Spatial Audio Reconstruction for VR Applications using a Combined Method based on SIRR and RSAO approaches

Atiyeh Alinaghi
Electronics and Computer Science (ECS)
University of Southampton
Southampton, UK
a.alinaghi@soton.ac.uk

Luca Remaggi
London, UK
lucaremaggi@gmail.com

Hansung Kim
Electronics and Computer Science (ECS)
University of Southampton
Southampton, UK
h.kim@soton.ac.uk

2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP) | 979-8-3503-3893-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/MMSP59012.2023.10337683

***Abstract*—In order to recreate a sound field as realistic and immersive as the original setting, it is important to preserve the properties of the room acoustics. The room acoustics are usually captured by room impulse responses (RIRs) which can be employed to generate sound that could lead to the same audio perception. In this paper, we compare and combine two main parametric approaches, Spatial Impulse Response Rendering (SIRR) and Reverberant Spatial Audio Object (RSAO) to encode and render the RIRs. We first discuss that SIRR synthesises the early reflections more precisely whereas RSAO is a better approach to render the reverberation. For our proposed combined method, each RIR is divided into three parts: The direct sound, early reflections, and late reverberation, as in RSAO. To estimate the boundary time between the early reflections and late reverberation, we employ the diffuseness factor as in SIRR and set the mixing time as the time when the diffuseness factor reaches a threshold. Then the early part is analysed and synthesised by SIRR while the reverberant tail is encoded and rendered using RSAO. We show that in general, the combined method benefits the positive aspects of the two baseline methods.**

***Index Terms*—spatial audio, room impulse response, virtual reality, parametric methods**

I. Introduction

Spatial audio has numerous practical applications from entertainment like video games and film industries to spatial music mixing and navigational aids [1]–[3]. It is also one of the building blocks of virtual reality (VR) to accompany the vision to make the experience more natural and immersive [4]–[6].

The main challenges for creating an immersive spatial audio experience are how to record the main features of the sound field easily, transmit them efficiently, edit them if required, and reproduce the sound field with general rendering systems.

There are three different types of approaches to represent an audio scene, known as channel-based audio (CBA) [7], scene-based audio [7], [8], and object-based audio [9], which are all recognized by ITU and supported by MPEG-H audio standard [10]. In CBA the loudspeakers' feed is directly transmitted which prevents the users from having any control over the sound components. Moreover, the mixed audio signal is transferred into the listener intended for a specific loudspeaker setup which is not generally what we have at the rendering point.

In object-based audio (OBA), the transferred data contains unmixed components known as audio objects along with their attributes such as their spatial positions and relative levels. The OBA rendering is then designed to be compatible with arbitrary loudspeaker configurations at the listening point.

There are two major parametric approaches that decode the audio scene, recorded as spatial Room Impulse Responses (RIRs), with some attributes which can be tuned by the end users and rendered with arbitrary loudspeaker setups. The first one is Spatial Impulse Response Rendering (SIRR) [11]–[13] which works in the time-frequency (T-F) domain. The B-format RIRs are mapped into T-F domain and then the direction of arrival (DoA), i.e. azimuth and elevation, and diffuseness of the sound which measures the reverberation, are calculated at each T-F unit. This process is aligned by the frequency-dependent nature of the reflections.

The other method is Reverberant Spatial Audio Objects (RSAO) [14], [15] that employs the time domain RIRs and divides it into three parts named as direct sound, early reflections, and late reverberation. The early part and the late reverb are separated at the mixing time, T_m, when the echo density [16] becomes greater than 1 as described in [17].

There is also another well-known technique called the Spatial Decomposition Method (SDM), which decomposes the spatial RIRs into a set of image sources [18]. Even though it is a valid representation of the direct sound and the early reflections, it is not an efficient way to parameterize the late reverberation [19].

In this paper, we compare these two parametric approaches and analyse their strengths and weaknesses in a VR environment. We discuss that SIRR method preserves the properties of the early reflections more accurately, while the RSAO is a

This work was partially supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction (EP/V03538X/1) and partially by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E32303).

better option for late reverberation modeling.

We then propose a new approach that combines these algorithms to get better performance with less data required to be transferred. In this way, we first employ the diffuseness factor in SIRR to estimate the mixing time, the time boundary between the early reflections and the late reverberation, T_m, for RSAO approach. Once the RIR is divided into the early part and late reverb, we use the SIRR to analyse and synthesise the early reflections, while the reverberation is encoded and rendered using RSAO parameters.

In section II the baseline algorithms, SIRR and RSAO for audio scene parameterization are explained, followed by section III which compares the methods. Then our proposed method is introduced in section IV. The experiments and results are described and discussed in sections V and VI, respectively.

II. Audio Scene Parameterisation

A. SIRR: Analysis and Synthesis

Spatial Impulse Response Rendering (SIRR) as presented in [11] maps the B-format audio signals into the time-frequency domain and estimates the azimuth, $\theta(m,\omega)$, elevation, $\phi(m,\omega)$ and diffuseness, $\psi(m,\omega)$ of the recorded signal at each time-frequency unit, (m,ω), where m is the time frame and ω is the frequency as in Short-time Fourier Transform (STFT). The analysis presented in [11] is based on the energy analysis of sound fields using B-format microphone measurements. Instantaneous intensity is defined as the product of the sound pressure $p(t)$ and the particle velocity vector $\mathbf{u}(t)$, $\mathbf{I}(t) = p(t)\mathbf{u}(t)$. The direction of arrival (DoA) can be directly estimated as the opposite of the direction of the sound intensity $\mathbf{I}(t)$. The diffuseness of the sound can be considered as the proportion of sound energy with no net flow and random local oscillations.

A B-format microphone records four channel signals named as $W(t)$, $X(t)$, $Y(t)$, and $Z(t)$. The omnidirectional signal $W(t)$ is proportional to the sound pressure $p(t)$ at the measurement position. As the relative values, and not the absolute values, of the sound intensity and energy density is what act on the equations, in this approach, the $W(t)$ is set as sound pressure, $W(t) = p(t)$. Moreover, the orthogonal signals $X(t)$, $Y(t)$, and $Z(t)$ are (ideally) proportional to the components of the particle velocity in the corresponding direction of a Cartesian coordinate system. Consider a vector $\mathbf{V}(t) = X(t)\mathbf{e}_x + Y(t)\mathbf{e}_y + Z(t)\mathbf{e}_z$, where $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ represent unit vectors in the direction of coordinate axes. By substituting the Fourier transforms of the B-format signals, the active intensity will be:

$$\mathbf{I}_a(m,\omega) = \frac{\sqrt{2}}{Z_0} Re\{W^*(m,\omega)\mathbf{V}'(m,\omega)\}, \tag{1}$$

and the diffuseness can be calculated as:

$$\psi(m,\omega) = 1 - \frac{\sqrt{2}||Re\{W^*(m,\omega)\mathbf{V}'(m,\omega)\}||}{|W(m,\omega)|^2 + |\mathbf{V}'(m,\omega)|^2/2}. \tag{2}$$

Having the intensity vector as in (1), the azimuth, $\theta(m,\omega)$ and the elevation $\phi(m,\omega)$ of the direction of arrival (DoA) can be written in the form of:

$$\theta(m,\omega) = \tan^{-1}\left[\frac{-\mathbf{I}_y(m,\omega)}{-\mathbf{I}_x(m,\omega)}\right], \tag{3}$$

$$\phi(m,\omega) = \tan^{-1}\left[\frac{-\mathbf{I}_z(m,\omega)}{\sqrt{\mathbf{I}_x^2(m,\omega) + \mathbf{I}_y^2(m,\omega)}}\right], \tag{4}$$

where $\mathbf{I}_x(m,\omega)$, $\mathbf{I}_y(m,\omega)$, and $\mathbf{I}_z(m,\omega)$ are the components of the active intensity in the directions of the corresponding Cartesian coordinate axes.

For the synthesis of a sound field through SIRR approach, the omnidirectional room impulse response is modified using the analysis parameters [11]. For the synthesis, similar to the analysis, the SIRR algorithm works in STFT domain and each time-frequency component is transmitted to the loudspeaker channels. The energy of each T-F component is first divided into nondiffuse and diffuse parts by applying the diffuseness parameter $\psi(m,\omega)$. The non-diffuse part of the omnidirectional signal $\sqrt{1-\psi(m,\omega)}W(m,\omega)$ is regenerated from the correct directions using amplitude panning, while the diffuse part is simulated by distributing the total diffuse energy $\psi(m,\omega)W^2(m,\omega)$ in a decorrelated form around the listener.

For rendering the reverberation, *Convolution diffusion* has been introduced in [11]. In this method, an exponentially decreasing function multiplied by a random Gaussian noise is convolved with the extracted diffuse part of the W-channel signal to decorrelate the signal for each loudspeaker.

For the nondiffuse rendering, a gain factor is calculated for each time-frequency component at each loudspeaker. This can be explained as applying different linear (zero) phase filters to the omnidirectional signal for each loudspeaker, where the filters are changing frame by frame. For this purpose, Vector Based Amplitude Panning (VBAP) [20] has been used.

B. RSAO: Encoding and Rendering

Another major parametric approach to encode the room acoustics is Reverberant Spatial Audio Object (RSAO) [14], [15] which characterizes the RIRs with different parameters that can be sent to a renderer with an arbitrary loudspeaker arrangement.

The main idea is to encode the recorded B-format RIR with a few parameters. To do so the RIR is divided into three parts namely, direct sound, specular early reflections, and late reverberation as represented in Fig. 1.

Four different parameters are calculated for each reflection peak: time of arrival (ToA), direction of arrival (DoA), attenuation (level), and the spectrum using B-format recordings. In [15], the ToAs are estimated using the dynamic programming projected phase-slope algorithm (DYPSA) [21] to identify the delays along the omnidirectional (W) channel of the B-format signal. For DoA, a virtual cardioid is steered by acting on the B-format signal $[W(t), X(t), Y(t), Z(t)]$.

The steered RIR is then employed for the level and spectrum estimation. The spectrum is modeled as 8 linear prediction

coefficients (LPCs). For the spectrum, the LPCs are then transformed to second-order sets of infinite impulse response (IIR) coefficients. Since LPC is equivalent to modelling the signal with an all-pole autoregressive model, its spectrum can be encoded with IIR filter coefficients.

The parameters used to describe the late reverberation are the mixing time, T_m, the late peak and decay time constant. The late delay or the mixing time is determined as the time when the echo density as in [16] becomes greater than 1. The mixing time is defined as the time when the acoustic space becomes fully mixed and reverberant so it is assumed that the impulse response will behave like Gaussian noise. The echo density measures the samples of RIR that lie outside of the standard deviation of a fitted Gaussian distribution to the windowed frame of the RIR. In [16], an echo density profile has been introduced that starts near zero and increases over time to around one, at the time when can be considered as the mixing time, T_m.

Once the mixing time has been evaluated, the late peak P_b, is calculated as the noise gain in the proximity of the mixing time at each octave sub-band, b [15]:

$$P_b = \sqrt{\sum_{t=T_m-I_b}^{T_m+I_b} W_b(t)^2}, \tag{5}$$

where W_b is the subband filtered W, the W channel, and $I_b = 2f_s/f_b$ is the frequency-dependent window size, with f_s and f_b being the sampling frequency and subband center frequency, respectively.

Then, the Schroeder energy decay curve (EDC) is estimated for the segmented signal at each octave band after the mixing time and an exponential curve is fitted to the EDC over the first 20 dB to estimate the time constant as a parameter.

For the rendering, each reflection is shifted (delayed), scaled, filtered and spatialised using VBAP based on the encoded parameters. For the late reverberation, a series of white Gaussian noise is first filtered to divide in each octave subband and then multiplied to a time envelope at each subband. The time envelope starts from zero before the first early reflection and then increases linearly to P_b at the mixing time, and then decreases exponentially with the encoded time constant. Finally, the signals are decorrelated using random-phase all pass FIR filter to send to the loudspeakers.

III. Comparing SIRR and RSAO

A. Early Reflections

As mentioned in II-A, the SIRR approach estimates the azimuths and elevations (DoAs) of the reflections at each T-F unit. On the other hand, in II-B, the RSAO evaluates the ToA, DoA, attenuation (level) and spectrum of each early reflection as a peak in the time domain.

Comparing these two techniques, the former one transforms the signal into time-frequency spectrum, whereas the latter one works in the time domain. As presented in [11], the early reflections are more apparent in the T-F domain since they are obviously frequency dependent. In that case, the DoAs are also different for each T-F unit, while in RSAO, DoAs are just related to the peaks in time.

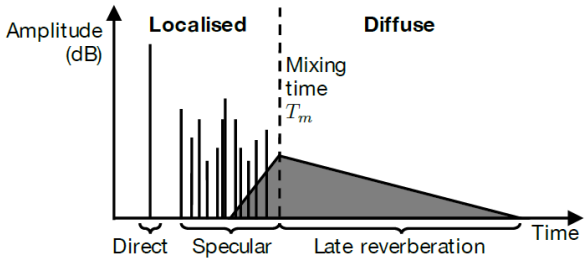

Fig. 1. RIR model as in RSAO algorithm [15].

Regarding ToA and attenuation (level) of each reflection, they are somehow embedded in the SIRR algorithm as can be simply spotted as onset times and amplitudes relative to the direct sound in the STFT spectrum.

As described earlier, the spectral properties of each reflection in RSAO is encoded using LPC coefficients, whereas SIRR works with Fourier transform as in STFT. The frequency components of the signal time frame in STFT is calculated with more coefficients, e.g. $N = 128$, compared to LPC, $N = 8$. On the grounds of this, it will be expected that SIRR conveys more detailed information about the early reflections spectrum compared to RSAO.

Another benefit of using SIRR is that SIRR parameters can be obtained from any recorded signal whereas RSAO needs RIRs to estimate the parameters. However, the main drawback of the SIRR is that it requires the W-channel signal to be transmitted, while RSAO only requires the encoded parameters which significantly reduces the size of data to be transferred.

B. Late Reverberation

In the SIRR analysis, a diffusion factor is the only factor which is calculated at each T-F unit to determine how dispersed and diffuse the sound energy is at that particular time and frequency. This parameter is then used to extract the reverberant energy from the sound for rendering implementation. The primary limitation of the SIRR approach is its reliance on the entire W-channel signal from the B-format RIR recording, in addition to the parameters, for reverberation rendering. In contrast, RSAO models offer the advantage of being able to render reverberation solely based on the provided parameters. These parameters, such as mixing time, late peak, and exponential decay time constant, can be utilized without the necessity of transmitting the W-channel or any other channels from the B-format recorded RIR alongside the parameters. On this account, we can argue that RSAO is considerably a better option for the late reverberation as it requires significantly less data, just a few parameters, to be transferred to the renderer compared to the whole W-channel signal of the B-format recording as in SIRR.

IV. Proposed Combined Approach

The first idea for the combined method is to exploit the diffuseness factor, $\psi(m,\omega)$, in SIRR to estimate the mixing time, T_m, for RSAO approach in order to divide the RIRs into early and late parts. Then SIRR is applied to the early segment, while the late tail is modeled by RSAO. Then both parts are fed to the loudspeakers together to regenerate the sound field.

There are other methods to estimate the diffuseness of a sound field such as coherence-based diffuseness estimation [22] and COMEDIE [23]. The former is based on the coherence between the eigenbeams, while the latter employs the covariance matrix. However, both of them are based on the recordings from spherical microphone arrays (SMAs) which typically consist of a few dozen microphones. Götz et al. [24] have also proposed a method to calculate the mixing time based on diffuseness, but again, they have employed SMAs. In contrast, the SIRR approach only requires a B-format microphone.

In addition to that, as mentioned earlier in II-B, the echo density approach [17] used in RSAO relies on the assumption that a reverberant field takes on a Gaussian distribution once acoustic space is fully mixed. However, It does not consider the directivity of the sound intensity. In comparison, the diffuseness factor in SIRR measures the proportion of the sound energy with no net flow and random local oscillations which is in fact the physical property of the reverberation.

For the combined method the signal is first mapped into T-F domain as in SIRR and the diffuseness, $\psi(m,\omega)$, is calculated. Then $\psi(m,\omega)$ is averaged over the frequency bins to obtain the variation of ψ over time $\psi(m)$. Having set a threshold, the time when the $\psi(m)$ reaches that, is considered as the time when the main acoustic energy has become diffuse. Then the first part before this mixing time is regarded as early reflections and the azimuths, $\theta(m,\omega)$, and elevations, $\phi(m,\omega)$, are calculated for each T-F unit as explained in section II-A. Then the first part of the W-channel, i.e. before T_m, along with θ, ϕ and ψ for time frames up to T_m are stored to be transmitted.

Fig. 2 shows the $\psi(m)$ variations over time for a random room, Kitchen, RIR. As you can see there is a sharp trough at the initial delay for direct sound and then it increases until it reaches some steady level. In this case it reaches a threshold of 0.9 at around 0.04 s.

At the same time, the late part of the RIR signal in the time domain is processed to evaluate the late reverb tail parameters, the late peak and the decay time constant, as described in section II-B. The parameters are then stored to be transmitted to a renderer.

At the renderer, $\psi(m,\omega)$ is applied to extract the non-diffuse energy from the transferred early part of W-channel signal. After that, having the $\theta(m,\omega)$ and $\phi(m,\omega)$, VBAP is applied to estimate each loudspeaker feed that contributes to the early reflections. Alongside that, the reverberation tail is reconstructed using the parameters and then decorrelated using random phase and is added to the loudspeaker feeds.

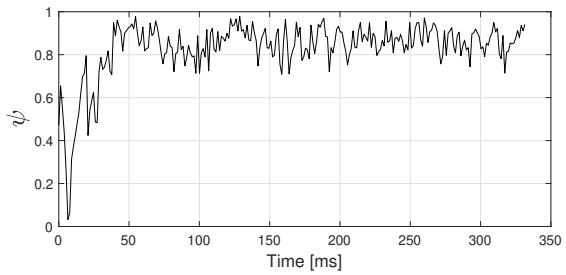

Fig. 2. This is an example of diffuseness factor, $\psi(m)$, variations over time for Kitchen RIR.

V. Experiments and Results

In order to objectively evaluate the performance, we measured the actual RIRs and compared with the regenerated RIRs in a corresponding VR space using the transmitted parameters estimated by the methods following the pipeline in [6].

Three different RIRs captured by B-format microphones in three different rooms are used for this experiment: Kitchen (KT), Meeting room (MR) and Courtyard (CY) [6]. Fig. 3 represents the Kitchen and its 3D model as an example. The RIR parameters for each room have then been estimated using SIRR, RSAO and combined methods and submitted to a renderer. The entire rendering procedure is performed in Unity, a game engine known for its ability to create 3D environments with virtual loudspeakers and microphones.

To replicate the listening environment, we positioned 64 virtual loudspeakers in Unity, arranged in a spherical configuration at elevations of $0°$, $\pm 28°$, $\pm 56°$, $\pm 80°$, as recommended in [13]. Please refer to Fig. 4 for a visual representation. Next, each approach was employed to calculate the loudspeakers' feed based on the transmitted RIR parameters. Then the Google Resonance Audio package was incorporated in the Unity platform to create spatial audio in the virtual environment. Following the simultaneous playback of the designated signals for each loudspeaker, the perceived regenerated RIR is recorded in the sphere's centre using a virtual B-format microphone in 4-channel.ogg format. Fig. 5 compares the original recorded RIRs with the regenerated ones in Unity. It appears that early reflections are better retrieved using the proposed approach compared to RSAO.

Different metrics have also been evaluated to compare the methods namely, the Xcorr, the reverberation time (RT60), the clarity index (C50), and the early decay time (EDT). We have chosen these metrics because, as explained in [25], these objective measurable metrics are well correlated with the subjective parameters. Xcorr is the cross-correlation between the early part of the recovered and the recorded RIR while RT60 is the time it takes that the sound level drops by 60 dB after the sound source has been turned off. The clarity index, C50, measures how clear and not smeared the sound is, which is defined as the ratio between the energy in the RIR before 50 ms and after that. The parameter EDT, similar to RT60, also estimates the rate of the decay, but based on the initial

Fig. 3. The Kitchen 3D model with one sound source.

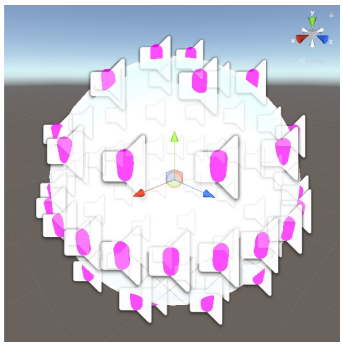

Fig. 4. The rendering 64 loudspeaker setup in Unity to regenerate the same acoustic properties of the room.

part from 0 dB to -10 dB and then extended to -60 dB [25]. The EDT and RT60 are calculated for each octave subband and then the averaged values are reported.

Table I shows the values for different rooms using different methods. At this stage, the W-channel is employed for evaluation, but SIRR will not give identical results as only 64 loudspeakers are used for rendering. For Xcorr, SIRR gives higher values as expected as it reconstructs the early part more precisely. For the combined method the Xcorr values are a bit lower due to the linearly rising diffuse energy added before the mixing time as shown in Fig. 1, but still considerably better than RSAO. For RT60, the combined method shows better performance than RSAO in all the rooms. For MR, both C50, and EDT have closer values to the GT using the combined method. Even though the combined method might show poor performance based on EDT, its early part match the original RIR better as shown in Fig. 6.

Table II reports the data required to be transferred to the renderer for each method. It is evident that the Recorded RIRs and SIRR approach require a substantial amount of data transmission, whereas, for the RSAO and combined methods, significantly less data is needed.

VI. Discussion

Comparing SIRR and RSAO, the early reflections are analysed and synthesised remarkably better by SIRR compared to RSAO according to Xcorr in Table I. Regarding the late reverberation, RSAO employs just a few parameters to regenerate the reverb tail with remarkably less data as shown in Table II.

For the Combined method, as long as it gives better results than RSAO, it is acceptable, since it is not fair to compare it with SIRR which has the whole W-channel along with the parameters. This is why we emphasized the results of the combined method in Table I whenever it demonstrates superior performance in comparison to RSAO, or even better, both RSAO and SIRR.

As presented in Table I, it can be seen that the combined method works better than RSAO with higher Xcorrs and closer RT60 values to the ground truth recorded values. As mentioned in [25], RT60 is considered as the baseline and most important objective parameter. Moreover, the combined method exploits significantly less data than SIRR as shown in Table II.

While Table I indicates that RSAO exhibits better values for EDT in KT and CY, it does not necessarily imply that RSAO accurately matches the ground truth. In contrast, as depicted in Fig. 6, the combined method demonstrates significantly higher precision than RSAO when calculating the initial portion of RIR, which is crucial for determining EDT. This discrepancy could be due to the EDT calculation method, an aspect that we plan to explore further in future investigations.

Overall, it can be concluded that in general the combined method is a better option than SIRR as it requires considerably less data to be transferred. It is worth mentioning that the data reported in Table II is only for one RIR while to regenerate the spatial sound of an environment hundreds of RIRs are needed to be recorded in different positions in the room and therefore, reducing the data is essential. The combined method also improves the RIR reconstruction compared to RSAO as measured by Xcorr and RT60. An informal listening test

TABLE I
THE METRICS XCORR, RT60 [S], C50 [DB], AND EDT [S] FOR RECORDED GROUND TRUTH RIR (GT) AND RENDERED RIRS USING SIRR, RSAO, AND THE COMBINED METHOD.

Metric	room	GT	SIRR	RSAO	Combined
Xcorr	KT	8.64	**7.10**	0.81	**4.27**
	MR	4.13	**1.56**	0.21	**1.24**
	CY	3.10	**2.37**	0.54	**1.44**
RT60	KT	0.396	0.377	0.380	**0.385**
	MR	0.224	0.255	0.199	**0.206**
	CY	0.952	**0.975**	1.106	**1.00**
C50	KT	6.53	9.53	**6.79**	7.44
	MR	16.56	15.17	13.61	**17.50**
	CY	7.81	9.54	**7.92**	10.98
EDT	KT	0.428	0.384	**0.415**	0.372
	MR	0.224	**0.221**	0.290	**0.211**
	CY	0.813	0.477	**0.773**	0.223

TABLE II
THE DATA REQUIRED TO BE TRANSFERRED FOR EACH APPROACH.

	KT	MR	CY
Recorded	3.66 MB	512 KB	3.66 MB
SIRR	3.66 MB	512 KB	3.66 MB
RSAO	12 KB	12 KB	12 KB
Combined	15 KB	12 KB	30 KB

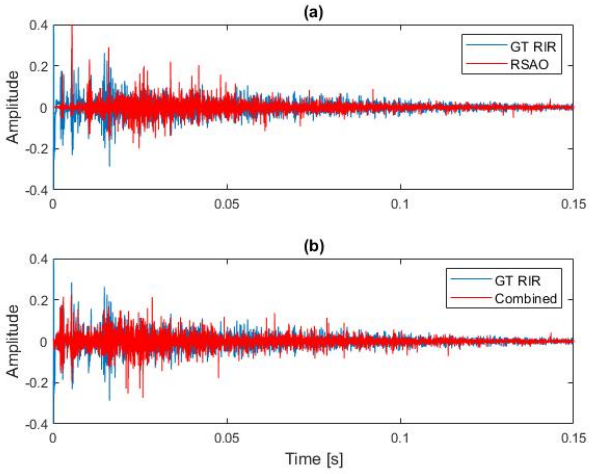

Fig. 5. The original Kitchen recorded ground truth (GT) RIR in blue, and regenerated RIR using (a) RSAO and (b) Combined methods in red.

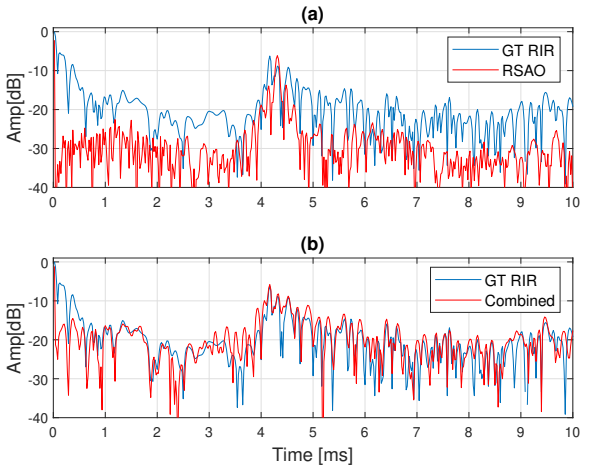

Fig. 6. The original Courtyard (CY) ground truth (GT) RIR in blue, and regenerated RIR using (a) RSAO and (b) Combined methods in red.

has also confirmed that the spatial audio regenerated by the combined method is closer to the original recorded sounds.

REFERENCES

[1] J. Yang, Y. Frank, and G. Sörös, "Hearing is believing: Synthesizing spatial audio from everyday objects to users," in *Proceedings of the 10th Augmented Human International Conference 2019*, ser. AH2019. New York, NY, USA: Association for Computing Machinery, 2019.

[2] L. Buck, M. F. Vargas, and R. McDonnell, "The effect of spatial audio on the virtual representation of personal space," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022, pp. 354–356.

[3] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[4] H. Kim, L. Remaggi, P. Jackson, and A. Hilton, "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360° images," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, pp. 120–126.

[5] H. Kim, L. Remaggi, A. Dourado, T. D. Campos, P. J. B. Jackson, and A. Hilton, "Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras," *Virtual Reality*, vol. 26, pp. 823–838, 2022.

[6] L. Remaggi, H. Kim, A. Neidhardt, A. Hilton, and P. J. B. Jackson, "Perceived quality and spatial impression of room reverberation in vr reproduction from measured images and acoustics," *Proceedings of the 23rd International Congress on Acoustics (ICA)*, pp. 3361–3368, Sep 2019.

[7] F. Olivieri, N. Peters, and D. Sen, "Scene-based audio and higher order ambisonics: A technology overview and application to next-generation audio, vr and 360 video," The European Broadcasting Union (EBU) Technology and Innovation, Technical review, Nov 2019.

[8] S. Shivappa, M. Morrell, D. Sen, N. Peters, and S. M. A. Salehin, "efficient, compelling, and immersive vr audio experience using scene based audio/higher order ambisonics," *journal of the audio engineering society*, september 2016.

[9] R. Bleidt, A. Borsum, H. Fuchs, and S. M. Weiss, "Object-based audio: Opportunities for improved listening experience and increased listener involvement," *SMPTE Motion Imaging Journal*, vol. 124, no. 5, pp. 1–13, 2015.

[10] ITU-R, *Recommendation BS.2076-0, Audio Definition Model*, International Telecommunication Union (ITU), June, 2015.

[11] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, december 2005.

[12] V. Pulkki and J. Merimaa, "Spatial impulse response rendering ii: reproduction of diffuse sound and listening tests," *Journal of the Audio Engineering Society*, vol. 54, no. 1/2, pp. 3–20, January/February 2006.

[13] L. Mccormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-order spatial impulse response rendering: investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *Journal of the Audio Engineering Society*, vol. 68, no. 5, pp. 338–354, May 2020.

[14] P. Coleman, A. Franck, P. J. B. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 66–77, January 2017.

[15] P. Coleman, A. Franck, D. Menzies, and P. J. B. Jackson, "Object-based reverberation encoding from first-order ambisonic rirs," *Journal of the Audio Engineering Society*, May 2017.

[16] J. S. Abel and P. Huang, "a simple, robust measure of reverberation echo density," *journal of the audio engineering society*, October 2006.

[17] A. Lindau, L. K. osanke, and S. Weinzierl, "perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," *journal of the audio engineering society*, vol. 60, no. 11, pp. 887–898, November 2012.

[18] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "spatial decomposition method for room impulse responses," *journal of the audio engineering society*, vol. 61, no. 1/2, pp. 17–28, january 2013.

[19] S. V. A. Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "optimizations of the spatial decomposition method for binaural reproduction," *journal of the audio engineering society*, vol. 68, no. 12, pp. 959–976, december 2021.

[20] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, June 1997.

[21] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[22] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, 2012, pp. 1–5.

[23] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, 2016.

[24] P. Götz, K. Kowalczyk, A. Silzle, and E. A. Habets, "Mixing time prediction using spherical microphone arrays," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. EL206–EL212, 2015.

[25] T. D. Rossing, *Springer Handbook of Acoustics*. New York, 2007.