Audio Engineering Society

Conference Paper

Presented at the AES 6th International Conference on Audio for Virtual and Augmented Reality and Immersive Games
June 30 - July 3, 2026, Sorbonne University/ d'Alembert & IRCAM/STMS, Paris, France

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 The complete manuscript was not peer reviewed.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.

Immersive recording with a virtual 3D microphone array using spatial information of virtual sound sources sampled in a target space.

Masataka Nakahara[1,2], Toru Kamekawa[3], Yasuhiko Nagatomo[4], and Akira Omoto[5,1]

[1] *ONFUTURE Ltd, Tokyo 141-0032 Japan*
[2] *SONA Corporation, Tokyo 164-0013 Japan*
[3] *Tokyo University of the Arts, Tokyo 120-0034 Japan*
[4] *Evixar Inc, Tokyo 104-0033 Japan*
[5] *Kyushu University, Fukuoka 815-8540 Japan*

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

ABSTRACT

This paper proposes a method called V2MA (VSVerb Virtual Microphone Array), which virtually generates spatial room impulse responses (SRIRs) captured by a conventional microphone array using only a set of four impulse responses (IRs) measured by an A-format microphone in a target space. Based on the principles of geometrical acoustics —specifically employing the mirror-image source method— V2MA detects "source intensities" containing acoustic information of sound sources from the *x*-, *y*-, and *z*-directions' sound intensities calculated from the four measured IRs. Subsequently, the locations, strengths, and phase characteristics of the sound sources are estimated. By applying geometrical acoustics, the spatial properties of these virtual sound sources can be updated to match a neighboring receiver position and a desired directivity, thereby yielding SRIRs at any receiver position with arbitrary directivity in the target room.

While immersive recording typically requires large microphone arrays, V2MA enables virtual immersive recording using only the initial A-format microphone measurements. To verify the plausibility of V2MA, this paper compares the responses of the proposed virtual array against a real conventional microphone array using measurements collected in a rehearsal hall under various conditions. The results demonstrate that the waveforms generated by V2MA exhibit a high correlation with those of the real microphone array, with reverberation times within ±5% of the just noticeable difference (JND). Regarding spatial properties, the correlation values between microphones (channels) show reasonable similarities. On the other hand, detailed temporal structures, such as C80, D50, and Ts, do not correspond well to those of the real array, partly due to differences in acoustic characteristics between virtual and real microphones. Future work will focus on addressing these limitations to improve accuracy.

1 Introduction

Immersive audio has emerged as a transformative trend in music production and reproduction. Capturing a musical performance alongside the unique acoustic signature of a specific environment, such as a concert hall, traditionally demands sophisticated microphone techniques. Conventional high-end techniques often require large-scale microphone arrays, necessitating significant equipment overhead and intensive manual labor by recording engineers.

To address these challenges, this paper proposes a method to streamline the workflow of large-scale spatial recording. This approach requires only a single measurement of four impulse responses (IRs) using an A-format microphone to emulate the spatial room impulse responses (SRIRs) of an arbitrary microphone array. By calculating instantaneous sound intensities in the *x*, *y*, and *z* axes from the four measured IRs, the positions and acoustic properties of virtual sound sources can be identified.

In geometrical acoustics, a virtual sound source represents a reflection as sound originating from a point source outside the physical boundaries of the room. The proposed method reconstructs 6DoF, 4π (full-sphere) reverberation by sampling these virtual sound sources, allowing for the synthesis of SRIRs as if they were captured by a physical microphone array. This technique is called V2MA (VSVerb Virtual Microphone Array) [1]. This paper examines the acoustical plausibility of V2MA by comparing the acoustic properties and waveforms obtained by V2MA against those of real microphone arrays.

2 Processing Flow of V2MA

Figure 1 illustrates an overview of the processing flow of V2MA. The SRIRs captured by virtually placed microphones are obtained by updating the VSVerb response, which is generated from four IRs measured in a target room using an A-format microphone. The processing flow is as follows:

1. **Measurement:** Four-channel A-format impulse responses (LFU, RFD, LBD, RBU) are measured.
2. **Filtering:** The responses are band-limited to suppress out-of-band noise.
3. **B-format Conversion:** The signals are converted to B-format (W, X, Y, Z), and the Hilbert transform is applied.
4. **Intensity Calculation:** Relative sound intensities (I_x, I_y, I_z) are calculated. Since the four capsules of the A-format microphone (LFU, RFD, LBD, RBU) exhibit nearly identical acoustic responses, the obtained W, X, Y, and 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 [2].
5. **Temporal Averaging:** A moving average with a 0.4ms window is applied.
6. **Source Detection:** Source intensities are identified using the Speed Detection method [3].
7. **Phase Estimation:** The phase characteristics (+/-) are estimated [3].
8. **VSVerb:** 4π SBA reverb, VSVerb, is generated.
9. **Postprocessing-1:** The positions of the virtual sound sources are updated with reference to the new receiver positions (i.e., microphone positions of a virtual microphone array). These updated positions are obtained by translating the original virtual sources in the opposite direction of the receiver's displacement.
10. **Postprocessing-2 (optional):** If moving the real sound source position (i.e., the position of the dry signal) is desired, re-measurement is typically required. However, if a relatively large margin of error is acceptable, the positions of the virtual sound sources can also be updated with reference to the new real source position [4]. This approach is treated as an optional method and is not employed in this study.
11. **Postprocessing-3:** The strengths of the virtual sources are updated in inverse proportion to the relative distance changes between the original and updated virtual sound source positions with respect to the new receiver positions.
12. **Postprocessing-4:** The strengths of the virtual sound sources are weighted to match the directivities of the microphones in a virtual microphone array:

Omni	1 (no change)	(1)
Figure 8	$\cos(\theta+\theta_0)\cos(\varphi+\varphi_0)$	(2)
Cardioid	$0.500+0.500\cos(\theta+\theta_0)\cos(\varphi+\varphi_0)$	(3)
Super Card.	$0.375+0.625\cos(\theta+\theta_0)\cos(\varphi+\varphi_0)$	(4)
Hyper Card.	$0.250+0.750\cos(\theta+\theta_0)\cos(\varphi+\varphi_0)$	(5)

(θ and φ : elevation and azimuth angles of virtual sources; θ_0 and φ_0 : elevation and azimuth angles of the receiver)

13. **V2MA:** The spatial properties of the virtual sound sources are converted into time-domain responses to yield the final V2MA responses. These are then convolved with a source signal in each channel.

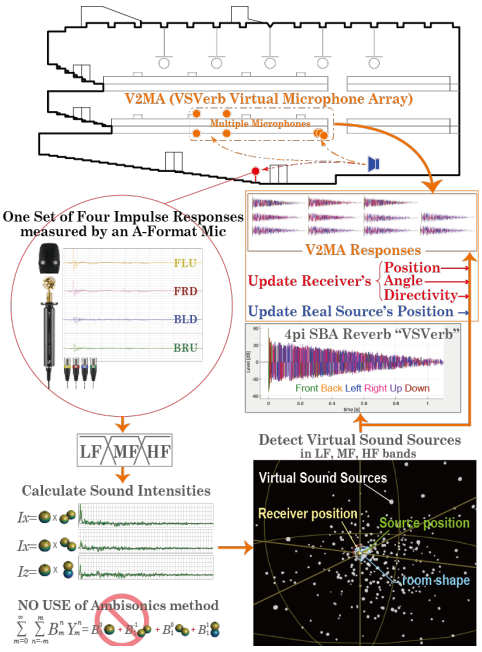


Figure 1. Overview of the V2MA processing flow.

3 Experiments

3.1 IR measurements

The impulse responses were measured using an A-format microphone and a 5.0.4ch microphone array in Hall 6 at the Tokyo University of the Arts (Figure 2). The 5.0.4ch configuration consisted of five mid-layer channels (L, R, C, Ls, Rs), no LFE channel, and four height channels (Ltf, Rtf, Ltr, Rtr).

Hall 6 is a wooden rehearsal hall featuring a flat floor across approximately two-thirds of the front area. The room dimensions, volume, and total surface area were (W, L, H) ≈ (15 m, 25 m, 9 m), V ≈ 3543 m^3, and S_{all} ≈ 1490 m^2, respectively. The reverberation time was approximately 1.5 s at 500 Hz. A Sennheiser Ambeo VR Mic was employed as the A-format microphone, and a swept-sine signal was played back from a B&K Type 4295 omnidirectional loudspeaker.

Measurements were conducted at five positions for the sound source (loudspeaker) and six positions for the receiver (A-format microphone and 5.0.4ch microphone array). For some of the center and height microphones in the 5.0.4ch array, the directivity types were varied during the measurements.

In this paper, three of these configurations — Measurement-A, -B, and -C— were selected to examine the similarities between the responses of V2MA and the real microphone array (Table 1). Compared to Measurement-A, the source position was moved 1 m to the left and closer to the receiver for Measurement-B (Figure 3). For Measurement-C, the directivities of the center and height microphones were altered. Additionally, in Measurement-C, the height microphones were lowered to the same height as the other microphones (Figure 4).



Figure 2. The Hall 6, Tokyo University of the Arts.

Table 1. Selected measurement conditions.

		Measurement-A	Measurement-B	Measurement-C
Loudspeker position		S4	S2	S1
A-format mic. position		P1	←	P4
5.0.4ch mic. array	L / R / Ls / Rs	Omni DPA 4006	←	←
	C	Omni AKG C414	←	Hyper-Card. AKG C414
	Ltf / Rtf / Ltr / Rtr	Cardioid SCHOEPS CMC64 main ch. +2m	← ←	← main ch. ±0m

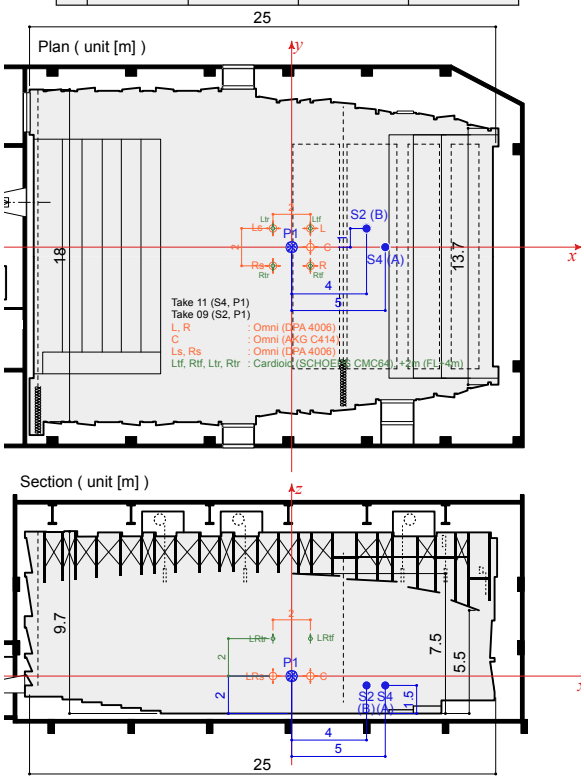


Figure 3. Measurement-A and -B.

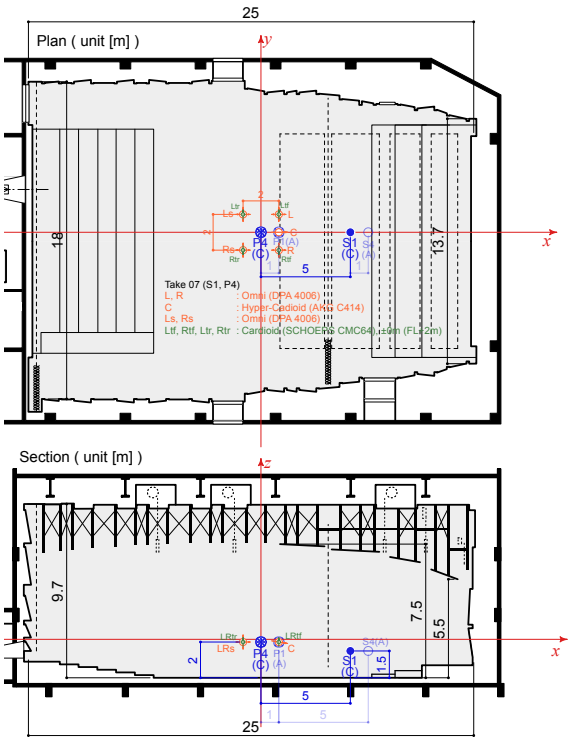


Figure 4. Measurement-C.

3.2 Generating VSVerb

The IRs measured by the A-format microphone were filtered into three bands: LF (90–500 Hz), MF (200 Hz–3 kHz), and HF (1 kHz–8 kHz). Subsequently, the 4π scene-based audio (SBA) reverb, VSVerb, was generated according to the processing flow of steps 1–8 in Section 2.

Figure 5 presents spatial scattergrams of the virtual sound sources obtained from Measurement-C. These scattergrams were derived from the *x*-, *y*-, and *z*-directions' sound intensities calculated from the measured IRs. In Hall 6, 927, 1487, and 1421 virtual sound sources were sampled in the LF, MF, and HF bands, respectively. According to the *x*-*z* views, reflections between the floor and ceiling contribute significantly to the longer reverberation time of Hall 6. By converting the spatial properties of the virtual sound sources into the time domain, the VSVerb is obtained (Figure 6).

The VSVerbs obtained here (Figure 6) demonstrate that the dynamic ranges between the direct sound and the lowest-level reflection are approximately 70 dB, and the time gap between the direct sound and the latest reflection is around 0.98 s. In this figure, six colors indicate the arrival directions of the reflection components: front, back, left, right, up and down. Certain differences in spatial characteristics are observable among the LF, MF, and HF bands.

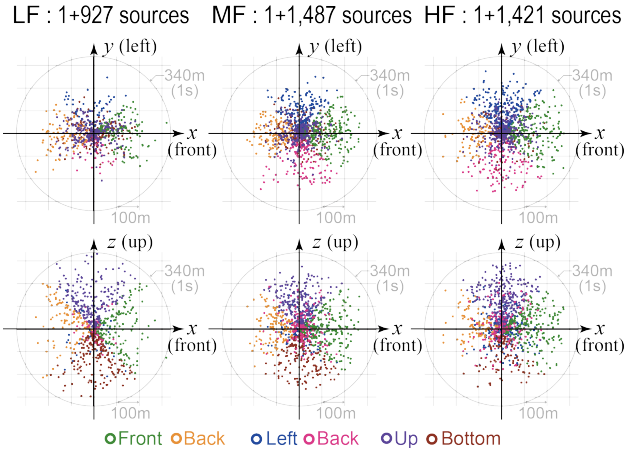


Figure 5. Virtual Sound Sources (Measurement-C).
from left, LF : 90-500Hz, MF : 200-3kHz, and HF : 1k-8kHz.

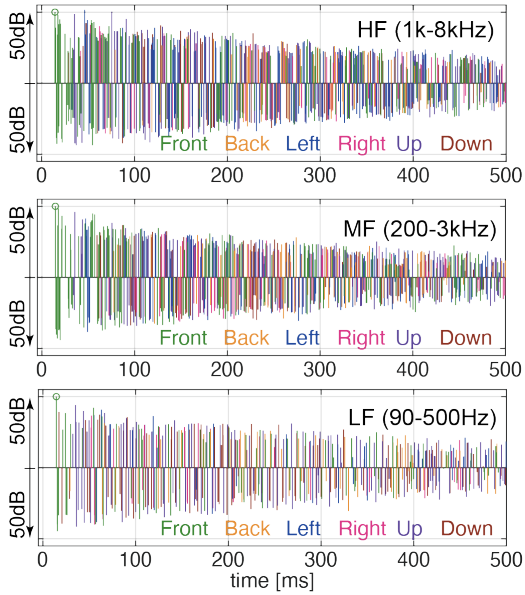


Figure 6. Generated VSVerbs (Measurement-C).
from Bottom LF : 90-500Hz, MF : 200-3kHz, and HF : 1k-8kHz.

3.3 Synthesizing V2MA

SRIRs captured by a directional microphone located at a different position from the initial measurement point can be reconstructed by updating the spatial information of the virtual sound sources (following the processing flow of steps 9–13 in Section 2). Figure 7 shows the V2MA responses in the MF band, representing the L and C channel responses of the 5.0.4ch microphone array used in Measurement-C.

The six colors (front, back, left, right, up, down) representing the reflection components indicate that both the L and C channels exhibit a high concentration of frontward reflections in the very

early stage, which subsequently transition to encompass reflections from all directions.
In Measurement-C, omni-directional and hyper-cardioid microphones are used for the L and C channels of the real microphone array, respectively. This directivity difference is clearly observable in the envelopes of the V2MA responses, characterized by a broader profile for the L channel and a narrower profile for the C channel.

4 Discussion

4.1 Waveform Correlation

V2MA is expected to generate realistic responses similar to those of an actual microphone array. Figure 8 illustrates a comparison of nine IRs, contrasting the reconstructed V2MA responses against the real microphone array measurements in the MF band. In this figure, the responses are presented as absolute values in the upper (V2MA responses) and lower (measured IRs) regions.
To quantify the waveform similarities between them, the cross-correlation Xcrr(n) was calculated using their absolute values:

$$X_{crr}(n) = \sum_i \left|\overline{p_V(i)}\right| |p_m(i+n)| \quad (6)$$

$$X_{crr}|_{max} = max\left(X_{crr}(n)\right) \quad (7)$$

where, $p_V(i)$ is the V2MA response.
$p_m(i)$ is the real microphone array's response.

Figure 9 shows the maximum cross-correlation values, $X_{crr}|_{max}$. In the LF and MF bands, the results exhibit relatively high correlations (≥ 0.75, indicated by the gray area) across all channels under all measurement conditions. Although the correlation is slightly lower in the HF band, it remains above 0.5.
Because V2MA eliminates measurement-specific artifacts —such as the frequency response of the transducer and measurement noise— and subsequently re-synthesizes the IRs based on a prescribed directivity model, discrepancies between the resulting waveforms and physical measurements are inherently expected. It should be explicitly noted that the current V2MA implementation models only idealized directivity patterns; it does not account for other physical constraints of real microphones, such as frequency- or distance-dependent directivity variations and on-axis frequency responses. Consequently, these results demonstrate that the V2MA responses achieve a reasonable and acceptable range of correlation, avoiding both unrealistically high agreements and poor correspondences.

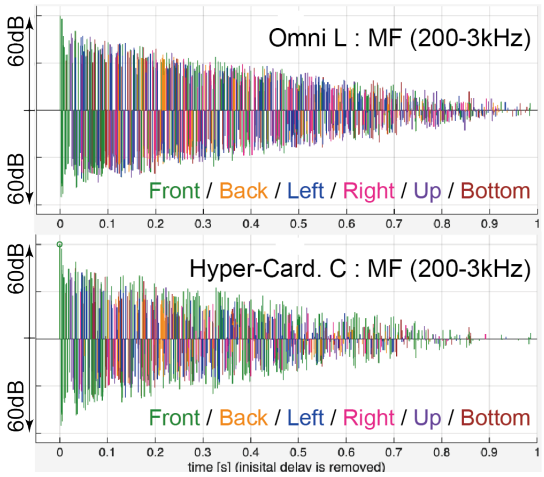


Figure 7. V2MA (Measurement-C, MF : 200-3kHz). from top, Left (Omni) and Center (Hyper-Card.) channels.

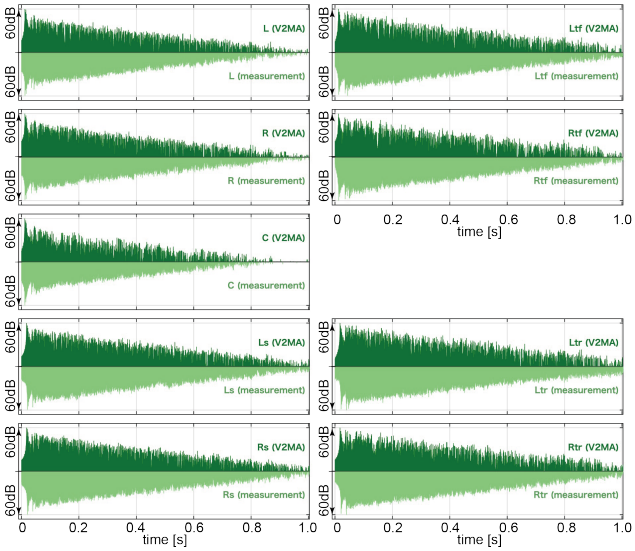


Figure 8. Responses by V2MA (upper dark graphs) and Real microphone array (lower light graphs). Measurement-C, MF : 200-3kHz.

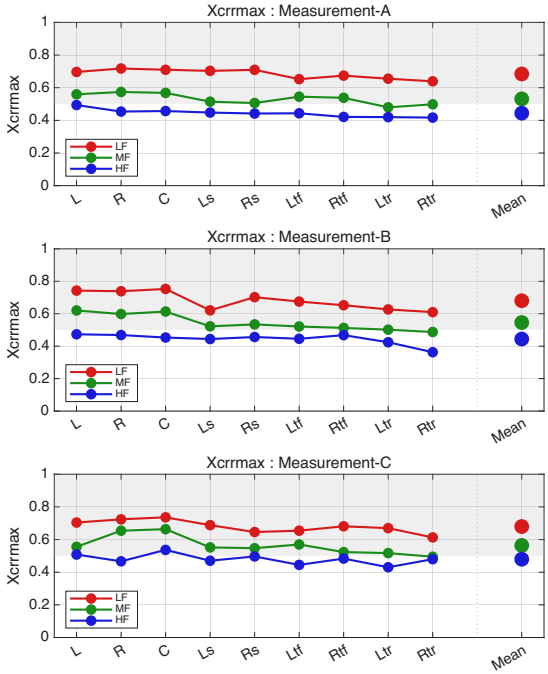


Figure 9. $X_{crr}|_{max}$
Maximum value of cross−correlations between V2MA and measured responses.

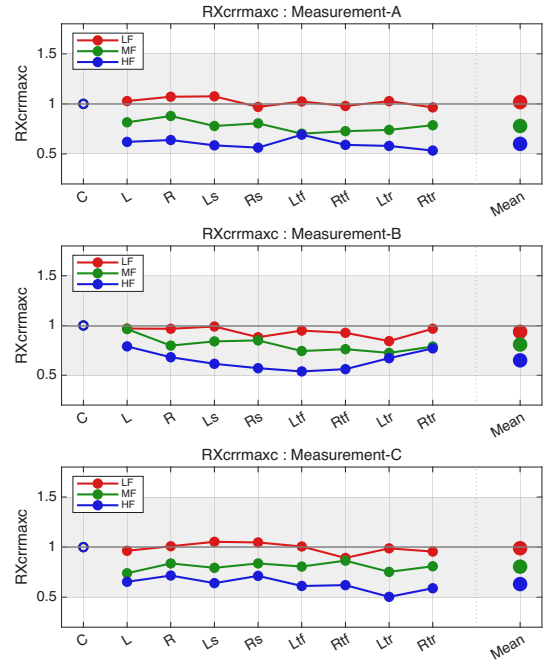


Figure 10. $R_{Xcrr,c}$
V2MA / measured ratio of the maximum values of cross−correlations between center and other channels.

4.2 Channel Correlation

Acoustic spaciousness and stereophonic images provided by a microphone array are closely related to the acoustical correlations between channels [5,6]. To compare the spatial characteristics of the responses provided by V2MA and the real microphone array, the cross-correlations between the C channel and other channel responses, $X_{crr,Vc}(n)$, $X_{crr,mc}(n)$, were calculated using the absolute values of the V2MA responses and the measured IRs, respectively.

$$X_{crr,Vc}(n) = \sum_i \left|\overline{p_{Vc}(i)}\right| |p_V(i+n)| \quad (8)$$

$$X_{crr,mc}(n) = \sum_i \left|\overline{p_{mc}(i)}\right| |p_m(i+n)| \quad (9)$$

$$R_{Xcrr,c} = \frac{max\left(X_{crr,Vc}(n)\right)}{max\left(X_{crr,mc}(n)\right)} \quad (10)$$

where, $p_V(i)$ is the V2MA response.
$p_m(i)$ is the real microphone array's response.
$p_{Vc}(i)$ is the center channel's response of $p_V(i)$.
$p_{mc}(i)$ is the center channel's response of $p_m(i)$.

Figure 10 shows the ratios of the maximum cross-correlation values of the V2MA responses to those of the real microphone array responses, RXcrr,c. In the figure, a value of 1 indicates no difference between the V2MA and real array responses. For instance, a value of 0.5 indicates that the correlation between the C channel and other channels in V2MA is half that of the real array. The gray area in the figure indicates that the differences between the V2MA and real array responses are within ±50%.

The results in Figure 10 demonstrate that V2MA provides almost the same channel correlations as the real array in the LF band. Although the discrepancies become larger at higher frequency bands (LF < MF < HF), the differences remain within 50%. This indicates that V2MA replicates the spatial characteristics of a real microphone array relatively well. Since most scores fall below 1.0, V2MA tends to generate slightly less correlated responses between channels than the real array produces, particularly at higher frequencies.

4.3Acoustic Parameters

To verify the temporal accuracy of the V2MA responses, four acoustic parameters—T20 [s], C80 [dB], D50, and Ts [ms]—were calculated for both the V2MA and real microphone array responses. The calculated results were evaluated based on their respective Just Noticeable Differences (JNDs) [7]. Figures 11–15 present the differences in these acoustic parameters between the V2MA and real microphone array responses, with the JND criteria indicated by the gray areas.

As Figure 11 (T20) indicates, V2MA replicates room reverberation well. Although some values in the LF band fall outside the JND ranges, these discrepancies are attributed to the insufficient diffuseness of the hall and/or acoustic diffractions around the receiver. It is well established that accurate in-situ measurements of reverberation remain challenging, particularly at low frequencies.

On the other hand, the other parameters (C80 [dB], D50, and Ts [ms]) do not exhibit close matching between the V2MA and real array responses. Focusing on the differences among measurement conditions A, B, and C, the performance appears to improve in the order of C > B > A. However, the underlying physical mechanism for this trend has not yet been identified, necessitating further investigation.

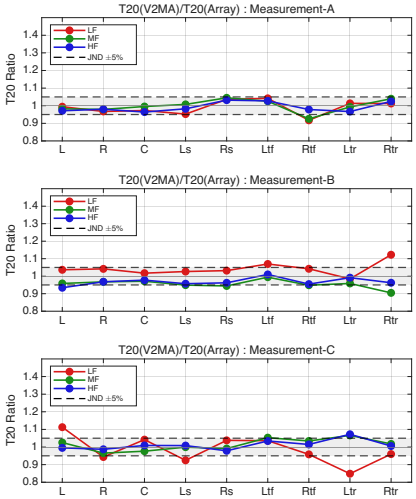


Figure 11. Ratio of T20 : V2MA / Mic. array.

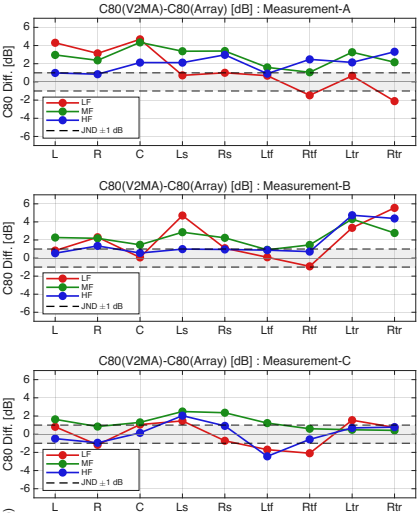


Figure 12. Difference of C80 : V2MA - Mic. Array.

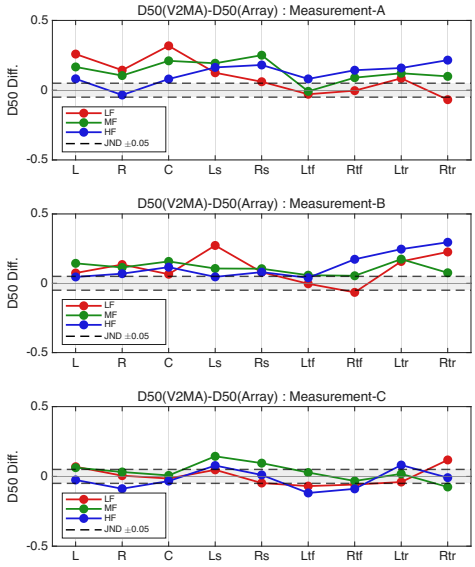


Figure 13. D50 : V2MA - Real array Difference.

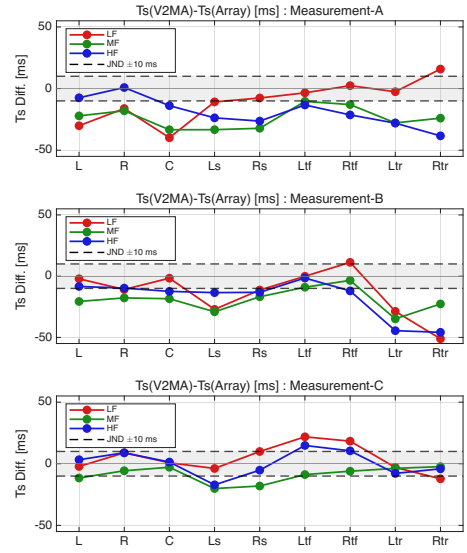


Figure 14. Ts : V2MA - Real array Difference.

4.4 Rating

Tables 2–5 present the similarity ratings between the V2MA and real microphone array responses. The performance is categorized into Good, OK, and NG based on the criteria illustrated in Figure 15. For the acoustic parameters (T20, C80, D50, and Ts), the rating values represent the proportion of calculated data points that fell within the JND threshold ranges. The values of $R_{Xcrr,c}$ in Table 2 exhibit high ratings across all measurement conditions. Consequently, it is confirmed that V2MA performs well in replicating the spatial properties of a real microphone array.

Similarly, $X_{crr}|_{max}$ in Table 3 demonstrates good ratings, confirming that V2MA provides waveforms that closely match those of a real microphone array. This waveform similarity directly contributes to achieving the high ratings observed in replicating the room reverberation time (T20), as presented in Table 4. On the other hand, lower ratings are observed for the remaining acoustic parameters under most conditions(Table 5).

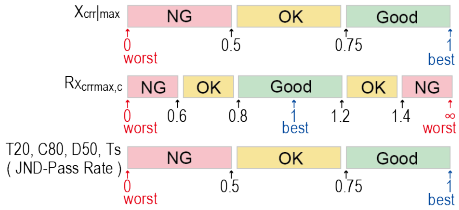


Figure 15. Rating Criteria.

Table 2. Rating results : $R_{Xcrr,c}$

Metric	Pos.	LF (90-500Hz)	MF (200-3kHz)	HF (1k-8kHz)
$R_{X_{crr,c}}$	A	1.018 [Good]	0.779 [OK]	0.601 [OK]
	B	0.937 [Good]	0.809 [Good]	0.650 [OK]
	C	0.989 [Good]	0.805 [Good]	0.631 [OK]

Table 3. Rating results : $X_{crr}|_{max}$

Metric	Pos.	LF (90-500Hz)	MF (200-3kHz)	HF (1k-8kHz)
$X_{crr}\|_{max}$	A	0.684 [OK]	0.531 [OK]	0.444 [NG]
	B	0.680 [OK]	0.546 [OK]	0.444 [NG]
	C	0.680 [OK]	0.564 [OK]	0.480 [NG]

Table 4. Rating results : T20 [s]

Parameter	Pos.	LF (90-500Hz)	MF (200-3kHz)	HF (1k-8kHz)
T20 (+-5%)	A	8/9 [Good]	8/9 [Good]	9/9 [Good]
	B	7/9 [Good]	5/9 [OK]	8/9 [Good]
	C	5/9 [OK]	7/9 [Good]	8/9 [Good]

Table 5. Rating results : C80 [dB], D50, Ts [ms]

Parameter	Pos.	LF (90-500Hz)	MF (200-3kHz)	HF (1k-8kHz)
C80 (+-1 dB)	A	3/9 [NG]	0/9 [NG]	3/9 [NG]
	B	4/9 [NG]	1/9 [NG]	6/9 [OK]
	C	3/9 [NG]	4/9 [NG]	7/9 [Good]
D50 (+-0.05)	A	2/9 [NG]	1/9 [NG]	1/9 [NG]
	B	1/9 [NG]	0/9 [NG]	3/9 [NG]
	C	5/9 [OK]	5/9 [OK]	4/9 [NG]
Ts (+-10ms)	A	4/9 [NG]	0/9 [NG]	2/9 [NG]
	B	3/9 [NG]	2/9 [NG]	3/9 [NG]
	C	6/9 [OK]	6/9 [OK]	6/9 [OK]

Table 6. Summary of the Rating results.

Metric	Description	LF (90-500Hz)	MF (200-3kHz)	HF (1k-8kHz)
$X_{crr}\|_{max}$	waveform similarity	OK	OK	NG
$R_{X_{crr,c}}$	channel correlation	Good	OK	OK
T20	Rev.time (+-5%)	OK	OK	Good
C80	Clarity (+-1 dB)	NG	NG	NG
D50	Defn. (+-0.05)	NG	NG	NG
Ts	Ctr.time (+-10ms)	NG	NG	NG

5 Conclusions

The comparison between the V2MA responses and physical measurements from real microphone arrays demonstrates the utility of V2MA as a virtual recording tool that leverages the spatial sampling and reconstruction capabilities of a microphone array. Table 6 summarizes the overall scores of the current version of V2MA evaluated in this study. V2MA reliably reproduces room-representative acoustic properties (T20) and the spatial structure of the sound field captured by a microphone array, confirming its effectiveness for auralizing immersive audio content without the need to construct physical microphone arrays. This demonstrates a clear reduction in the intensive equipment setup and manual labor required of recording engineers on site. On the other hand, non-negligible differences were observed in detailed temporal structures, such as C80 [dB], D50, and Ts [ms]. Future work will focus on developing physical modeling of the frequency- and distance-dependent directivities of real microphones to further improve the accuracy and plausibility of V2MA.

References

[1] M. Nakahara, A. Omoto, and Y. Nagatomo, “Virtual 3D microphone arrays by virtual sound sources detected from an FOA response,” Proc, AES International Conference on Spatial & Immersive Audio, Huddersfield (2023)

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

[3] M. Nakahara, Y. Nagatomo, and A. Omoto, “Generate 4pi SBA reverberation from virtual sound sources detected from x-y-z sound intensities. - Improvement of the source detection method.,” Express Paper, AES 160th Convention, Copenhagen (2026)

[4] 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)

[5] R. Kiyan, S. Preihs, and J. Peissig, “Signal Dependencies of the Immersion Sweet Area in Multichannel Loudspeaker Reproduction,” J. Audio Eng. Soc., Vol. 73, No. 5 (2025)

[6] T. Kamekawa and A. Marui, “Correlation between middle and top layer loudspeaker signals and the listening area in 3D audio reproduction,” Paper, AES 158th Convention, Warsaw (2025)

[7] “Acoustics -Measurement of room acoustic parameters- Part 1: Performance spaces,” ISO Standard 3382-1 (2009)