Source distance modelling in the context of Audio Augmented Reality

Vincent Martin, Isabelle Viaud-Delmon, Olivier Warusfel

SOURCE DISTANCE MODELLING IN THE CONTEXT OF AUGMENTED REALITY

Vincent Martin[1] **Isabelle Viaud-Delmon**[1] **Olivier Warusfel**[1]

[1] Sciences et Technologies de la Musique et du Son, STMS-Lab
Sorbonne Université, IRCAM, CNRS,
F-75004 Paris, France

vmartin@ircam.fr

ABSTRACT

Audio-only Augmented Reality (AAR) consists in enhancing a real environment with virtual sound events. A seamless integration of the virtual events within the environment requires processing them with artificial spatialization and reverberation effects that simulate the acoustic properties of the room. It is known that the room effect has a strong influence on the perception of sound source distance. Consequently, in the context of AAR, an important issue is to guarantee that the perceived distance of the virtual events is coherent with the room geometry and with the intended position of the virtual sound sources. This calls for the choice of a distance rendering model and the access to a priori information about the acoustical properties of the room.

This study focuses on the perceptual evaluation of a room effect rendering model informed by a single Spatial Room Impulse Responses (SRIR). The model simulate the missing positions/distances by generating a new SRIR through modifications of the temporal envelope measured one. The perceptual performances of the proposed model is comparable to performances observed on reference situations based on Binaural Room Impulse Responses (BRIR) measured at each reconstructed distance.

1. INTRODUCTION

Audio-only augmented reality (AAR) consists in adding virtual sound events to the real environment. The concept of AAR can be used in practice to: associate digital sounds to physical objects, blend digital sounds and acoustic sounds (sound reinforcement ...), populate the environment with virtual sound events, overlay audio information onto the real world (museum tours, navigation . . .), or achieve telepresence (musical instruments, conversation. . .). The aspiration of such applications is that the virtual sound sources are seamlessly integrated in the real environment, which calls for applying spatialization processing to the virtual auditory events. More specifically, the reverberation effect applied to the virtual sound events should resemble that originating from the sounds emitted by real acoustical sources. It is known that the reverberation effect will contribute significantly to the perceived location of the virtual sound events. Thus, the reverberation processing needs to be carefully designed to guarantee that the perceived location of a virtual event conforms to the intended one. The addressed problem is twofold, a) selecting an appropriate spatialization model to control the location of the virtual auditory events and the associated room effect, b) accessing a priori information about the acoustical or architectural properties of the real environment to tune the model accordingly. The choice of the spatialization model will have a direct influence on the reproduction of acoustic cues conveyed by the room effect. It will affect the spatial perception of a sound source and, more generally, contribute to the perceptual representation of the global sound scene.

Among the different spatial dimensions, the present paper focuses on the distance perception of virtual events situated in front of the listener. Numerous studies have been dedicated to the identification of both auditory and non-auditory cues that affect the auditory distance perception [1–4]. However, the localization performance of a listener in a particular AAR context may be difficult to predict. Many factors may contribute to the perceived distance of the virtual auditory events. Concurrent real acoustical sources may give useful cues to gauge their relative depth. The auditory distance perception may be enhanced or biased by spatially congruent or discrepant visual anchors [5, 6]. The vision of the real environment may also elicit acoustical expectations of the listener [7] and influence the perceived distance.

The objective of this study is to evaluate whether different sound distance rendering models in the context of a simple AAR scenario generate differences in distance perception in terms of accuracy. The study focuses on models that produce sound spatialization with limited a priori acoustical information, i.e. provided by a single impulse response measured in the environment. Participants are requested to judge the distance of a frontal source. The mean perceived distance compared to the actual position of the source as well as intra and inter-subject variations are compared.

2. PERCEPTION OF SOUND SOURCE DISTANCE

On average, listeners underestimate the distance of distant sources ($>$ 10m) and overestimate the distance of nearby sources ($<$ 1m). These systematic biases originate from the interpretation of multiple cues to provide a distance information of a sound source. The perceived auditory distance depends on two different types of cues whose presence and reliability depends of the stimulus and properties of the environment [2,8]: acoustic cues and cognitive cues. The overall sound level is a relative distance cue. The perceived distance generally increases with decreasing sound level. However, the distance judgement cannot rely only on this cue since a variation of the level can be induced by the sound intensity emitted by the source itself. The presence of reverberation is also an important cue for distance judgement, more precisely the Direct-to-reverberant ratio (DRR) has been demonstrated to provide absolute distance information [3, 6, 9]. Different definitions of the DRR are proposed in the literature, the common hypothesis being that the human brain cannot separate perfectly the direct sound from early reflections. Subsequently, the DRR is generally calculated using an integration time window for the direct sound, which length depends on authors and studies: 2.5ms [10], or 7ms, [7]. Bronkhorst [3] suggested that the calculation of the transition time separating what is referred to as the extended direct sound and the reverberation could be made according to the time profile of the interaural time differences (ITD). The direct energy would integrate all incoming sounds with ITDs that differ less than a given value.

The sound spectrum yields additional distance cues. For large propagation distances ($> 15m$), the frequency dependent air absorption becomes noticeable, with increasing attenuation at high frequencies [11]. For sound sources in the near-field ($< 2m$), head shadowing provides important monaural as well as binaural spectral cues [4, 12].

Subjective cues also contribute to distance estimation, and can vary greatly from a listener to the other. Familiarity with the sound source and the sound environment, expectation according to the semantic content or the nature of the sound source contributes to the interpretation of acoustic cues providing distance information. Visual information can also affect the auditory distance judgements, through a visual capture effect [13]. More generally, the presence of congruent visual information tends to be beneficial to the accuracy of auditory distance judgements [14]. Besides the analysis of the bias between the perceived distance of the auditory event and the actual distance of the source, studies have also revealed large intra- and inter-individual variations [2, 15].

Zahorik [5] suggests that a compressive power function is a good approximation to most psychophysical functions, auditory distance included. The function takes the form:

$$D = k * d^a \tag{1}$$

D is the perceived auditory distance, d is the sound source distance, k the linear compression (< 1) or expansion (> 1) coefficient (equivalent to the intercept when the perceived distance is represented on a logarithmic scale) and a the non-linear compression coefficient ($a < 1$, equivalent to the slope when the perceived distance is represented on a logarithmic scale).

In this study, the aim was to simulate a simple AAR situation where the listener is located in a real environment enhanced with a virtual sound source. This context is used to evaluate the performances of different auditory distance rendering models in terms of perceptual accuracy and variability. The participants were seating eyes opened in a damped room. Therefore, they had access to a global visual information about their environment. However, no visual cues were spatially or semantically congruent with the virtual auditory sources. The auditory stimuli were not rendered so that they could be perceived as emanating from a specific visual anchor in the environment. The stimuli were speech samples rendered via binaural reproduction on headphones. We were expecting that the visual access to the global environment would constraint the auditory distance judgements (i.e. distance range limited by the walls), and that accuracy of auditory distance judgements would be improved by the use of speech stimuli [16].

3. DISTANCE RENDERING MODELS

The seamless integration of a virtual sound source in the real environment calls for the implementation of a spatial audio processing framework. The spatialization engine shall simulate the acoustic properties of the environment and allow for real-time control of the virtual source location, e.g. according to the movement of the listener or of the source. In the context of AAR, the virtual sound scene will generally be played back using hear-through headphones and shall be ideally rendered in binaural format using the listener's individual HRTFs. Several approaches exist for implementing such a spatial audio processor. They rely on different room acoustic representations and require different a priori information about the real environment.

Numerical modeling of sound propagation (e.g., image source model, beam tracing, radiosity etc.) can be applied given the geometrical description of the real room as well as the sound absorption and diffusion coefficients of its surface materials [17]. This so-called auralization approach, is powerful but requires substantial computer resources to render and update the virtual sound scene in real time according to the current positions of the source and listener [18]. Algorithmic reverberation synthesis, based on feedback delay network, may provide a more efficient solution [19]. In this approach, the parametric room effect can be tuned according to a prior characterization of the real room, using conventional acoustic criteria. However, most of these criteria show a strong dependence according to the relative position and orientation of the listener and the source. Therefore, using such an approach calls for the implementation of a distance rendering model that automatically modifies the room effect parameters to provide a plausible auditory distance impression [20]. Although efficient, this approach may suffer from an over simplification

of the room effect characterization, which will possibly miss perceptually important effects linked to a particular spatial and temporal pattern of the early reflections. Alternatively, one could use real-time convolution processing that exploit one or several room impulse responses measured in the real environment. Today, the availability of spherical microphone arrays (SMA) allows to exploit high-spatial resolution Spatial Room Impulse Responses (SRIR) that will faithfully reproduce the actual room effect and can be further decoded on headphones using a set of individual HRTFs. As for the above mentioned FDN approach, a distance rendering model is needed to control how a measured SRIR shall be reshaped in order to convey the desired distance impression.

The present study investigates the latter approach in the specific case where the room acoustics is characterized by a single SRIR measured at a given source-receiver distance. Two different distance rendering models are tested to alter the measured SRIR in order to continuously control the perceived auditory distance of a virtual sound source.

3.1 Initial Impulse Response

The experimental room was characterized with a single SRIR measured with an Eigenmike® EM32 at a source-receiver distance of $1\,\mathrm{m}$. The motivation for choosing to characterize the room with a SMA was the perspectives offered by this solution. While a BRIR, recorded for instance with a dummy head, would only permits modifications in the time domain, SRIR can be subject to spatial modifications, which could influence the auditory distance perception [3, 21]. Moreover, such SRIR can also be decoded using individualized HRTFs. However, for this first study, only time domain transformations were applied and no individual adaptation was proposed during the listening test. Hence, the initial SRIR was first decoded into binaural format using a Neumann KU100 dummy head HRTFs.

From this initial BRIR, 9 others BRIR corresponding to increasing distances were calculated with two models described here after.

3.2 Envelope-based model

The first model is based on a simplified representation of the energy envelope of the impulse response, here divided into two temporal segments: the early energy E_s comprising both the direct sound and early reflections, and the late reverberation energy E_{rev} (Fig. 1).

Several perspectives can be considered to demarcate the time limit between these two segments. This transition time can be linked to perceptual cues, regardless of the room geometry. The clarity of speech (C_{50}) considers a transition time of $50ms$ between the early and late energy. In order to delineate the 'useful sound' in opposition to the 'detrimental sound', Lochner and Burger consider a weighting function equal to 1 until $35ms$ and then linearly decreasing up to $95ms$ [22].

This transition time can also refer to the physical properties of the room and of its impulse responses. After

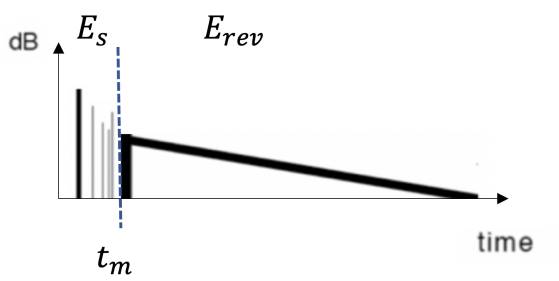

Figure 1. Mixing time is considered as the boundary separating early energy from late reverberation.

reaching a sufficiently high echo density and modal overlap, the room reverberation exhibits an exponentially decaying stochastic behaviour. This lower time limit is referred to as the mixing time. Several estimators of the mixing time have been suggested in the literature. The estimation $t_m = \sqrt{V}$ (with t_m the mixing time in ms and V the volume of the room in m^3) was proposed in [23]. Other methods rely on the evaluation of the diffuseness of the sound field from the statistics of the echos observed in the impulse response. This estimation may either be conducted in the time domain [24, 25] or in the spatial domain, when a spatial room impulse response (SRIR) is available [26].

The model applied to alter the initial BRIR to control its apparent source distance is inspired from previous work described in [20]. In the proposed approach, the sound source distance is driven through the control of two entities, the direct sound energy E_{dir} and the reverberated energy E_{rev}.

The level of the direct sound according to the source distance d may be expressed as follows:

$$E_{dir}(f,d) = S_\phi(f)^2 \frac{\mu(f)^d}{4\pi c d^2} \tag{2}$$

with c the sound celerity, f the frequency, $\mu(f)$ the frequency dependant sound absorption for $1\,\mathrm{m}$ propagation in the air, and S_ϕ(f) the free field transfer function of the source in the direction ϕ of the receiver.

The level of the reverberation after the time τ in the impulse response may be expressed as follows:

$$E_{rev}(\tau,f) = \frac{Tr(f)}{13.81}\frac{S_d(f)}{V} e^{-13.81\frac{(\tau+d/c)}{Tr(f)}} \tag{3}$$

with $Tr(f)$ the reverberation time, V the volume of the room, $S_d(f)$ the diffuse-field transfer function. The dependence of the reverberation energy E_{rev} with the distance d agrees with Barron's revised theory on energy relations in the room response [27].

In the present study, some further simplifications are made. The frequency dependence of the reverberation time and the air absorption are not considered. The spatial dependence of the free field transfer function S_ϕ(f) is also neglected as the source will always be heard from its frontal direction. Moreover, the attenuation law of the direct sound E_{dir} will be extended to the whole early energy

E_s. Under these assumptions, and using equations (2) and (3), the gains that should be applied to the early and late segments of the initial impulse response measured at a distance d_{ref} to derive the new impulse response at distance d can be written as follows :

$$E_s(d) = E_s(d_{ref}) * \frac{{d_{ref}}^2}{d^2} \tag{4}$$

$$E_{rev}(d) = E_{rev}(d_{ref}) e^{-13.81 \frac{(d-d_{ref})}{cTr}} \tag{5}$$

In the following, this method will be referred to as the "envelope-based model".

3.3 Intensity-based model

The second model, extrapolates the impulse responses corresponding to different distances by applying a global gain to the initial BRIR. For each distance, the gain was tuned so that the loudness of the speech stimulus, convolved with the transformed BRIR, corresponded to the loudness of the stimulus generated with the BRIR measured at the same location in the experimental room (see §3.4). The loudness criteria used here is EBU R128. This model will be referred to as the "intensity-based model". This second model is tested to quantify the importance of room related acoustical cues apart from the perceived loudness.

3.4 Reference

In order to serve as a reference, ten SRIRs were measured in the experimental room at ten different distances in front of the source, using the Eigenmike® EM32 SMA. The source-microphone distances were ranging from 1 to 7 m (1 , 1.5 , 2 , 2.5 , 3 , 3.5 , 4 , 5 , 6 , 7 m). These SRIRs were converted from a 4th order ambisonic signal to BRIRs, using the HRTF of the Neumann KU100 dummy head. These ten resulting BRIRs will be referred to as "reference" in the following. As described in §3.1, the BRIR measured at 1m served as the initial impulse response from which the nine other BRIR were extrapolated using the two above described models.

3.5 Objective comparison

The experimental room is a classroom at IRCAM, (semi-damped, dimensions: 8.7 x 4.7 x 3.5 m^3 – L x W x H, $RT60$ at 1kHz of 0.55s). The geometrical estimation of the mixing time corresponding to this room is 12 ms. The time estimation of the diffuseness from the analysis of the reference BRIRs ranged from 15 ms (time domain analysis) to 25 ms (spatial domain analysis). The mixing time that was considered for the study was fixed to $15ms$.

The differences between the two models and the reference for each source distance are depicted on Fig. 2. For the envelop-based model, the main differences occur for the early energy E_s which becomes significantly underestimated as the distance increases. This behavior comes from the $1/d^2$ attenuation law that was applied uniformly to the whole section. In contrast, the late energy E_{rev} tends to exceed that of the reference as the distance increases. Conversely, the intensity model shows a significant underestimation of the late energy E_{rev} and a slight overestimation of the early energy E_s as the source distance increases.

Both models show a slight underestimation of the total energy compared to the reference. This underestimation increases with distance but remain lower than 2dB.

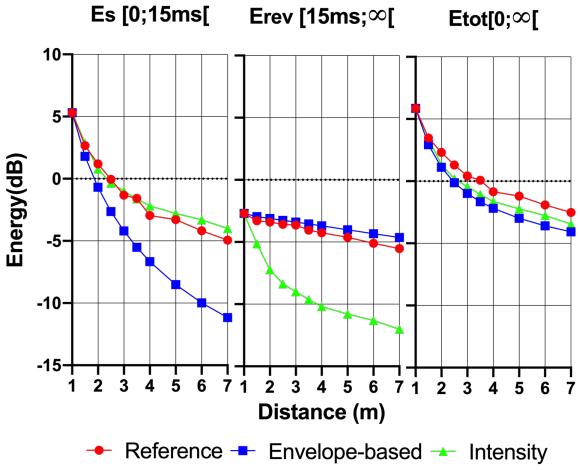

Figure 2. Evolution of the early energy E_s (left), reverberation E_{rev} (middle) and total energy E_{tot} (right) according to the source distance for the two models and the reference.

4. MATERIAL AND METHODS

4.1 Participants

A total of 20 (8 women) participants, age ranging from 19 to 35 (mean age: 26.05), took part in the study. An initial interview insured that none of them had hearing problems or uncorrected vision. All the participants had little to no knowledge about audio processing or room acoustics. Informed written consent was signed by each participant before the undergoing of the experiment's procedure.

4.2 Listening room

The listening tests were conducted in the classroom used for the measurement session. The participant was seating at the place corresponding to the location of the SMA. Twelve chairs were distributed every 45 cm along a line facing the participant at 0° azimuth and starting 1.2 m from the participant position, in order to guide the visual distance perception. However, the participant was informed that the spatial distribution of the chairs did not correspond to the actual spatial distribution of the measured and modeled stimuli. A visual fixation cross was drawn at a height of 1.2 m on the wall facing the participant, who was asked to look at it during the playback of the stimuli.

4.3 Stimuli

The original stimulus was a 5-second anechoic recording of a sentence in Swiss German pronounced by a male

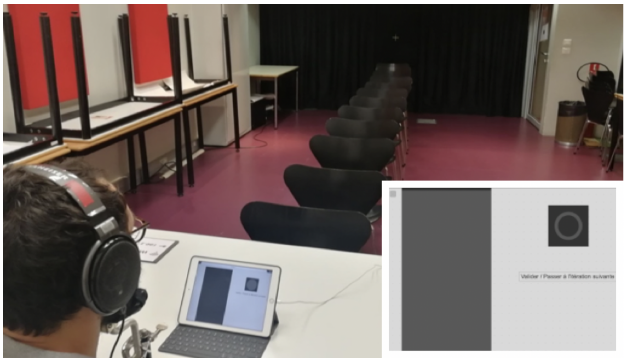

Figure 3. Configuration of the room and GUI.

speaker. The choice of the language, unknown to all participants, was made to avoid focusing on the semantic content of the sentence. The stimuli were generated by convolving the speech sample with each of the measured or modeled BRIR. The playback level was set by calibrating the level of the speech stimulus convolved with the reference BRIR measured at $1\,\mathrm{m}$. For this BRIR, the stimulus was reproduced with headphones placed on the Neumann KU100 dummy head, and the level was adjusted to match that of a standard male speaking standing at $1\,\mathrm{m}$ in front of the dummy head (68dB (LAeq)).

Stimuli were rendered through circumaural open headphones (Sennheiser HD 650), no head-tracking system was used. The participant's head was immobilised using a chin rest during the trials, to prevent inadvertent movements.

4.4 Perceived auditory distance collection method

The participant reported the perceived sound source distance with a graphical slider presented on a touchscreen tablet. This method was used to prevent bias from personal interpretation of an absolute scale in meters. The software *MAX/MSP* was used for the rendering of the stimuli, the creation of the graphical interface and the data collection.

4.5 Procedure

The participant was given the tablet and was introduced to the graphical interface used for reporting distance judgements. It was explained that the minimum of the slider corresponds to the participant's position and the maximum to the back wall. After an indication of the expected duration of the experiment (1 hour), the participant started a training session of 27 stimuli, composed of all the different possible conditions (9 distances x (2 model + reference)). The goal of the training was to familiarize the participant with the distance reporting method and to ensure that the procedure was understood. After the training, the experiment was divided into three blocks, each of them containing 81 stimuli. Each stimulus was repeated 3 times within each block. The order of the stimuli within a block was randomized. During the trials, the participant could trigger the stimulus playback when she/he wanted but it was played only once. The trial response was collected through the graphical interface given to the participant. A final questionnaire was filled at the end of the 3 blocks to collect additional information related to the localization of the source (externalization, direction), realism, problems with the interface, global attention of the participant, noticeable differences between the different stimuli apart from the distance.

5. RESULTS

5.1 General results

Initial attention was focused on the normalization of the responses of each participant on every condition (9 distances x (2 models + reference)) in order to identify possible statistical outliers. A Jarque Bera-test indicated that one participant showed non normal distribution of the responses for a majority of the tested conditions. This participant was excluded from the pool of subjects. The following analysis is based on the data of 19 subjects. To analyze the performance of each model and reference, the geometric mean of the perceived distance for each subject on each conditions was computed. For comparison purposes, the mean perceived distance considered here, results from a linear conversion of reports made by subjects on the visual analogic scale (0% on the slider corresponding to $0\,\mathrm{m}$, and 100% to $7\,\mathrm{m}$).

$$D_g(d) = \prod_{k=1}^{n} \sqrt[n]{D_k(d)} \tag{6}$$

(D_g) : Geometric mean perceived distance for a sound source at a distance d, n the number of subjects. This mean was used because it is admitted that distances are perceived following a power function [2].

A repeated measures one-way ANOVA applied to the geometric mean distances of each subject was carried out, with the within-subject factors DISTANCE (10 levels from 1 to 7m) and MODEL (3 levels: 2 models and the reference). The main effect DISTANCE was significant ($F(1,8) = 283,08$, $p < 0.01$, $Partial - \eta^2 = 0,9402$) as well as the MODEL ($F(1,2) = 12,87$,$p < 0.01$, $Partial - \eta^2 = 0.4168$) and the interaction DISTANCE × MODEL ($F(1,15) = 5,34$, $p < 0.01$, $Partial - \eta^2 = 0.2287$). The analysis of the reference conditions confirms that the perceptual distance is globally overestimated for short distances, here from 1 to $5\,\mathrm{m}$ and underestimated for longer distances. This behaviour is also observed for the envelope-based model. For the intensity model, the perceived distance is always underestimated although close to the actual distance between 2 and $3\,\mathrm{m}$.

The similarity of the enveloped-based model and the reference was further investigated using a post-hoc analysis (Fisher LSD). For each distance conditions, no significant differences between the reference and the envelope-based model were found. Besides, this post-hoc analysis reveals the presence of a distance beyond which no significant effect of the distance can be observed anymore. This auditory horizon appears at $5\,\mathrm{m}$ for the envelope-based model and the reference and at $4\,\mathrm{m}$ for the intensity-based model.

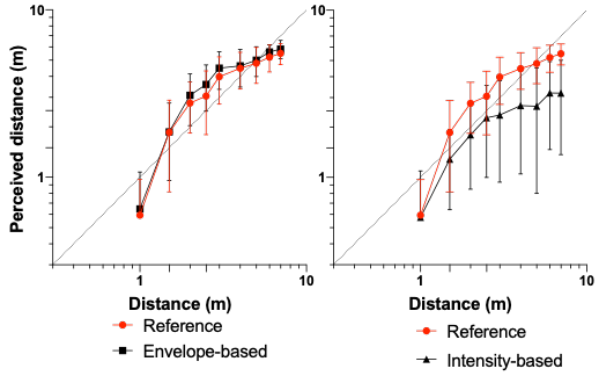

Figure 4. Geometric mean of perceived distances according to the model used to generate the sound stimuli (27 conditions : (2 models + reference) x 9 distances). Error bars correspond to the standard deviation over the geometric mean perceived distance of all subjects (N=19).

5.2 Individual results

The general results following the ANOVA demonstrate a similarity between the reference and the envelope-based model. The similarity between the envelope-based model and the reference can also be found at an individual level. Zahorik's compression law Eqn. (1) was fitted following a linear regression model (with k equivalent to the intercept, and a to the slope) to the logarithm of the geometric mean perceived distance, for each subject on every models (19 subjects, 3 models).

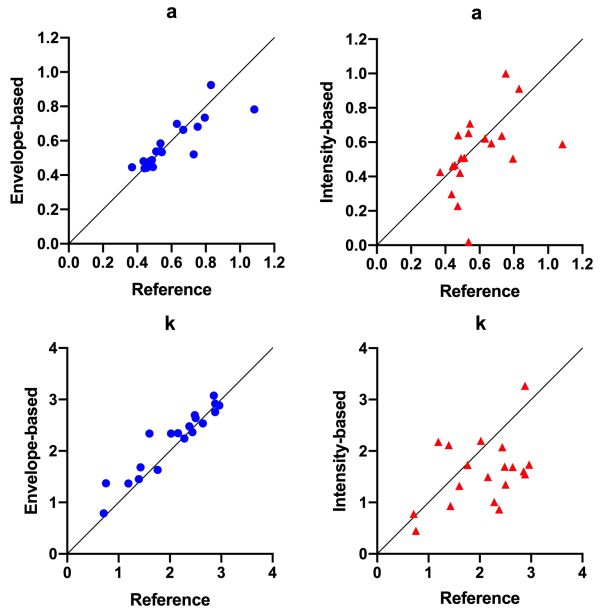

Figure 5. Comparison of the individual fitting parameters a (above) and k (below), between the reference and the models. (left) Envelope-based, (right) Intensity.

The distribution of dots in Fig. 5 confirms the presence of a similarity between the envelope-based model and the reference for each subject, in terms of auditory distance perception. In contrast, the distribution of both parameters for the intensity-based model is more dispersed, showing no correlation at an individual level. The result for the intensity-based model also highlights that most non-linear compression coefficients a are lower than those obtained for the reference. The graphical user interface had a limited range, which could introduce a bias in the collected perceived distances. Thus, for trials corresponding to distances from $5\,\mathrm{to}\,7\,\mathrm{m}$, the normality of the responses was affected. Hence, no further analysis on the individual variability (intra-subject) could be conducted.

6. DISCUSSION

The study aimed at evaluating the performances of models, based on the measure of a single impulse response, to control the distance of a virtual sound source in a context of AAR applications. Distance perception was assessed on the basis of accuracy, quantified with the mean values of fit parameters, and inter-subject variability (standard deviations).

6.1 Evaluation of the accuracy and variability of the envelope-based model

Despite objective differences illustrated in Fig. 2, the results indicate a general similarity between the envelope-based model and the reference, both in terms of mean perceived distance and inter-subject variability. These conclusions are based on the results illustrated in Fig. 4 and from the power function fit parameters obtained for the two models and the reference (see Fig. 5). In contrast, the intensity-based model is less accurate with lower non-linear coefficient a and higher inter-subject variability. For the envelope-based model as well as for the reference, the perceived distance reaches its horizon beyond $5\,\mathrm{m}$. However, considering the experimental conditions it cannot be concluded whether this limit was driven by the size of the room or by the indication given to the participants that the maximum of the slider was corresponding to the distant wall and fixation cross.

Results also indicated a similarity between the reference and the envelope-based model on the basis of individual data. The perceived distance profile of each subject can be represented by its individual Zahorik's power function parameters [6]. Fig. 5 shows that for the envelope-based model and the reference each subject's fit parameters are close. This indicates that regardless of individual perceptual profiles, the reference and the envelope-based model can be considered as similar. In contrast, the intensity-based model shows the presence of different perceptual profiles among the subjects. Subjects mainly basing their auditory distance judgement on intensity will show a perceptual similarity between the intensity-based model and the reference. Previous studies illustrate the fact that listeners weight differently the acoustic cues to produce an auditory distance judgement [6], especially the direct-to-reverberant ratio. It can be assumed that the more a subject relies on cues other than intensity, the less she/he can differentiate the different stimuli generated by the intensity-

	Calcagno [14]	Anderson [15] Audio Condition	Anderson AV Condition	Zahorik [2]	Ref.	Energy Distr.	Intensity
k	1.14 +- 0.12	2.22 +- 1.99	1.38 +- 0.91	1.32	1.24 +-0.50	1.16 +-0.49	1.75 +-1.22
a	0.89 +- 0.06	0.61 +- 0.30	0.87 +- 0.27	0.54	0.85 +-0.24	0.87 +-0.33	0.83 +-0.55
R^2	n/a	0.64 +- 0.22	0.84 +- 0.18	0.91	0.79 +-0.1	0.78 +-0.08	0.71 +-0.24
Sound rendering	Speaker	Binaural		n/a	Binaural		
Stimuli content	Noise Bursts	Gaussian Noise		n/a	Speech		
Room type	Semi-reverberant	Concert hall		n/a	Damped room		
Distance Range	2-6m	0.3-9.8m		n/a	1-7m		
Report method	Verbal	Verbal		Verbal	Visual analogic scale		

Table 1. Mean values of fitting parameters with standard deviation and R^2 reported in different studies.

based model. Thus, the more she/he tends to underestimates larger distances, producing smaller non-linear compression coefficient (see Fig. 5).

6.2 Influence of the perceptual context of the study

To evaluate the influence of the specific perceptual context of the current study on accuracy and variability, the results are compared with previous studies using similar protocols. Tab. 1 compares the mean value of the Zahorik's power fit function parameters obtained in several studies : a meta-analysis realized by Zahorik [2] over 81 different studies dedicated to the perception of auditory distance, a study from Anderson [15] comparing auditory distance judgements in audio-only and audio-visual condition and a study from Calcagno [14] studying auditory distance judgements in presence of visual cues. For the two models and the reference of the current study the non-linear compression coefficient a is shown to be closer to 1, compared to the values obtained in the meta-analysis realized by Zahorik. The linear coefficient k is also closer to 1 for the reference and envelope-based model. These values can be interpreted as a better global accuracy in distance judgements in our case study. This enhancement of the auditory distance perception accuracy obtained for the reference and envelope-based model can be caused by the presence of visual cues and use of speech, while the meta-analysis ran by Zahorik [2] is mainly based on studies implying blind auditory distance judgements and various type of stimuli (from noise bursts to speech signal).

With both coefficients closer to 1, the results are similar to what can be found in other experiments involving visual cues [14, 15]. Results in terms of accuracy are consistent with a study of Calcagno [14]. One of the protocol studied in their paper is actually similar to the current study (see Tab. 1), as the subjects could have access to the visual configuration of the room prior to the experiment. Although their experiment was using a real loudspeaker, in contrast with our study which uses binaural reproduction on headphones, the mean fit parameters are very similar. However, the comparison in terms of variability must be done cautiously considering the difference in fitting methods (Calcagno uses fitting on raw data with a least square method instead of using linear fitting on logarithmic data) and number of trials per condition (3 instead of 9).

Accuracy results are also consistent with those obtained in the audio-visual condition of the study of Anderson & Zahorik [15], although the visual capture effects are different. In their study, each auditory stimulus condition is associated with a simultaneous projection of a loudspeaker image at the same location, whereas in our study the distribution of the visual anchors (chairs) does not coincide with the auditory stimuli. The analysis of inter-subject variability can be made by comparison of the standard deviation calculated on the fit parameters. Results indicate lower inter-subject variability in our study. However, the use of a different scale by Anderson & Zahorik (verbal report in meters), and the inherent noise induced by this method could led to an overestimation of the real perceptual noise. Contrarily, the report method used in our study induces an underestimation of inter and intra-subject variability for distances beyond $5\,\mathrm{m}$. To conclude, as it was expected, the presence of visual cues and the use of speech signal enhanced the accuracy and reduced the variability of auditory distance reports [6, 8], even if conclusions and possible comparison on inter-subject variability have to be done cautiously considering the protocol used.

7. CONCLUSIONS

The study shown the possibility to emulate efficiently an auditory distance effect using a single BRIR measured in the room and modified using a simplified model of the energy envelop. In this particular case, it has been demonstrated to be as efficient as the reference based on measured BRIRs, in terms of perceptual accuracy. The comparison with previous studies also shows that in the context of audio-only augmented reality applications, the perceptual conditions (presence of visual cues, suggestion that the source stays within the room) enhances the accuracy of the auditory distance judgements. In further studies, this model will be refined in order to be generalised to larger rooms as well as more complex geometries, taking full advantage of the use of SRIR to analyse the acoustic envi-

ronment. More specifically, the intent is to provide better insights on the role of the spatial distribution of early reflections on the auditory distance perception.

8. REFERENCES

[1] A. J. Kolarik, B. C. 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," *Attention, Perception, & Psychophysics*, vol. 78, no. 2, pp. 373–395, 2016.

[2] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *ACTA Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[3] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, 1999.

[4] N. Kopčo and B. G. Shinn-Cunningham, "Effect of stimulus spectrum on distance perception for nearby sources," *JASA*, vol. 130, no. 3, pp. 1530–1541, 2011.

[5] P. Zahorik, "Estimating sound source distance with and without vision," *Optometry and vision science*, vol. 78, no. 5, pp. 270–275, 2001.

[6] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *JASA*, vol. 111, no. 4, 2002.

[7] D. L. Valente and J. Braasch, "Subjective expectation adjustments of early-to-late reverberant energy ratio and reverberation time to match visual environmental cues of a musical performance," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 840–855, 2008.

[8] J. M. Loomis, R. L. Klatzky, J. W. Philbeck, and R. G. Golledge, "Assessing auditory distance perception using perceptually directed action," *Perception & Psychophysics*, vol. 60, no. 6, pp. 966–980, 1998.

[9] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975.

[10] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *JASA*, vol. 112, no. 5, pp. 2110–2117, 2002.

[11] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[12] D. S. Brungart, "Control of perceived distance in virtual audio displays," in *Proc. of the 20th Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Society. Vol. 20 Biomedical Eng. Towards the Year 2000 and Beyond*, vol. 3, IEEE, 1998.

[13] D. H. Warren, R. B. Welch, and T. J. McCarthy, "The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses," *Perception & Psychophysics*, vol. 30, no. 6, pp. 557–564, 1981.

[14] E. R. Calcagno, E. L. Abregu, M. C. Eguía, and R. Vergara, "The role of vision in auditory distance perception," *Perception*, vol. 41, no. 2, pp. 175–192, 2012.

[15] P. W. Anderson and P. Zahorik, "Auditory/visual distance estimation: accuracy and variability," *Frontiers in psychology*, vol. 5, p. 1097, 2014.

[16] P. Zahorik, "Auditory display of sound source distance," in *Proc. Int. Conf. on Auditory Display*, pp. 326–332, 2002.

[17] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *JASA*, vol. 138, no. 2, pp. 708–730, 2015.

[18] D. Poirier-Quinot, M. Noisternig, and B. F. Katz, "Evertims: Open source framework for real-time auralization in vr," in *Proc. of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, pp. 1–5, 2017.

[19] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty years of ircam spat: looking back, looking forward," in *41st ICMC*, 2015.

[20] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engineering Society Convention 103*, 1997.

[21] L. Wallmeier and L. Wiegrebe, "Ranging in human sonar: effects of additional early reflections and exploratory head movements," *PloS one*, vol. 9, no. 12, p. e115363, 2014.

[22] P. Lochner and J. Burger, "The subjective masking of short time delayed echoes by their primarysounds and their contribution to the intelligibility of speech," *Acustica*, vol. 8, no. 1, 1958.

[23] J.-D. Polack, "Modifying chambers to play billiards: the foundations of reverberation theory," *Acta Acustica united with Acustica*, vol. 76, no. 6, pp. 256–272, 1992.

[24] J. S. Abel and P. Huang, "A simple, robust measure of reverberation echo density," in *Proc. of AES convention 121*, 2006.

[25] R. Stewart and M. Sandler, "Statistical measures of early reflections of room impulse responses," in *Proc. of the 10th Int. Conf. DAFx-07 France*, 2007.

[26] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "A robust denoising process for spatial room impulse responses with diffuse reverberation tails," *JASA*, vol. 147, no. 4, 2020.

[27] M. Barron and L.-J. Lee, "Energy relations in concert auditoriums," *JASA*, vol. 84, no. 2, 1988.