of virtual sounds using low computational cost spatialization algorithms for hearables

Vincent Grimaldi, Gilles Courtois, Laurent Simon, Hervé Lissek

EXTERNALIZATION OF VIRTUAL SOUNDS USING LOW COMPUTATIONAL COST SPATIALIZATION ALGORITHMS FOR HEARABLES

Vincent Grimaldi[1] **Gilles Courtois**[2] **Laurent S. R. Simon**[3]
Hervé Lissek[1]
[1] École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2] Sonova AG, Stäfa, Switzerland
[3] University of Zürich, Zürich, Switzerland
vincent.grimaldi@epfl.ch

ABSTRACT

In binaural sound reproduction, it has been shown that externalization improves the listening comfort. Using individualized binaural room impulse responses, it is possible to simulate sound sources in a given room for a listener wearing headphones. However, in some real-time binaural sound applications such as miniaturized hearables, it is not always possible to use such optimal filtering. This potentially results in the perception of virtual sources inside the head rather than externalized. Being able to be aware of surroundings in space, referred as spatial awareness, is another crucial feature in this type of application. This study assessed three sound spatialization algorithms that aim to optimize externalization while preserving spatial awareness. Those algorithms were designed to be implementable on wearable devices, using low computational power and little memory. These algorithms were evaluated in terms of externalization as well as spatial awareness. The results show that a convincing externalization can be achieved with those low computational cost algorithms while preserving spatial awareness.

1 INTRODUCTION

In certain applications, hearables can be used in combination with a remote microphone (RM). The first goal of this process is to optimize speech intelligibility in a challenging auditory situation such as a noisy environment or when the distance between the speaker and the listener is large. In most cases, the voice of a speaker is picked up by a microphone, and transmitted directly to the hearables. However, when the speech is played diotically, important binaural cues for sound localization such as interaural time differences (ITDs) and interaural level differences (ILDs) [1] are not available in this signal. Therefore the perceived location of the speech source does not match its physical location in the environment. Moreover, if only the speech source is played to the listener, safety concerns are raised, as sounds other than the main speech may not be heard by the listener, e.g. an alarm, another speaker or any unexpected event. This is referred to as spatial awareness, i.e. the listener's ability to be aware of themselves and of their surroundings in space.

In the field of hearing aids, where RM systems are often used, sound localization and spatialization have been topics of interest in several studies. In [2], it was shown that localization performances were better with lower gain on the RM signal. Methods to estimate the localization of the sound source have been proposed in [3, 4]. While successfully providing sound localization to the listeners, the spatialization method based on anechoic head-related transfer functions (HRTFs) proposed in [3] usually fails to provide an externalized sound image, i.e. sounds are perceived inside the head rather than surrounding the subject. In [5,6], the authors have shown that early reflections (ER) contribute largely to sound externalization. In the context of RM systems, several studies have demonstrated that the superimposition of ER to the RM signal can significantly improve externalization. Those studies used ER that were either extracted from the hearing device microphones signals [7, 8] or artificially synthesized [9].

This article describes a study in which three different algorithms in the context of RM systems for hearables were subjectively evaluated by a panel of normal-hearing listeners. The performance of the three algorithms was evaluated both in terms of externalization and spatial awareness.

2 DESCRIPTION OF THE ALGORITHMS

This section aims at providing a brief description of the three algorithms evaluated in this study

2.1 Remote Microphone + ambient Microphone (RMM)

This is the baseline implementation as depicted in Fig. 1.

This approach was introduced in [3] and consists of a generic spatialization, using minimum-phase 128-sample (5.8 ms) head-related room impulse responses (HRIRs), as measured on a Knowles' Electronic Manikin for Acoustic Research (KEMAR, G.R.A.S) in an anechoic chamber. The ITD is simulated by a pure delay. The input from the hearable microphones is superimposed with a certain gain

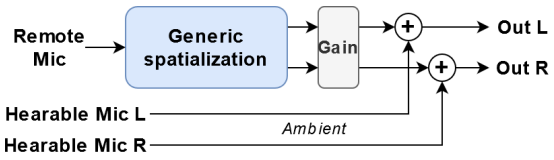

Figure 1. Block diagram of the RMM algorithm

to the main output. This gain depends on the estimated signal to noise ratio (SNR): high gain for high SNR and conversely. This implementation is designed for an optimal speech intelligibility.

2.2 Early Reflections extraction and Cleaning (ERC)

The first goal of this algorithm (Fig. 2) is to add ER picked up in the hearable microphone signals to the aforementioned generic spatialization. This is done using a proprietary algorithm which is designed to extract ER from the hearable microphone signals and clean the ambient noise.

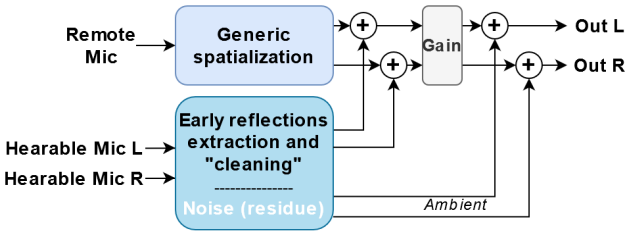

Figure 2. Block diagram of the ERC algorithm

Limiting the algorithm to the addition of the extracted ER would not be optimal for spatial awareness. Hence, a complementary filter is introduced, to bring back some of the ambient sound (called here: residue). Based on the filter computed by ERC module, a new filter is computed as:

$$A_{res}(\omega) = (1 - A_{erc}(\omega)) * G_{res} \quad (1)$$

Where A_{res} is a coefficient of the residue filter, A_{erc} is a coefficient of the ERC filter, ω is the frequency index and G_{res} is the general gain (between 0 and 1) applied to the residue part. A gain of 1 would lead to an original restitution of the signal (unprocessed). The RM signal, the ER and the ambient sound can be tuned with independent gains, allowing to optimize the trade-off between speech intelligibility, externalization and spatial awareness.

2.3 Partitioned Convolution (PConv)

This algorithm consists in introducing synthesized ER as depicted in Fig. 3. The ER are generated by using a uniform partitioned convolution algorithm [10], as implemented in [11]. It consists in partitioning an impulse response into a series of smaller blocks. Those blocks can be seen as separate impulse responses or subfilters that can be run in parallel in a real-time processing.

The RM signal was convolved with a 50-ms truncated pair of binaural room impulse responses (BRIRs). BRIRs

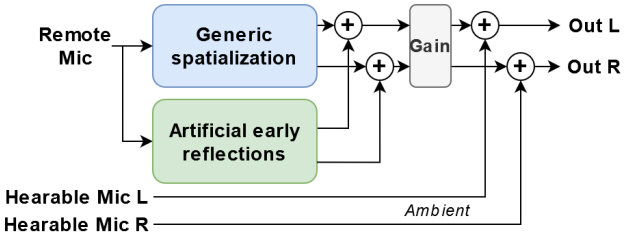

Figure 3. Block diagram of the PConv algorithm

measured in a classroom were used for this study, regardless of the room used for the experiment. Indeed, in a real usecase, BRIRs could not be measured in every room the hearable user walks in. A pair of BRIRs corresponding to a source at 0° and a distance of 2 m was used in the experiments. This limits the required computational cost and memory usage. To enable spatial awareness, an additional signal from the hearable microphones is superimposed. This implementation allows to tune independently the gain of the main speech (RM signal), the ER and the ambient sound.

3 EXPERIMENT 1: EXTERNALIZATION

The goal of this experiment was to study if the additional ER introduced in the ERC and PConv algorithms improve the perception of externalization.

3.1 Setup

The experimental setup was installed in a listening room (volume = 125 m^3, RT_{60} = 0.17 s). Three loudspeakers (Genelec 1029A), were placed at an azimuth of 30° on the right side of the listener, at a distance of 67 cm, 113 cm and 200 cm respectively from the listener's position, as depicted in Fig. 4. They were numbered from 1 to 3. During the test, the listener was sitting at the center of the room and had their head immobilized by means of a chin rest.

3.2 Stimuli

Five stimuli were evaluated in this experiment. One was a diotic reproduction (anchor), one was the reference, convolved with the individual BRIRs of loudspeaker 3, and the three remaining stimuli corresponded to the RMM, ERC, and PConv algorithms. The stimuli consisted of 15-second speech sequences, obtained by concatenating short phonetically-balanced random sentences from the French HINT database [1] .

In the baseline algorithm (RMM), the main path was set 10 dB louder than the ambient path. When adding ER to the direct signal, the direct-to-reverberant ratio (DRR) is an important parameter regarding externalization [12]. For this part of the experiment, the DRR was fixed to 5 dB for the ERC and PConv algorithms based on preliminary informal tests. All stimuli were presented at a level of

[1] Collège National d'Audioprothèse 2006

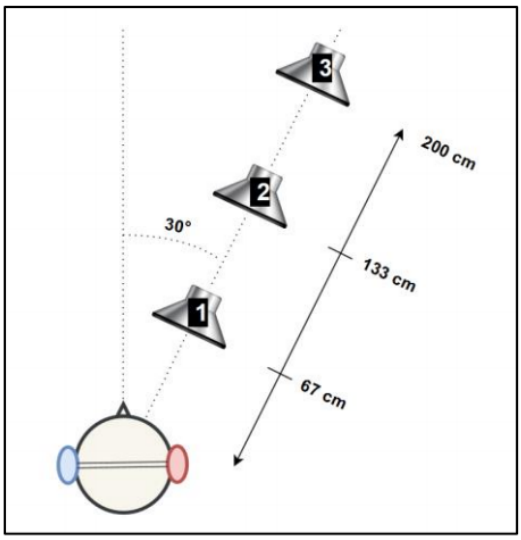

Figure 4. Schematic representation of the setup for Experiment 1

65 dB SPL. For all participants, the stimuli were compensated with their individual headphone-to-ear impulse responses (HPIRs) and low-pass filtered (cut-off frequency = 6.5 kHz).

3.3 Procedure

For every participant, individual BRIRs corresponding to the three loudspeakers, as well as HPIRs were measured using a pair of in-the-ear binaural microphones (Sound Professionals MS-TFB-2). During the experiment, all stimuli were played through a pair of open headphones (Audeze LCD-2C) driven by a headphones amplifier (Lake People HPA RS 02).

Using a MUSHRA-like (MUltiple Stimulus with Hidden Reference and Anchor [13]) graphical user interface (GUI) displayed on a touch-pad, the participants were asked to rate the auditory externalization perceived for the five stimuli. The stimuli were available simultaneously and it was possible for the listener to cycle through the stimuli and listen to them as many times as they wanted. They were instructed to answer the question: "How far do you perceive each stimulus from your position?". They used a continuous scale displayed as a slider with the following markers: Center of the head (0), Boundary of the head (20), At Loudpseaker 1 (40), At Loudpseaker 2 (60), At Loudpseaker 3 (80) and Further than Loudpseaker 3 (80 to 100). The task was repeated over 4 runs.

The experiment was preceded by a short training phase in which each listener could listen to speech sound examples processed with their individual BRIRs of loudspeakers 1, 2 and 3. This helped the participants to get accustomed with the task and gave them an a priori knowledge of the reproduction level used in the experiment. This also served to ensure that their auditory impression matched the visual location of the loudspeakers.

The allocation of the different tested stimuli to the "play" buttons was randomized across runs to avoid any bias due to the order of presentation of the various stimuli. It was also randomized across participants, i.e. every participant experienced different allocations during the experiments.

3.4 Results

Externalization ratings are reported in Fig. 5. For every listener, only the last three runs are considered. The first run was considered as training and therefore not taken into account. Three subjects (out of 25) were not retained in the results because they did not perceive the reference as externalized, or perceived the anchor as the furthest sound.

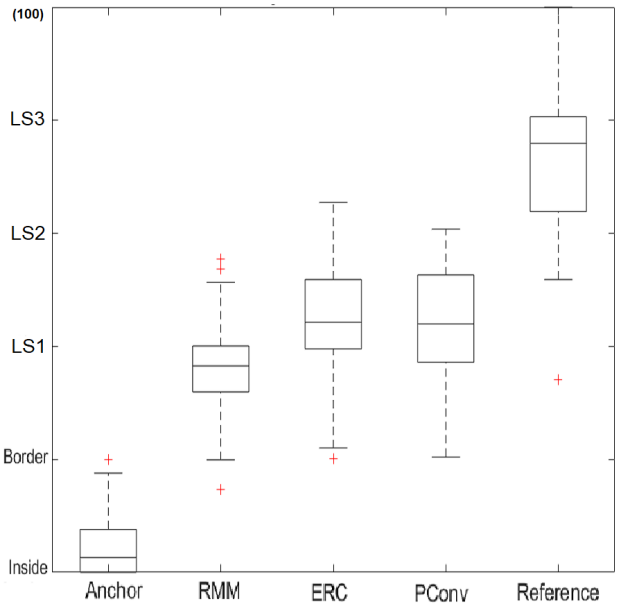

Figure 5. Degree of externalization evaluated on a continuous scale with the following markers : Inside of the head (0), Border of the head (20), At Loudpseaker (LS) 1 (40), At LS 2 (60), At LS 3 (80) and Further than LS 3 (80 to 100).

A Friedman test among repeated measures revealed significant differences in the perceived degree of externalization between the five stimuli, $\chi^2(4) = 226.65$, $p < 0.001$. A large effect size was found using Kendal's W: W = 0.859. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction were conducted. The results are reported in Tab. 1. The reference was perceived significantly more

	Anchor	RMM	ERC	PConv
RMM	> 0.001			
ERC	> 0.001	0.001		
PConv	> 0.001	0.001	1.000	
Reference	> 0.001	> 0.001	> 0.001	> 0.001

Table 1. Degree of externalization, Wilcoxon signed-rank tests results (significant p-values are in red)

externalized than the anchor, RMM, ERC and PConv stimuli. The anchor was perceived significantly more internal-

ized than the reference, RMM, ERC and PConv stimuli. No significant difference was observed in the externalization ratings between the ERC and PConv stimuli. Both the ERC and PConv stimuli were perceived as more externalized compared to the RMM stimuli.

It should be noted that the PConv algorithm used BRIRs from a classroom with a significantly different and denser reflection content than the ones of the listening room in which the test was conducted. Additionally, the BRIRs in the PConv algorithm corresponded to a source located at 0°. Despite this, the PConv algorithm gave a similar perceived degree of externalization compared to the ERC algorithm.

4 EXPERIMENT 2: SPATIAL AWARENESS

The second experiment aimed at studying if the ERC and PConv algorithms affect spatial awareness compared to the RMM strategy.

4.1 Setup

The test took place in the same room and with the same hardware as in Experiment 1. A complex auditory situation was presented over headphones. This auditory scene is schematically depicted in Fig. 6.

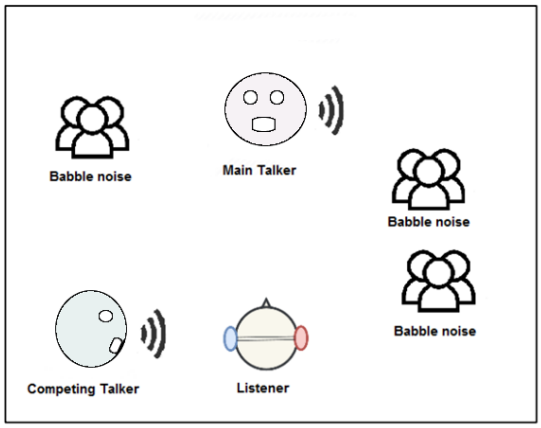

Figure 6. Auditory scene for Experiment 2, simulated over headphones using binaural synthesis

4.2 Stimuli

The stimuli consisted of complex auditory scenes including a main talker in the front (0°), conversational noise with a SNR of 0 dB at the listener's position, and a competing talker located on the right side of the listener (90°). Binaural synthesis was used to simulate the main and competing talker, using BRIRs for a source at 2 m, 0° in a listening room and 2 m, 90° in an anechoic room, respectively. Binaural recordings from an internal database were used for the babble noise. The stimuli were processed using the full implementation of the PConv and ERC algorithms as well as the RMM strategy. The ERC algorithm was presented with two settings: one with a DRR of 5 dB and one of 2 dB. Similarly, the PConv algorithm was presented in two versions, one with a DRR of 5 dB and one of 2 dB. The RMM algorithm is used with the same setting as in Experiment 1: the RM signal was amplified by 10 dB compared to the ambient path. Those algorithms are denoted: ERC5, ERC2, PConv5, PConv2 and RMM respectively. All stimuli were presented at a level of 65 dB SPL.

4.3 Procedure

The listener was asked to complete a dual-task performance. The listener had to repeat one of the short consecutive sentences pronounced by the main talker (the one followed by a "beep"). Simultaneously, they had to pay attention to the continuous speech of the competing talker and click a button on the screen when one segment was presented backwards (time-reversed). In the example depicted in Fig. 7, the time-reversed segment was necessarily overlapping with the sentences of the main talker. For any sequence, the sentence to repeat could be anywhere between the 3rd and the 6th sentence, and the time-reverse speech could happen at any time between the first main talker sentence and the sentence to repeat. The test included 10 repe-

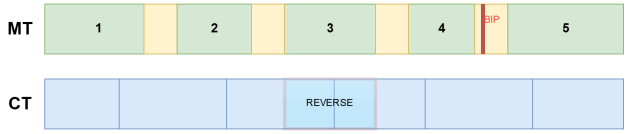

Figure 7. Example of stimuli structure in time for Experiment 2 (MT: main talker, CT: competing talker)

titions for each algorithm and parameterization. Thus, the listeners had to evaluate a total of 50 stimuli. The order of the sentences and the order of the algorithms applied to each sentence were randomized across participants. A short training preceded the test in order to familiarize the listener with the task and the stimuli.

4.4 Results: number of detections of the time-reverse speech

The number of times a time-reverse speech signal was correctly detected was analyzed. Detection before it was presented or after the sentence to repeat was considered as a non-detection. Boxplots for the scores of the various algorithms are reported in Fig. 8. One listener who presented outlier ratings was removed from the analysis to ensure normality of the distribution. Using a repeated measures ANOVA, no significant difference was found in the scores of the various algorithms regarding the number of time-reverse speech detections, $F(4, 92) = 1.571$, $p = 0.189$.

The number of filling sentences presented between sentences to repeat ranged from 2 to 5. This could yield a "tension" effect affecting the performance in detection. To assess the possibility of such an effect, a repeated measure ANOVA was run, looking at the interaction between

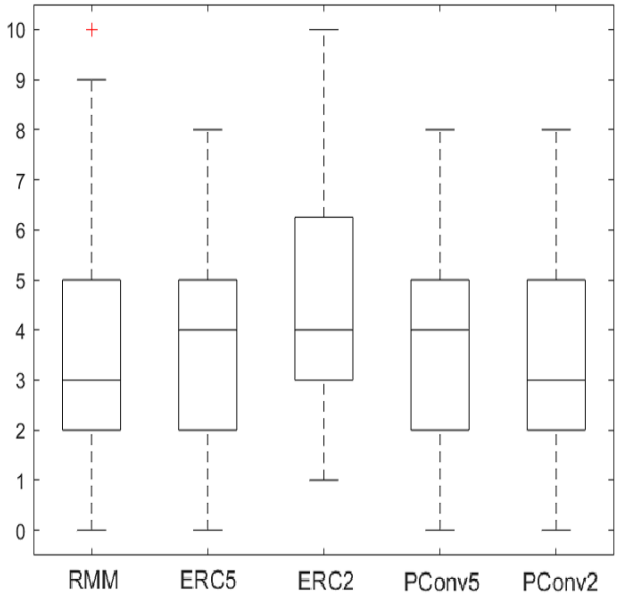

Figure 8. Scores for the number of detections

the tension (number of filling sentences) and the detection scores. No significant interaction was found, $F(4, 192) = 0.726$, $p = 0.575$.

The speech intelligibility task could be considered as a control task, which had a very high success rate for every participant and algorithm, hence results are not reported in this article for the sake of conciseness.

5 CONCLUSION

In this work, two new algorithms were proposed to improve the perception of externalization for hearables with RM, while preserving spatial awareness and speech intelligibility. Both algorithms allow to independently tune the gain applied to the RM signal, the ER and the ambient sound. This enables to look for an optimal trade-off between speech intelligibility, auditory externalization and spatial awareness.

Two experiments were conducted and showed that a significant improvement in the perceived degree of externalization was obtained with the ERC and Pconv algorithms compared to the RMM algorithm. The proposed DSP strategies, ERC and Pconv, do not affect spatial awareness.

To follow up, similar tests should be conducted with hearing impaired subjects, to assess the generalization of the results to the specific case of hearing aids.

Aknowledgements

This research was partly funded by the Swiss Innovation Promotion Agency Innosuisse with grant number 25760.1 PFLS-LS, and partly by Sonova AG, Stäfa, Switzerland.

6 REFERENCES

[1] J. Blauert. Spatial Hearing the Psychophysics of Human Sound Localization, 2nd ed.; *The MIT Press*: Cambridge, MA, USA, 1995.

[2] J. G. Selby, A. Weisser, E. MacDonald. Influence of a remote microphone on localization with hearing aids. In *Proceedings of the International Symposium on Auditory and Audiological Research*, Vol. 6. The Danavox Jubilee Foundation. 2017.

[3] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande. Binaural hearing aids with wireless microphone systems including speaker localization and spatialization. In *Proc. of the 138th Audio Eng. Soc. Convention*, Warsaw, Poland, 2015.

[4] M. Farmani, M. S. Pedersen, and J. Jensen. Sound source localization for hearing aid applications using wireless microphones In *2018 IEEE 10th Sensor Array and Multichannel SignalProcessingWorkshop(SAM)*, IEEE, pp.455–459, 2018.

[5] W.M. Hartmann, A. Wittenberg. On the externalization of sound images. In *J. Acoust. Soc. Am.* 99(6), pp 3678–3688, 1996.

[6] A.W. Bronkhorst, T. Houtgast. Auditory distance perception in rooms. In *Nature*, 397(6719), pp 517–520, 1999.

[7] G. Courtois, V. Grimaldi, H. Lissek, P. Estoppey, E. Georganti. Perception of Auditory Distance in Normal-Hearing and Moderate-to-Profound Hearing Impaired Listeners. In *Trends in Hearing*, Vol. 23, pp 1–18, 2019.

[8] V. Grimaldi, H. Lissek, G. Courtois, E. Georganti, P. Estoppey. Auditory externalization in hearing-impaired listeners with remote microphone systems for hearing aids. In *Proc. of the 26th ICSV Conv.*, Montreal, Canada, 2019.

[9] J.M. Kates, K.H. Arehart. Integrating a remote microphone with hearing-aid processing. In *J. Acoust. Soc. Am.* 145(6), pp 3551-3566. 2019.

[10] W.G. Gardner. Efficient convolution without input-output delay. In *J. Audio. Eng. Soc.* 43(3)3, pp. 127–136, 1995.

[11] A. Torger, A. Farina. Real-time partitioned convolution for Ambiophonics surround sound In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, pp 1-4, 2001.

[12] P. Zahorik, D.S. Brungart, W.A. Bronkhorst WA. Auditory Distance Perception in Humans: A summary of Past and Present Research. In *Acta Acust. united Ac.* 91(3), pp 409–420. 2005.

[13] ITU-R recommendation BS.1534-3, 2015.