Audio Engineering Society
Convention Express Paper 456

Presented at the 160th Convention
2026 May 28-30, Copenhagen, Denmark

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Generate 4pi SBA reverberation from virtual sound sources detected from x-y-z sound intensities. - Improvement of the source detection method.

Masataka Nakahara[1,2], Yasuhiko Nagatomo[3], and Akira Omoto[4,1]

[1] *ONFUTURE Ltd, Tokyo 141-0032 Japan*
[2] *SONA Corporation, Tokyo 164-0013 Japan*
[3] *Evixar Inc, Tokyo 104-0033 Japan*
[4] *Kyushu University, Fukuoka 815-8540 Japan*

Correspondence should be addressed to Masataka Nakahara (nakahara@onftr.com)

ABSTRACT

Accurate reconstruction of 4pi reverberant sound fields is essential for immersive audio, and Scene-Based Audio (SBA) reverb is a vital tool for this purpose. We have developed VSVerb (Virtual Source reVerb), which generates 4pi reverberation based on virtual sound sources detected from three-dimensional (*x*-*y*-*z*) sound intensities measured in a target space. In VSVerb, the reliable extraction of virtual sound sources —an acoustic concept in geometrical acoustics also known as mirror or imaginary sources— is critical. While our previous "Speed Detection method" separated source-related components based on spatial moving speed, its performance was limited by a fixed threshold model that failed to account for environment-dependent variability. This paper proposes an adaptive thresholding scheme derived from the local statistical behavior of spatial moving speed. This approach enables robust separation of coherent and incoherent components without relying on predefined sound propagation models. Experimental results demonstrate that the proposed method improves the temporal and spatial consistency of detected virtual sound sources, resulting in an SBA reverb, VSVerb, which more closely matches measured impulse responses. These results indicate that adaptive, locally-informed thresholding is essential for physically consistent virtual source detection and provides a reliable foundation for VSVerb.

1 Introduction

We have previously developed VSVerb (Virtual Source reVerb), which generates 4pi SBA reverberation based on virtual sound sources detected from three orthogonal sound intensities measured in a target space [1]. Building upon this, we introduced V2MA (VSVerb Virtual Microphone Array), a technology that enables virtual immersive recording by virtually constructing a desired microphone array in a target space [2]. A fundamental challenge in these developments is the accurate identification of the acoustic properties of sound sources from measured sound intensities. In this paper, we present a significant advancement in this core technique—the first major improvement in nine years.

2 Virtual sound source

A virtual sound source (mirror source, imaginary source) is a fundamental concept in geometrical acoustics, representing a reflection path as an equivalent source located outside the room. In the ideal geometrical model, each reflection is associated with a single specular path, and the corresponding virtual source emits sound toward the receiver at a specific time (Figure 1). This approximation provides a convenient representation of reflection structure but neglects the wave nature of sound. In measured sound fields, reflections are not generated by a single point but by a finite region on a boundary surface. Therefore, the contribution of a virtual sound source should be interpreted not as a point, but as a spatially extended region over a short time interval.

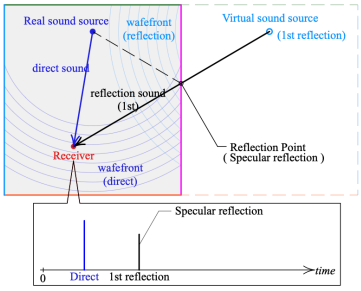


Figure 1. Conventional geometrical acoustics.

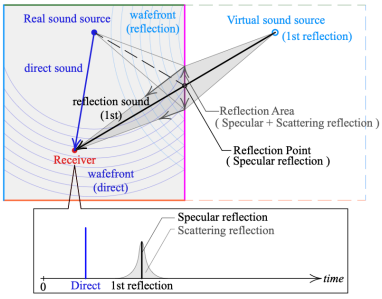


Figure 2. Extended geometrical acoustics.

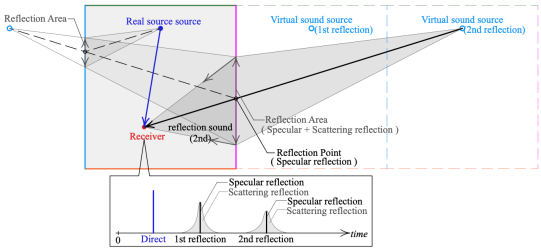


Figure 3. Reflection area size and reflection order.

3 Extension of geometrical acoustics

To account for the wave nature of sound, we extend the concept of virtual sound sources by considering the spatial extent of reflection.

A finite reflective surface produces not only specular reflection but also scattering components. As a result, a reflection arriving at the receiver is generated by a finite region on the surface, rather than a single point (Figure 2). Furthermore, as the wavefront propagates, its spatial extent increases with distance. Consequently, the effective reflection area associated with a virtual sound source grows with reflection order (Figure 3). This implies that the contribution of a virtual sound source is distributed over a short time interval and a finite directional region. In the time domain, this leads to a cluster of temporally continuous and directionally coherent components, rather than isolated impulses.

This interpretation provides the basis for distinguishing coherent reflection components from incoherent ones in the Speed Detection method.

4 Speed Detection

To extract virtual sound sources, it is necessary to distinguish coherent reflection components from incoherent or stochastic behavior in the sound intensity field. We interpret this distinction in terms of temporal coherence and spatial stability of the instantaneous sound intensity vector.

Reflections originating from a localized region produce intensity vectors that remain directionally stable over short time intervals. In contrast, diffuse components and noise exhibit rapid and irregular directional changes.

Based on this, we defined:

- **Source Intensity:**
  coherent, directionally stable components
- **Drifting Intensity:**
  incoherent, rapidly varying components

To quantify this behavior, we introduce the spatial moving speed v(t), representing the rate of directional change. The Speed Detection method assumes that: Source Intensities have low v(t), while Drifting intensities have high v(t). Thus, applying a threshold to v(t) enables separation of coherent reflections from incoherent components (Figure 4) [3]. This can be regarded as coherence-based filtering in the directional domain. In contrast, conventional methods for detecting virtual sound sources have primarily relied on peak amplitude values [4, 5]. However, these approaches encounter difficulties in correctly identifying the relationships between the peaks of the *x*, *y*, and *z* sound intensities. This is particularly evident in the late reverberation stage, where a high peak density and a low signal-to-noise ratio prevail (Figure 5).

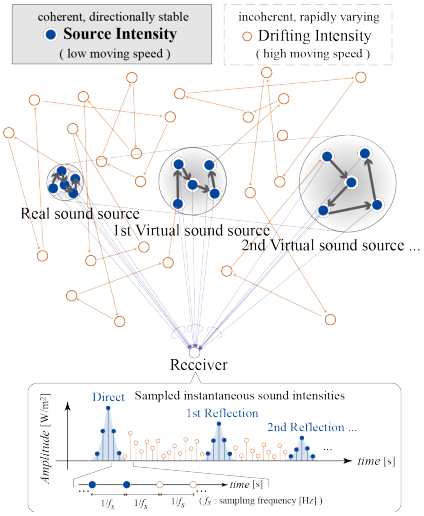


Figure 4. Source Intensity and Drifting Intensity.

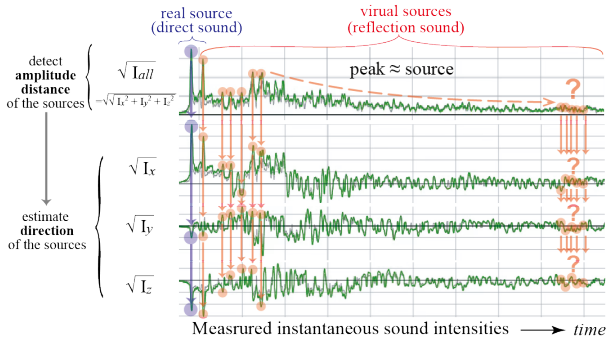


Figure 5. Difficulty of the peak detection method.

5 Threshold for spatial moving speeds

5.1 Limitation of fixed threshold

In previous work [3], the threshold was defined as:

$$v_t(t) = 40(100t + 10)^{1.5} \ \text{[m/s]} \qquad (1)$$

where, t indicates arrival times of sound intensities [s].

This model assumes smooth temporal evolution of spatial moving speed based on propagation distance. Since the threshold is proportional to the 1.5th power of the travel time ($\propto t^{1.5}$), it implies that the sound propagation exhibits characteristics intermediate between those of a point source ($\propto t^2$) and a line source ($\propto t$) (Figure 6). However, in practice, measured sound fields exhibit strong local variability due to room geometry, source directivity, and interference. Consequently, $v(t)$ often shows non-smooth behavior that cannot be captured by a predefined fixed function. This leads to missed detections and over-segmentation, indicating that a fixed threshold is inherently insufficient. Therefore, an adaptive thresholding approach that accounts for local field variations is necessary.

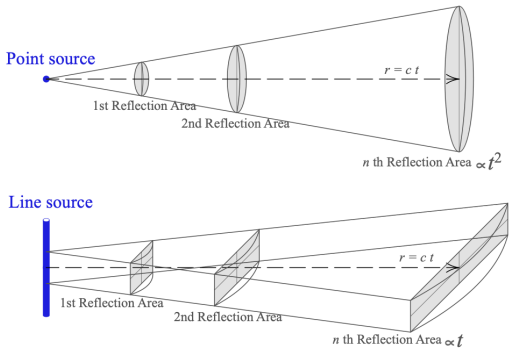


Figure 6. Source types and wavefront propagation.

5.2 Adaptive threshold

We propose an adaptive threshold derived from the local statistics of $v(t)$:

$$v_t(t) = \frac{1}{T(t)} \int_{t-\mathrm{T}(t)/2}^{t+\mathrm{T}(t)/2} v(t)dt \ \text{[m/s]} \qquad (2)$$

where, $T(t)$ is a time-varying window.

The instantaneous speed is smoothed as:

$$\bar{v}(t) = \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t)dt \ \text{[m/s]} \qquad (3)$$

The smoothing interval is set to $\Delta t=10^{-4}$ s, a value empirically determined to provide a stable estimate.

Source Intensities are then detected using the following criterion:

$$\bar{v}(t_s) \leq v_t(t_s) \qquad (4)$$

This formulation adapts to the local structure of the sound field, enabling the separation of coherent (low variability) and incoherent (high variability) components without relying on a predefined propagation model.

6 Experiments

6.1 Measurement conditions

Instantaneous sound intensities in the x, y, and z directions were measured using an A-format microphone (Zoom H3-VR) in a lecture room at Tokyo University of the Arts.
The room dimensions were 6.45 m × 6.95 m × 2.69 m. Most surfaces were reflective, with the exception of the ceiling, which was finished with rockwool. An omnidirectional loudspeaker (B&K Type 4295) was positioned near the center of the room, while the microphone was placed at a height of 1.0 m with a 0.7 m horizontal offset from the source (Figure 7).

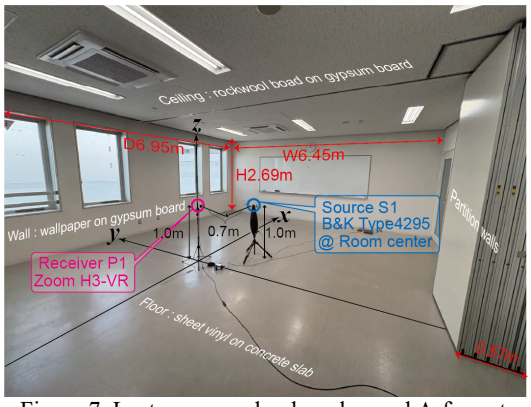


Figure 7. Lecture room, loudspeaker and A-format microphone for the measurements.

6.2 Processing flow

The virtual sound sources of the lecture room were obtained from the sound intensities calculated from four measured impulse responses using an A-format microphone. For Speed Detection, both the conventional (fixed) and proposed (adaptive) thresholds were employed. The 4pi SBA reverb, VSVerb [1], was generated from the detected virtual sound sources. The calculation flow is as follows:

1. **Measurement:** Four-channel A-format impulse responses (LFU, RFD, LBD, RBU) were measured.
2. **Filtering:** The responses were band-limited to suppress out-of-band noise.
3. **B-format Conversion:** The signals were converted to B-format (W, X, Y, Z), and the Hilbert transform was applied.
4. **Intensity Calculation:** Relative sound intensities (Ix, Iy, Iz) were calculated. Since the four capsules of the A-format microphone (LFU, RFD, LBD, RBU) exhibit nearly identical acoustic responses, the obtained W, X, Y, Z components are proportional to the sound pressure (p) and the particle velocities in the x, y, and z directions (u_x, u_y, u_z), respectively [6].
5. **Temporal Averaging:** A moving average with a 0.4 ms window was applied.
6. **Source Detection:** Source Intensities were identified using Speed Detection method.
7. **Phase Estimation:** The phase characteristics (±) were estimated.
8. **Reconstruction:** 4pi SBA reverb, VSVerb, was generated.

7 Results

7.1 Source Intensity detection

Figure 8 displays the calculated results of preprocessing steps 1 through 5, as described in the previous section. The spatial moving speed of the sound intensity is calculated from the time-averaged instantaneous sound intensities shown at the bottom of the figure.

Source Intensities, which contain sound source information, are identified by detecting sound intensities that move more slowly than a threshold value. In Figure 9 (a), the light- and dark-green lines represent the spatial moving speed $v(t)$ and its averaged value $\bar{v}(t)$, respectively. Figure 9 (a) also illustrates two types of threshold values: the conventional and the proposed thresholds. Under the conventional (fixed) method, the time ranges where the light-green lines fall below the blue dash-dotted line represent the time indices of the Source Intensities. For the proposed (adaptive) method, the time ranges where the dark-green lines fall below the red line represent these indices. The resulting detections, shown in (b1) and (b2), represent the Source Intensities of $I_{all}=\sqrt{(I_x^2+I_y^2+I_z^2)}$. In these figures, "pre-sources"—undesirable sources detected before the direct sound—are shaded in gray.

The conventional method produces fragmented detections and fails to capture significant reflections, particularly in the 10–100 ms range. This is primarily due to the inability of the fixed threshold to adapt to abrupt changes in spatial moving speed. In contrast, the proposed method yields more continuous and structured detections, indicating improved extraction of coherent reflection components.

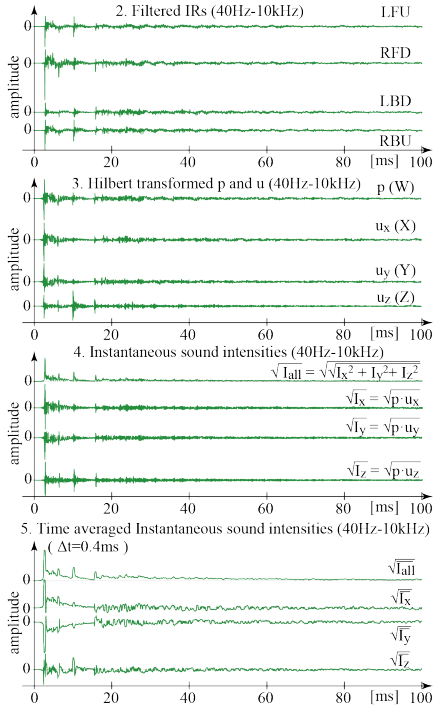


Figure 8. Preprocessing flow

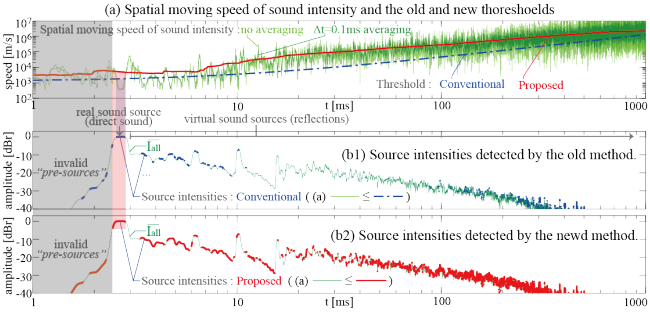


Figure 9. Spatial moving speed of sound intensities and detected Source Intensities.

7.2 Virtual sound sources

The red bold segments in Figure 10 indicate Source Intensities of the I_{all}, Ix, Iy, and Iz within the first 100 ms. These detected Source Intensities are labeled as Isi_{all}, Isi_x, Isi_y, and Isi_z (i = 0, 1, 2, …). The acoustic properties of the virtual sound sources were calculated from Isi_{all}, Isi_x, Isi_y, and Isi_z as follows:

1. **Distance from the source to the receiver (d_i)**

   Conventional method: d_i = mean (t_i) ·c [m]
   Proposed method: $d_i = t_i$ of max (Isi_{all}) ·c [m]

   where, t_i [s] denotes the time index of sound intensities within Isi_{all}, and c [m/s] is the sound speed.

2. **Amplitude (Is_i)**

   Conventional method: Is_i = mean (Isi_{all}) [W/m^2]
   Proposed method: Is_i = sum(Isi_{all}) [W/m^2]

3. **Elevation (θ) and Azimuth (ϕ) angles**

   For both methos:
   θ=tan^{-1}{mean(Isi_z) / $\sqrt{}$(mean(Isi_x)^2+mean(Isi_y)^2)} [rad]
   ϕ=cos^{-1}{mean(Isi_x) /$\sqrt{}$(mean(Isi_x)^2+mean(Isi_y)^2)} [rad]

The detected virtual sound sources are shown in Figure 11. The arrival times and relative amplitudes (referenced to the direct sound) of the latest reflection—representing the most distant virtual sound source—were (0.999 s, -51.4 dB) for the conventional method and (1.000 s, -50.9 dB) for the proposed method.

While the conventional method detected 3,638 sources, the proposed method identified 1,710 sources from the same dataset. This indicates that the conventional method tends to over-detect sources, whereas the proposed method produces a more compact and structured distribution. The spatial distributions obtained by the proposed method exhibit better agreement with expected reflection patterns.

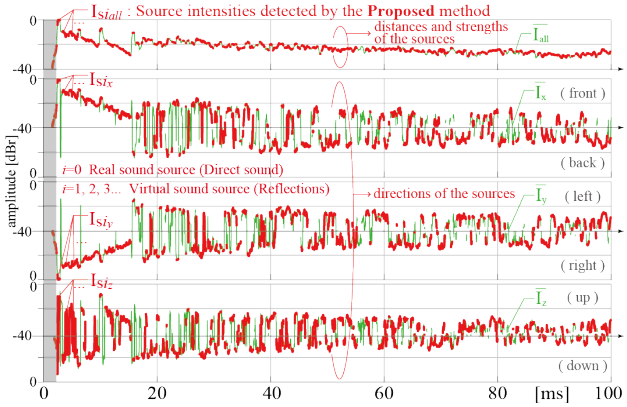


Figure 10. Source Intensities of the I_{all}, I_x, I_y, and I_z in the first 100ms, detected by the new method.

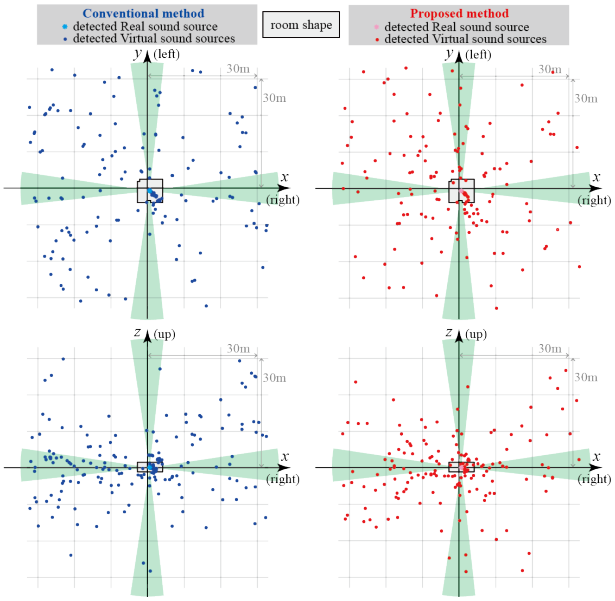


Figure 11. Detected virtual sound sources in the range of 30m (≈100ms).

7.3 Reconstruction of the reverb, VSVerb

The phase characteristics of the virtual sound sources—specifically the plus/minus signs of the reflections—were estimated from the sound pressure impulse response (p value rather than intensity) [1]. As shown in Figure 12 (b1), the impulse response was partitioned into small segments, including the direct source (p_0) and the virtual sources (p_1, p_2 …), using the time indices derived from the Source Intensities. Subsequently, as illustrated in (b2), these segments were deconvolved with the direct source segment (p_0). Finally, as shown in (b3), the summed value for each segment was calculated, and the phase of each virtual sound source was determined by the resulting plus/minus sign.

Figure 13 displays the impulse responses reconstructed from the detected direct and virtual sound sources, incorporating the estimated phases. The line colors indicate the arrival directions of the reflections: front, back, left, right, up and down. Because each reflection is associated with a specific arrival direction, the reconstructed impulse responses constitute an SBA reverberation. We designate this synthesized reverb as VSVerb (Virtual Source reVerb).

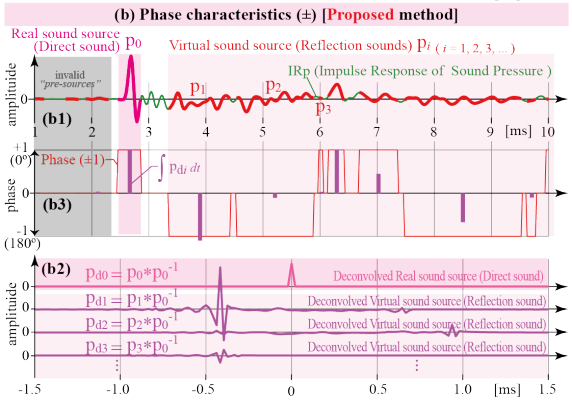


Figure 12. Calculating flow of the phase characteristics

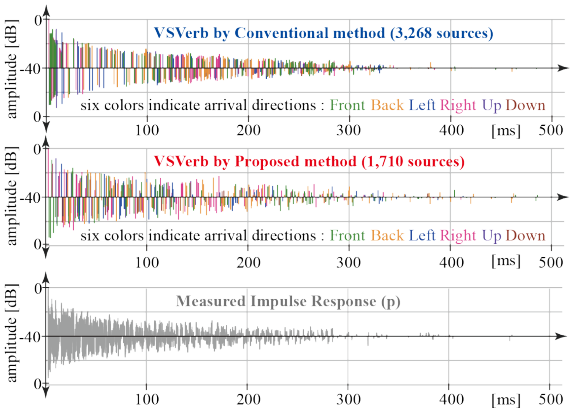


Figure 13. Generated VSVerb and measured IR.

8 Discussion

In this section, we evaluate the differences between the conventional (fixed) and proposed (adaptive) methods to determine which is more effective at extracting plausible virtual sound sources.

Figure 14 (b1) and (b2) display the "dominant" reflections within the first 50 ms of the VSVerb responses. These reflections were selected based on their magnitudes relative to the reverberation curves shown in (a1) and (a2). Time responses (b1) and (b2) illustrate the selected reflections with amplitudes normalized by these RT curves. In the responses, the green lines represent the geometrical direct and reflected sounds calculated via the mirror image method.

Figure 15 presents spatial scattergrams of the virtual sound sources derived from the time responses shown in Figure 14 (b1) and (b2). The gray dashed lines illustrate the mirror image of the room, while the green circles depict the virtual sound sources corresponding to first- and second-order reflections, as calculated by the mirror method. The green scopes indicate regions where virtual sound sources are expected to align, assuming the boundaries provide specular reflections.

While Figures 14 and 15 do not definitively prove the superiority of one method over the other, the proposed method appears to replicate the mirror-method results more accurately than the conventional method.

Figure 16 shows the spatial scattergrams of the virtual sound sources detected within the first 500 ms. The size of each circle is proportional to the source amplitude, and the six colors indicate the arrival directions: front, back, left, right, up and down. Notably, the conventional method produces a circular distribution, whereas the proposed method yields a rectangular shape resembling the actual room geometry. This result indicates that the conventional method tends to detect virtual sound sources randomly, particularly those distant from the receiver.

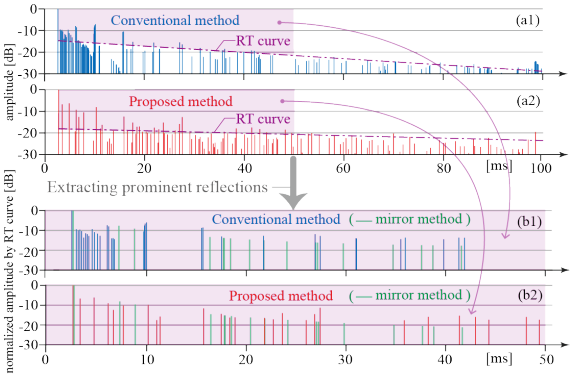


Figure 14. Comparison between old and new methods: Dominant reflections in the first 50ms.

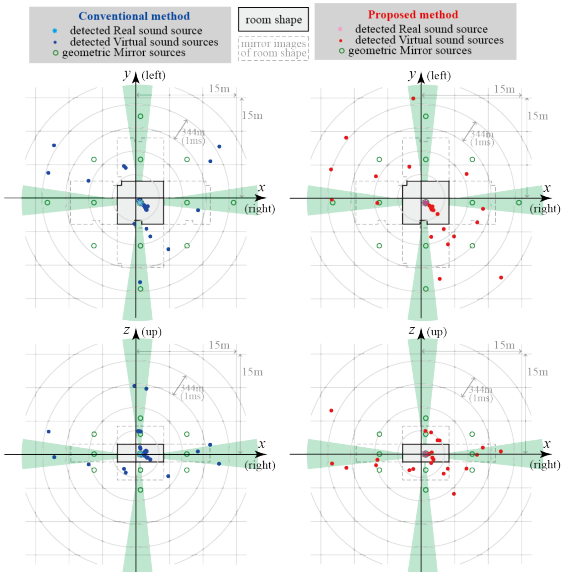


Figure 15. Comparison between old and new methods: Dominant virtual sound sources in 15m.

Figure 17 (a1) and (a2) illustrate the results shown in Figure 16 from a different perspective. The blue and red circles represent sources detected by the conventional and proposed methods, respectively. Compared with the measured impulse response (green lines), the conventional method (blue circles) appears to merely trace the envelope of the impulse response, neglecting the underlying acoustic properties hidden within the waveform. Consequently, the conventional method over-detects sources, resulting in a reverberation characteristic that decays more slowly than the measured impulse response, as shown in Figure 17 (b). Conversely, the proposed method provides a decay curve that closely resembles the measured data.

Figure 18 illustrates the cumulative total number of sound sources over time. The blue and red lines represent the conventional and proposed methods, respectively, while the green dash-dotted and dashed lines show theoretical results for 2D reflection (absorptive ceiling and five reflective surfaces) and 3D reflection (all reflective surfaces) [7]. We find that the conventional method yields an excessive number of sources within the first 300 ms. In contrast, the proposed method initially yields a source count nearly identical to the theoretical 3D value and later approaches the 2D theoretical value.

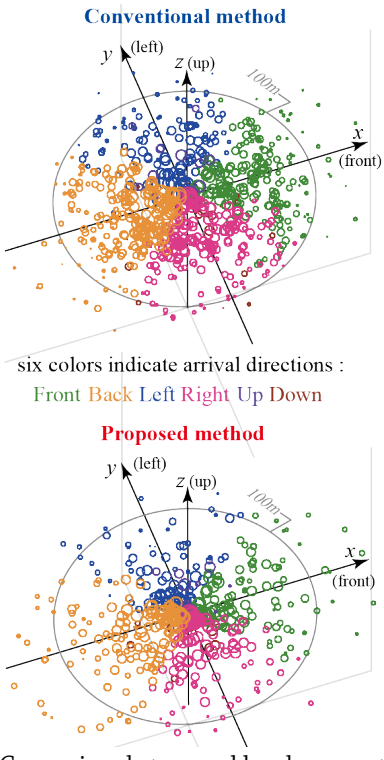


Figure 16. Comparison between old and new methods: Virtual sound sources in 150m (≈500ms).

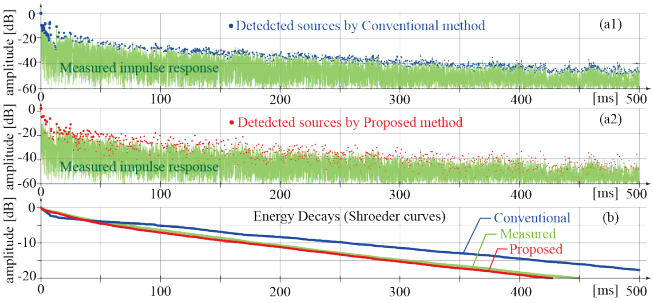


Figure 17. Comparison between old and new methods: Virtual sound sources with IR and decay curves in 500ms.

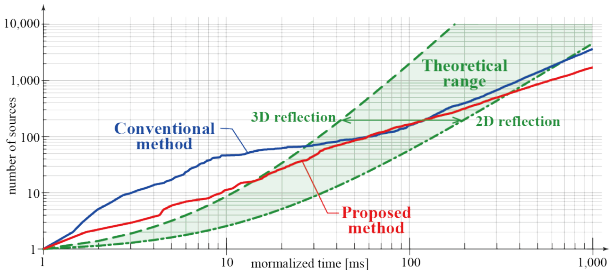


Figure 18. Comparison between old and new methods: Numbers of the virtual sound sources over time.

Acoustic quantity	JND (ISO 3382-1)	Measured IR	Conventional method	Proposed method
T20 [s]	(±5%)	1.49	1.74	1.38
		-	(17%)	-(7%)
C80 [dB]	(±1.0dB)	8.8dB	9.0dB	9.0dB
		-	(0.2dB)	(0.2dB)
D50	(±0.05)	0.818	0.865	0.826
		-	(0.047)	(0.008)
Ts [ms]	(±10.0ms)	33.0ms	28.8ms	31.6ms
		-	-(4.2ms)	-(1.4ms)

Table 1. Comparison between old and new methods: Calculated results of acoustic quantities.

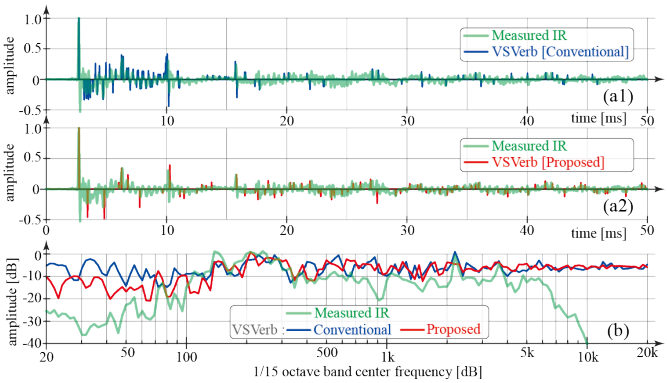


Figure 19. Comparison between old and new methods: Time (first 50ms) and frequency (all 1s) responses.

Table 1 summarizes the calculated acoustic quantities—T20, C80, D50, Ts—from the measured impulse response and the VSVerbs. Focusing on the Just Noticeable Differences (JNDs), both methods provide values within the JND ranges, except for T20. Although the T20 values do not meet the 5% JND criterion, the proposed method yields a value closer to this threshold. For all acoustic quantities, the proposed method yields values closer to, or equal to, those of the measured impulse response compared to the conventional method.

Figure 19 compares the time and frequency responses of the VSVerbs generated by the conventional and proposed methods with the measured impulse response. The time responses highlight the initial characteristics within the first 50 ms. Significant differences are observed between the two methods: while the conventional method (a1) captures the detailed waveform of the impulse response immediately following the direct sound (within 10 ms), its accuracy diminishes after 10 ms. In contrast, the proposed method (a2) reconstructs the impulse response waveform with consistent accuracy throughout the analyzed duration.

Regarding the frequency response (Figure 19 (b)), it is observed that the measured impulse response includes not only the room's acoustic characteristics but also the response of the measurement chain. Specifically, the measured impulse response exhibits a loss of low- and high-end energy due to the frequency limits of the loudspeaker used. This is a typical limitation of IR-based sampling reverbs.

Conversely, VSVerb provides a flat frequency response regardless of whether the conventional or proposed method is used. This demonstrates that VSVerb, which reconstructs the IR using discrete virtual sound sources, can isolate and restore the reverberant atmosphere of the room without any unintended coloration from the measurement equipment. Based on these examinations, we confirm the superiority of the proposed method, particularly regarding temporal and spatial characteristics.

9 Concluding remarks

In this paper, we proposed an adaptive thresholding method within the Speed Detection framework to improve the extraction of virtual sound sources from measured sound intensity data.

The conventional approach, based on a fixed threshold model, was shown to be insufficient for capturing the non-stationary and environment-dependent behavior of real sound fields. In contrast, the proposed method defines the threshold based on the local statistical properties of the spatial moving speed, enabling the robust separation of coherent and incoherent components.

Experimental results demonstrated that the proposed method improves both the temporal and spatial consistency of detected virtual sound sources. This leads to the generation of an SBA reverberation, VSVerb, that more closely matches measured impulse responses. These findings indicate that adaptive, locally-informed thresholding is essential for physically consistent virtual sound source detection and is a key requirement for reliable VSVerb generation.

Future work will focus on extending the method to more complex acoustic environments and further investigating the perceptual impact of the spatial characteristics of the detected sources.

References

[1] M. Nakahara, Y. Nagatomo, and A. Omoto, "VSVerb, a practical method of reproducing spatial reverberation from virtual sound sources captured in a room," Proc, 24th ICA, Gyeongju (2022)

[2] M. Nakahara, T. Kamekawa, and A. Omoto, "Generate acoustic responses of virtual microphone arrays from a single set of measured FOA responses. -Apply to multiple sound sources," Express Paper, AES 157th Convention, NY (2024)

[3] M. Nakahara, A. Omoto, and Y. Nagatomo, "Accurate extraction of dominant reflections from measured sound intensity responses in a room," eB, AES 142nd Convention, Berlin (2017)

[4] Y. Yamasaki and T. Itou, "Measurement of spatial information in sound field by closely located four point microphone method.," Acoust. Sci. & Tech.10, 101-110 (1989)

[5] K. Sekiguchi, S. Kimura S, and T. Hanyu, "Analysis of sound field on spatial information using a four-channel microphone system based on regular tetrahedron peak method," Appl. Acoust. 37, 305-323 (1992)

[6] T. Hanyu and K. Hoshi, "Method for analyzing sound pressure and sound particle velocity using cardioid microphones," Acoust. Sci. & Tech. 45, 2 (2024)

[7] T. Hanyu and K. Hoshi, "Relationship between reflected sound density and mean free path in consideration of room shape complexity," Acoust. Sci. & Tech. 33, 3 (2012)