Necessary attributes for integrating a virtual source in an acoustic scenario

Yu, Wangyang; Kleijn, W. Bastiaan

DOI
10.1109/IWAENC.2018.8521276

Publication date
2018

Document Version
Final published version

Published in
16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018 - Proceedings

NECESSARY ATTRIBUTES FOR INTEGRATING A VIRTUAL SOURCE IN AN ACOUSTIC SCENARIO

Wangyang Yu* *W. Bastiaan Kleijn*†*

* EEMCS, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
† ECS, Victoria University of Wellington, Kelburn, Wellington 6012, New Zealand

ABSTRACT

We investigate what information about a room is necessary to integrate a new source into an existing scenario. In particular, we consider the effects of the reflection order, the order of ambisonics signals and reverberation time. We conducted a series of listening tests and used the control variates method to determine the quantitative relevance of the selected attributes. In terms of integration and accurate localisation, at least third order ambisonics description of a source, is required for integration of that source. In addition, a finite number of early reflections can perform equally well to a full room impulse response when a new source is integrated into an existing scenario. However, the room impulse response with only the correct reverberation time is not sufficient.

Index Terms— Ambisonics, localisation, reflection order, order of ambisonics signals, reverberation time.

1. INTRODUCTION

Head-set based virtual reality (VR) is a specific immersive audio-visual environment that simulates a user's physical presence in an artificial scenario with corresponding VR headsets. Virtual reality will play an increasingly important role in numerous aspects of daily life, such as entertainment, education and health care. Spatial audio aims to create a 3D audio experience, which is an important component for a believable VR system.

Our goal is to examine what information about a room is necessary to integrate a new source into an existing acoustic scene. This knowledge will allow us to synthesize a realistic, convincing audio component. We are not aware of existing work on the problem. To understand the integration problem better, we first review the composition of a head-set based VR system. We will discuss accurate environment simulation and soundfield reproduction separately.

An accurate environment simulation is essential for perceptually acceptable sound in a VR system. To model the acoustics environment, we need to consider several physical attributes of sounds in a room, such as reflections and reverberation time. The image-source method is used to model reflections in a room [1, 2]. However, the computational load increases with an increasing number of reflection walls and it can only handle convex room shapes [3]. The high complexity of modelling reflections in acoustics environments makes efficient methods important [4–6]. Reverberation time, RT_{60}, is the time that the sound drops 60 dB below the original level [1]. Reverberation time is considered to be an important attribute in acoustic environment simulation. Several methods [7–9] exist to estimate the reverberation time.

Besides accurate environment simulation, a high quality soundfield reproduction system is of great importance. Ambisonics [10–12] has become the de-facto standard representation for VR systems. Ambisonics is particularly suitable for VR systems as head rotations are easily modelled as the rotation of sound fields in the spherical harmonics domain. It describes the sound field by means of a small set of temporal signals. In recent work ambisonics often refers to higher order ambisonics, which includes more signals than the four that are used in the original method as developed by Gerzon [11]. With an ambisonics representation of sufficient order, a high quality binaural audio rendering system can give listeners a realistic spatial audio experience. Hence it allows us to demonstrate our work on spatial audio. A number of techniques can be used for binaural rendering of ambisonics [13–15].

The main contribution of our paper is that we investigate how one can integrate a new source into an existing immersive environment with finite information of the environment. We study what is required to make a new sound source integrate into an acoustic scene so that people can perceive the new source as a natural component of the acoustic scene and in the correct direction. In this work we assume the head is in a fixed location. Through listening tests, we found at least third order ambisonics is required to integrate a new source. In addition, a finite number of early reflections can perform equally well to a full room impulse response when a new source is added to an existing scenario. However, only correct reverberation time is not sufficient.

The paper is organised as follows. In section 2, we describe our hypothesis of the integration of virtual objects in an acoustic scene. In section 3, we discuss our experiments in detail and analyse the results. Finally, we conclude our paper in section 4.

978-1-5386-8151-0/18/$31.00 ©2018 IEEE

2. INTEGRATING A VIRTUAL SOURCE

When we describe a soundfield, what is the necessary information of a room to make the sound natural and believable? We focus on the acoustics-only scenario, which implies that we omit the visual part of VR systems. The specific problem that we study here is the integration of a new sound source into an existing acoustic scenario. In a VR system, we already have an immersive environment. When we want to add a new source, like a virtual cat, we want to know what is required to make the new source perceptually plausible.

We study what aspects we can hear when we make specific modifications to a given acoustic scene. There exists a set of possibly relevant perceived attributes of a sound source in a room, such as room geometry and direct path direction to the source. In this paper, we focus on the order of ambisonics signals, reflection order, and reverberation time. When we consider the reverberation time, we also take the direct path distance, direct path direction and room size into account.

In the first subsection, we first briefly review ambisonics. We then discuss the selected attributes for integration in the second subsection.

2.1. Ambisonics

As ambisonics is the standard tool to reproduce a soundfield, we use it as the basis for our study. Ambisonics represents the sound field for the so-called interior case, where all sources lie outside the region of interest. Let the temporal frequency be denoted by $k = \frac{\omega}{c}$, where ω is frequency in rad/s and c is the speed of sound. Furthermore, let p represent the sound signal pressure at any point in space. Then the sound field can be represented by

$$p(r, \theta, \phi, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} B_n^m(k) j_n(kr) Y_n^m(\theta, \phi), \qquad (1)$$

where the $B_n^m(k)$ are the *ambisonics coefficients*, $j_n(kr)$ is the spherical Bessel function of the first kind and $Y_n^m(\theta, \phi)$ are the spherical harmonics, which are defined as in [16].

We can model a real source as a point source approximately and describe the spherical harmonic expansion of a point source [16]. By mode matching, we can derive the *ambisonics coefficients* $B_n^m(k)$. For the specific case of a single primary point source at location (r_q, θ_q, ϕ_q),

$$B_n^m(k) = -jkh_n^{(2)}(kr_q) Y_n^{m*}(\theta_q, \phi_q) S_q(k), \qquad (2)$$

where $h_n^{(2)}(kr_q)$ is the n'th spherical Hankel function of the second kind, and $S_q(k)$ denotes the driving signal.

2.2. Attributes for integration

We seek the necessary information to integrate a new sound source into an existing acoustic scenario. We are interested in reflection order, the order of ambisonics signals and reverberation time, which we will discuss separately below.

An important question is what order of ambisonics signals is necessary to make the integration of a new sound object believable. The most commonly used ambisonics signals are first-order ambisonics signals and third-order ambisonics signals. For head related transfer functions, [18] shows that an ambisonics order as low as four is sufficient, which indicates people do not perceive fine details during listening. Does this suggest that we do not need ambisonics signals of high order, such as order seven, to reproduce the soundfield? However, it is reasonable to explore the accuracy of the commonly used first order ambisonics and third order ambisonics. When (1) is truncated to a particular N, the sound field will be accurate within a spherical region near the origin, which is commonly called the *sweet zone*. Using $\mathcal{D}_R^{3D}$ to denote the dimensionality of three-dimensional ambisonics signals, we have $\mathcal{D}_R^{3D} = (N + 1)^2$. Furthermore, let R denote the radius of the sweet zone and f denote the frequency of the signal. Then we arrive at [19],

$$\mathcal{D}_R^{3D} = (\lceil \frac{4\pi R}{\lambda} \rceil + 1)^2 \approx 73R^2 \frac{f^2}{c^2}. \qquad (3)$$

Consequently, for third order ambisonics signals, if we assume the diameter of our head is 0.1m, the sound is correctly rendered at our ears up to 1600 Hz, which is too small comparing with the human hearing range. In addition, lower order ambisonics signals results in low angular resolution of soundfield reproduction. Is first-order or third-order enough for a believable VR system? Our hypothesis is that ambisonics signals of lower than order three are not sufficient for a believable VR system.

An important question with respect to reflections is whether we can use direct sound and a finite number of early reflections to replace the room impulse response to make a new sound source integrate into an existing acoustic scene. With an increasing number of reflections, the computational load of room impulse response increases [3]. Since real-time soundfield reproduction is required for a VR system, the computational load is a significant problem although efficient algorithms exist [4–6]. The room impulse response is composed of direct-direction sound, early reflections, and late reverberation. Early reflections are relatively sparse first echoes and influence the spatial impression [20, 21]. Late reverberation is a dense decayed succession of echoes [22] and can degrade automatic speech recognition [23]. It is unclear if the late reverberation makes a difference when we integrate a new sound source into an existing scenario. Our hypothesis is that we can use direct sound and a finite number of reflections to replace the room impulse response and still obtain perceptually acceptable integration.

Reverberation time is considered to be one of the important attributes in acoustic environment simulation. We study the question if this measure is sufficient for the integration. It is commonly quantified in the form of Sabine's formula:

$$RT_{60} = \frac{24ln10}{c_{20}} \frac{V}{Sa} \approx 0.1611\mathrm{sm}^{-1} \frac{V}{Sa}, \qquad (4)$$

where c_{20} is the speed of the sound in the room for 20 degrees Celsius, V is the room volume, S is the total surface area of the room and a is the average absorption coefficient of room surfaces. From (4), reverberation time is related to the room volume, surface area, surface absorption, and direct path length. However, it does not vary with the positions of the sources and listeners [24].

If we only have correct reverberation time when we integrate a new source, is it sufficient? We divide the problem into two categories to examine the room volume, surface area, and direct path length. Firstly, for a fixed reverberation time and fixed room geometry, we want to know if different positions affect the integration. We assume we have one room impulse response of a room, which is generated with a fixed reverberation time. If we use this room impulse response, we only replace the direct path with the true direct path and keep other pathways fixed, is it perceptually acceptable for a VR system? Moreover, is the distance or the direction of the direct path important? Our hypothesis is that a room impulse response with a correct reverberation time and a correct direct path is sufficient to integrate a new sound source. Secondly, for a fixed reverberation time, we are interested if listenerscan hear the effect of different room sizes. We hypothesise that listeners can hear the difference in the different room sizes.

3. EXPERIMENTS

We conducted listening tests to answer the questions asked in section 2.2. We used the control variates method to determine the quantitative relevance of the above selected attributes and used statistical analysis to analyse the experimental results. We first describe our experimental setup in section 3.1. We then present our experimental results and finally discuss these results.

3.1. Experimental setup

In this subsection, we give a general description of our experiments. Each artificial scenario lasts for ten seconds. In each scenario, there was one woman speaking in an empty rectangular room for four seconds. Then we added another woman as a new source to speak in this scenario, which lasts for six seconds and whose location is chosen randomly.

We choose the room size to be $6 \times 4 \times 3$m and the acoustic environment was modelled by the image-source method [2]. We used the room impulse generator of [25] for our experiment. The speed of sound was set to $c = 342$ m/s. The reverberation time RT_{60} was set to be 0.4 s. We used HRTFs from MIT Media Lab [26]. The headphone used for the listening test was Beyerdynamic$^{\text{TM}}$ DT 990 pro.

Ambisonics signals of order nine were used to reproduce the soundfield as a reference. We first resampled the input wav file with 16 kHz. After resampling, we constructed a four times oversampled Gabor frame and applied square-root Hann windows to satisfy the condition of perfect reconstruction. Based on the stationarity of the source signal and the length of room impulse, we chose a window support of 32 ms, which corresponds to 512 samples.

We used the commonly used audio rendering technique. We simulated playback over a given physical loudspeaker array, where each virtual loudspeaker signal is filtered with appropriately adjusted head related transfer functions (HRTFs) [17]. In our experiments, 598 secondary sources were used, the layout of which was same as that used for the HRTF database . We assumed the radius of a human head is 0.1 m and the center of the listener's head was located at $(3, 2, 1.7)$.

There were twelve participants for the experiments, which included two women and ten men. The subjects were not experts in spatial audio. Test subjects were allowed to listen to each scenario multiple times and change the volume in between. The experiments lasted approximately 30 minutes overall. The subjects answered questions for 16 scenarios. For each scenario they were required to answer if the new source is in the same scenario in the reference scenario and point out the azimuth and elevation of the new source with our user interface (the angular resolution is 10 degrees).

3.2. Description of Experiments

We conducted three sets of experiments to examine the three selected attributes, i.e., reflection order, the order of ambisonics signals, and reverberation time. We describe these three sets of experiments in detail below.

Our first experiment aimed to examine the relationship between the integration quality and the order of ambisonics signals. The reference scenario was reproduced with ninth-order ambisonics signals. The new source to be added to the scenario was reproduced with ambisonics signals of order one, three, five, seven, and nine respectively.

Our second experiment examined the influence of reflection order. In the reference scenario, the length of the room impulse response was set to be 340 ms, which included the early reflections and the late reverberation. To simplify the notation, we refer to the 340 ms room impulse response as *full response*. To examine the necessary reflection order, we changed the reflection order of the new sound source as zero, one, five, and nine. In addition, the full response was added as a contrast.

Our third experiment aimed at studying reverberation time. We first computed one room impulse response with the predefined reverberation time and a random position in the room, which is referred as the measurement point later. We assumed the measurement point is 1 m distant from the listener.

We only changed the direct path signal in room impulse responses according to the source positions. Four modified room impulse responses were used to convolve with the new source at four different positions. Two of the positions (position 1 and 2) are at the same direct path distance as the measurement point (1 m) but with two different direct path directions. One (position 3) is nearer to the listener than the measurement point (0.7 m) and one (position 4) is farther (1.4 m).

In addition, we investigated if room impulse response with correct reverberation time and incorrect room size can integrate a new sound source into an existing scenario. Hence, we changed the size of the room to 4m × 2m × 3m and to 8m × 6m × 3m and computed the corresponding room impulse responses.

3.3. Statistical analysis

We used the chi-square test to investigate if each test object is sufficient for integration. The full response case is our reference for integration. Since eight out of 12 people answered "yes" to this full response case, our null hypothesis is the source is considered to be integrated into the existing scenario where we expect eight out of 12 people answered "yes". The critical value is 2.706 with level of significance $\alpha = 0.10$ of a 1 degree of freedom test. When the computed value exceed the critical value, we can reject the null hypothesis. Consequently, if there are less than six out of 12 test subjects who answered "yes", we can claim that the corresponding information is not sufficient in terms of integration.

3.4. Experimental result and discussion

In this subsection, we present our experimental results. The experimental results of integration problem is shown in Figure 1, where we show the number of "yes" responses for each case and the error bar represents the Wilson scored interval for a 95% confidence interval. In addition to the integration, we are also interested in the localisation accuracy when a new source is integrated into an existing scenario. The mean absolute error is shown in Figure 2 and the error bar represents the standard deviation.

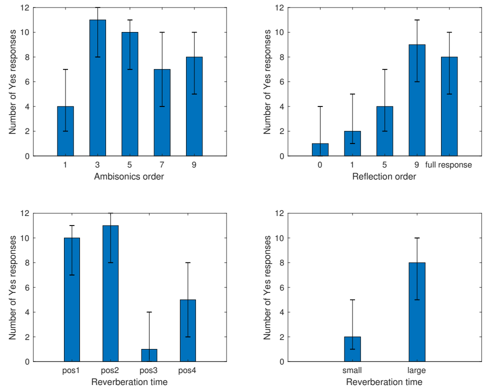

Fig. 1: Integration experimental result.

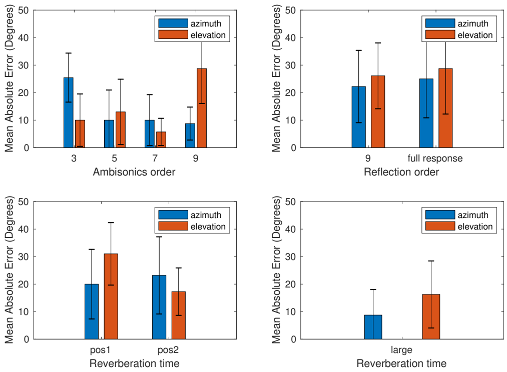

Fig. 2: Localisation accuracy.

When we observe the experimental results of the order of ambisonics, in terms of integration, ambisonics signals of order three to nine are sufficient to reproduce the sound field. We can conclude that an ambisonics order as low as three is sufficient for integration. Ambisonics of order nine shows lower elevation localisation accuracy than ambisonics of order five and seven, which may result from that late reverberation is clearer with ninth-order ambisonics signals and it can reduce the localisation accuracy.

As for the reflection order, we conclude that if a new source is integrated into an existing scenario, reflection order nine or full response is sufficient. In addition, reflection order nine shows approximately equal localisation accuracy as full response. We found that localisation accuracy depends on source location. While we not consider this effect in the present paper, this explains the differences in the ambisonics and reflection order experiments. To conclude, a finite number of reflections can replace the full room impulse response in terms of integration.

When we observe the experimental result of reverberation time, we conclude that a room impulse response with only correct reverberation time is not sufficient to guarantee good integration. Only with the same direct path distance, the source is perceived to be in the same scenario. Similar to the reflection order experiments, we claim that listeners can approximately point out the correct direction of the integrated new source. Combining this result with the results of a preliminary suggests that when the room size is larger than the reference room but smaller than twice reference room size, listeners perceive the new sound source as integrated into the existing scenario and the localisation is also relatively accurate.

4. CONCLUSION

In this paper, we used ambisonics signals to reproduce soundfield. We conducted a series of listening tests to examine the necessary information to integrate a new sound source into an existing acoustic scene and analysed the accuracy of localisation. We arrive at three conclusions. Firstly, with ambisonics signals of order three or higher, a new source can be integrated into an existing scenario. Secondly, a finite number of early reflections, for example ninth order reflection, can perform equally well in terms of integration and localisation as full room impulse responses. Finally, only using correct reverberation time to generate room impulse responses is not sufficient for integration and accurate localisation. To add a new source into an existing scenario, more information is required, such as direct path distance.

5. REFERENCES

[1] H. Kuttruff, *Room acoustics*, CRC Press, New York, 2014.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[3] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[4] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 187–187, 2007.

[5] J. Yan and W. B. Kleijn, "Fast simulation method for room impulse responses based on the mirror image source assumption," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2016, pp. 1–5.

[6] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, Aug 2010.

[7] M. Lee and J. H. Chang, "Blind estimation of reverberation time using deep neural network," in *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, Sept 2016, pp. 308–311.

[8] N. Faraji, S. M. Ahadi, and H. Sheikhzadeh, "Reverberation time estimation based on a model for the power spectral density of reverberant speech," in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug 2016, pp. 1453–1457.

[9] F. Lim, P. A. Naylor, M. R. P. Thomas, and I. J. Tashev, "Acoustic blur kernel with sliding window for blind estimation of reverberation time," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2015, pp. 1–5.

[10] D. Jerome and M. Sebastien, "Further study of sound field coding with higher order ambisonics," in *Audio Engineering Society Convention 116*, May 2004.

[11] F. Hollerweger, "An introduction to higher order ambisonic," 2013.

[12] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of Audio Engineering Society*, pp. 1004–1025, 2005.

[13] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 385–388.

[14] W. B. Kleijn, A. Allen, J. Skoglund, and F. Lim, "Incoherent idempotent ambisonics rendering," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 209–213.

[15] W. B. Kleijn, "Directional emphasis in ambisonics," *ArXiv e-prints*, Mar. 2018.

[16] J. Ahrens, *Analytic Methods of Sound Field Synthesis*, Springer Science & Business Media, Berlin, 2012.

[17] J. G. Tylka and E. Choueiri, "Comparison of techniques for binaural navigation of higher-order ambisonic soundfields," in *Audio Engineering Society Convention 139*, Oct 2015.

[18] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 921–930, Aug 2015.

[19] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2542–2556, June 2007.

[20] A. Warzybok, J.Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, 2013.

[21] D. R. Begault, B. U. McClain, and M. R. Anderson, "Early reflection thresholds for virtual sound sources," in *Proc. 2001 Int. Workshop on Spatial Media*, 2001.

[22] M. Karjalainen and H. Jarvelainen, "More about this reverberation science: Perceptually good late reverberation," in *Audio Engineering Society Convention 111*. Audio Engineering Society, 2001.

[23] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[24] W. B. Joyce, "Sabines reverberation time and ergodic auditoriums," *The Journal of the Acoustical Society of America*, vol. 58, no. 3, pp. 643–655, 1975.

[25] International Audio Laboratories Erlangen, "RIR generator," 2014.

[26] MIT Media Lab, "HRTF measurements of a KEMAR dummy-head microphone," 1994.