Audio Engineering Society

# Convention e-Brief 623

Presented at the 149th Convention
Online, 2020 October 27-30



# Sound Quality Improvement of MPEG-H 3D Audio Encoder

Akifumi Kono[1], Hiroyuki Honma[1], and Toru Chinen[1]

[1] *Sony Corporation, R&D Center Tokyo Laboratory 20, Tokyo Japan*

Correspondence should be addressed to Author (Toru.Chinen@sony.com)

**ABSTRACT**

In 2019, Sony launched 360 Reality Audio, which provides a new music experience using object-based spatial audio technology. Object-based audio contains information on time-varying object loudness and location and audio data, which are transmitted to playback devices, and then rendered and played back. It was reported in [1] that object locations affect the subjective sound pressure perception depending on the direction of the sound source. In this e-brief, we present an approach to increase the sound quality by considering the loudness and locations of objects. We perform a subjective listening test for three test items. The results indicate that two items had statistically significant differences in sound quality.

## 1 Introduction

Sony announced 360 Reality Audio (360RA) [2] in CES 2019, and the service was launched in October 2019. 360RA empowers the creation of immersive music experience, such as feeling as if you were at a live concert, which uses object-based audio technology. In object-based audio, vocals, guitars, and such instruments are considered "objects," as shown in Figure 1. The object comprises audio data and metadata, i.e., the information of the position in a three-dimensional (3D) space and gain information. Therefore, metadata and audio data are transmitted to the playback devices and rendered for playback.

Furthermore, it was reported in [1] that the subjective perception of sound pressure differs depending on the direction of arrival of the sound source in the 3D space. In [1], there is an experiment regarding the subjective perception of sound for 29 subjects. Each subject subjectively adjusts the volumes of the other 31 directions to make them equal level to the volume of the front center directions. The results are shown in Figure 2. We refer to this difference in subjective loudness sensitivity as 3D psychoacoustics and refer to the mean of adjusted volumes in Figure. 2 as compensation gain. In this e-brief, we consider improving the sound quality by applying 3D psychoacoustics to the bit-allocation of encoder. First, we describe how to improve the sound quality of the encoder by 3D psychoacoustics. Second, we describe the experiments and results of the encoder using 3D psychoacoustics. Finally, we discuss the experimental results.

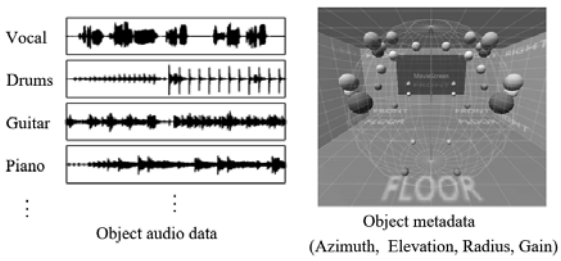


Figure 1. Object-based audio

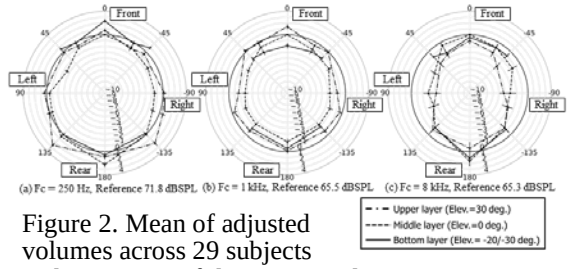


Figure 2. Mean of adjusted volumes across 29 subjects and its 95% confidence intervals per CF
(Nakai, A. ***AES Convention:*** *148 (May 2020)* ***eBrief:****581 [1]*)

## 2 Proposed Method

In general, encoders use perceptual entropy (PE) [3] for bit-allocation. In the case of two-channel encoder, PE is calculated by applying the equivalent loudness curve to the L and R channels of the stereo data. Similarly, one can constitute an object-based audio encoder by applying the equivalent loudness curve to all the objects for PE calculation. In this e-brief, for comparison with other codecs, we refer to this object-based audio encoder as "General," which conforms to the MPEG-H 3D Audio standard [4]. However, considering 3D psychoacoustics, bit-allocation might become more efficient by using each equivalent loudness curve along the direction of each object.

In this e-brief, we consider improving the sound quality of an object-based audio encoder by compensating an equal loudness curve that utilizes the position information and gain information of the metadata. This object-based audio encoder is referred to as "Proposal" for comparison with "General." The block diagram of "Proposal" is shown in Figure. 3.

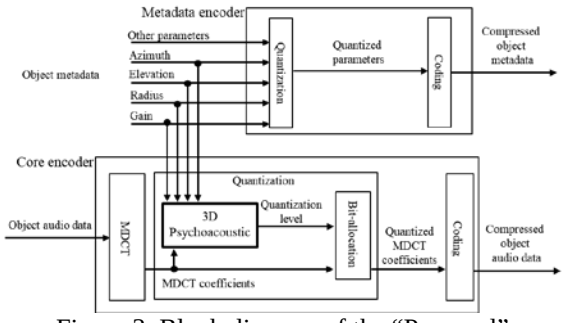


Figure 3. Block diagram of the "Proposal"

### *Interpolation in frequency and direction*

In [1], there are 93 compensation gains for three frequencies (250 Hz, 1 kHz, and 8 kHz) and 31 directions. The compensation gains are interpolated in frequency and direction to cover the entire frequency and direction ranges. The interpolated gains are applied to the encoder, which then becomes "Proposal." We do not describe how to interpolate compensation gains owing to space limitations.

## 3 Preliminary Experiment

We conducted a preliminary experiment to confirm that the bit-allocation of "Proposal" changed from that of "General," as we had intended. For this experiment, we prepared three test items, each of which included three objects and was encoded at the bitrate of 192(64 * 3) kbit per second (kbps). The audio data of object for each test item comprises three instruments, namely, vocal, bass, and hi-hat (see Figure 4). From Table 1, it is seen that different metadata of object are used for the test items. An object is located along the direction that has the highest subjective sound-pressure sensitivity, i.e., +90 degree/+6.0 dB for test item 1, +60 degree/+6.0 dB for test item 2, and +90 degree/+6.0 dB for test item 3 (called "target object"), (see Table 1).

We encoded each test item using both "General" and "Proposal," respectively, and we then obtained the PE and bitrate of each object. The PE and bitrate values are shown in Figure 5. We calculated the average PE for three objects. The ratio of the PE of each test item to the average PE is shown in the upper part of Figure 5. Similarly, the bitrate for each object is shown in the lower part of Figure 5.

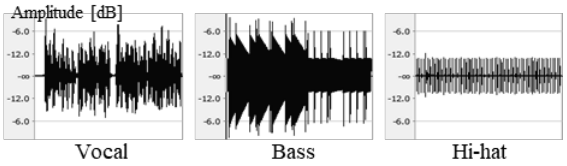


Figure 4. Audio waveform

Table 1. Conditions of metadata

| Item | Metadata | Object | | |
|---|---|---|---|---|
| | | Vocal | Bass | Hi-hat |
| 1 | Azim. [°] | +90 | | |
| | Elev. [°] | 0 | 0 | |
| | Gain [dB] | +6.0 | | |
| 2 | Azim. [°] | | +60 | |
| | Elev. [°] | 0 | 0 | 0 |
| | Gain [dB] | | +6.0 | |
| 3 | Azim. [°] | | | +90 |
| | Elev. [°] | 0 | | +30 |
| | Gain [dB] | | | +6.0 |
| All | Radius | | 1.0 | |

From Figure 5, it is seen that the PEs for the target object increase from "General" to "Proposal." The highest increase in bitrate is +13.6 kbps, which is hi-hat for test item 3. Therefore, we confirmed that the bit-allocation of "Proposal" changed from "General," as we had intended.

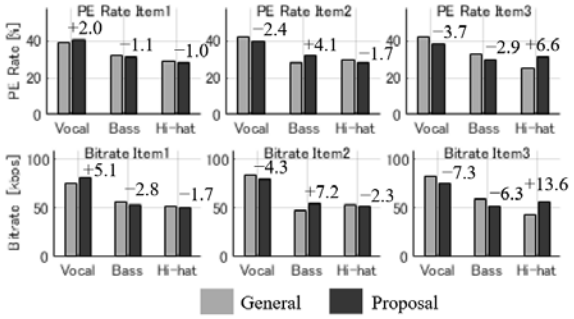


Figure 5. Result of PE rate and bitrate

## 4 Subjective Listening Test

We conducted a subjective listening test to confirm whether the subjective sound quality changed from "General" to "Proposal." The condition of this test, layout of the loudspeakers used, and image of the listening room are presented in Table 2, Table 3, and Figure 6, respectively. The test methodology was according to the Recommendation ITU-R BS.1116-3 [5], in which the subject is requested to listen to three sound sources, "Reference" (original sound), "A," and "B," and then score the subjective sound qualities of A and B in comparison with the Reference. Either A or B is the Reference and the other is a coded sound (In this e-brief, General or Proposal), and the subject is requested to conduct a listening test without knowing what he/she is listening to, A or B, according to subjective five-grade scale presented in Table 4.

We used three music sources from 360RA services. A total of 22 subjects participated in this subjective listening test.

Table 2. Conditions of the subjective listening test

| | | |
|---|---|---|
| Test methodology | ITU-R BS.1116-3 (same as MPEG audio) | |
| Presentation | 13 Loudspeakers | |
| No. of subjects | 22 (Expert listeners) | |
| Bitrate | 1536 kbps (24objects) | |
| Two systems under the test | 1 | - Hidden reference (Original)<br>- General |
| | 2 | - Hidden reference (Original)<br>- Proposal |
| Three test items | Pop music (24 objects) | 4. Verse (Normal)<br>5. Climax (Loud)<br>6. Solo (Quiet) |
| Listening room | Length<br>Width<br>Height | 7.4 [m]<br>5.0 [m]<br>3.5 [m] |

Table 3. Layout of 13 loudspeakers

| Index | Azim. [°] | Elev. [°] | Rad. [m] |
|---|---|---|---|
| Middle-1 | 0 | 0 | 1.85 |
| Middle-2 | 30 | 0 | 1.85 |
| Middle-3 | −30 | 0 | 1.85 |
| Middle-4 | 110 | 0 | 1.85 |
| Middle-5 | −110 | 0 | 1.85 |
| Upper-1 | 0 | 35 | 1.85 |
| Upper-2 | 30 | 35 | 1.85 |
| Upper-3 | −30 | 35 | 1.85 |
| Upper-4 | 110 | 35 | 1.85 |
| Upper-5 | −110 | 35 | 1.85 |
| Bottom-1 | 0 | −20 | 1.95 |
| Bottom-2 | 30 | −20 | 1.95 |
| Bottom-3 | −30 | −20 | 1.95 |

Table 4. Subjective five-grade scale
Recommendation ITU-R BS.1116-3

| Impairment | Grade |
|---|---|
| Imperceptible | 5.0 |
| Perceptible, but not annoying | 4.0 |
| Slightly annoying | 3.0 |
| Annoying | 2.0 |
| Very annoying | 1.0 |

Figure 6. Listening room

The results of the subjective listening test are shown in Figure 7. The mean scores and their 95% confidence intervals for 22 subjects are shown in the eight graphs of Figure 7: two for test item 4, two for test item 5, two for test item 6, and two for all the test items. The graphs of each test item show the evaluation results of Reference and General, and Reference and Proposal. In addition to the evaluation results, we obtained bitrates: 24 of "General" and 24 of "Proposal" for each test item. For the three test items, in total, there are 144 bitrates (= 48 bitrates * 3 test items). In Figure 7, as for the 95% confidence intervals, the upper values of "General" did not cross 5.0 for test items 4, 5, and 6. The upper values of "Proposal" did not cross 5.0 for test item 4 but did cross 5.0 for test items 5 and 6. For all the test items, the mean of "Proposal" was higher than that of "General." For each test item, the highest increase and highest decrease in bitrate from "General" to "Proposal" are presented in Table 5.

Table 5. Bitrate between "General" and "Proposal"

| Item | Object | General [kbps] | Proposal [kbps] | Difference [kbps] |
|---|---|---|---|---|
| 4 | Vocal | 89.4 | 94.8 | +5.4 |
| | Brass | 32.4 | 16.8 | −15.6 |
| 5 | Vocal | 47.2 | 56.2 | +9.0 |
| | Brass | 43.3 | 27.3 | −16.0 |
| 6 | Vocal | 90.3 | 107.9 | +17.6 |
| | Drums | 215.3 | 184.9 | −30.4 |

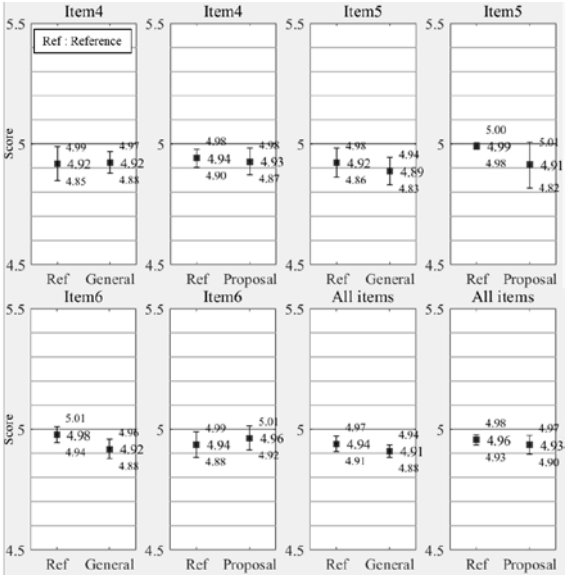


Figure 7. Mean of the score across 22 subjects and its 95% confidence intervals

## 5 Discussion

### *Listener post-screening*

In the subjective listening test in this e-brief, the test methodology of BS.1116-3 was used, in which it is recommended to apply "listener post-screening" to the results of the listening test to evaluate the listener's expertise. However, in our listening test, almost all the subjects were excluded when we applied post-screening to the results of our listening test. Here we consider why we obtain the results like this and cite two reasons. The first is that the scores of "General" and "Proposal," respectively, were nearly 5.0. The second is that many subjects scored below 5.0 for Reference. Therefore, we made the following assumptions.

First, the sound qualities of "General" and "Proposal," respectively, are close to that of the original sound. Second, many subjects may attempt to score below 5.0 on "General" or "Proposal." If the assumptions hold, it is invalid to say that the subjects did not score according to the test methodology. In this e-brief, considering these assumptions and the result of the post-screening, we

decided not to apply post-screening to the results of our listening tests. However, there is still no evidence for this assumption, and we need more future studies to confirm whether this assumption holds. A further study might develop a different subjective listening test methodology for obtaining a high-quality audio codec.

### *Sound quality improvement*

As for "Proposal," the upper value of the 95% confidence intervals for test items 5 and 6 crossed 5.0. However, for "General," it did not cross 5.0. This indicates that "Proposal" improves the sound quality by incorporating 3D psychoacoustics, i.e., equivalent loudness compensation designed for frequencies and directions. The improved sound qualities for test items 5 and 6 for "Proposal" might be attributed to the significant increase in the bitrate for some objects for those test items. However, form Table 5, it is seen that the bitrates for some objects are decreased. If the listeners mainly listen to objects with reduced bitrates, the upper value of the 95% confidential interval may not cross 5.0. As for the Pop music that we used in our experiment, vocal is main object, thereby improving the sound quality. The results for Jazz or Classic might get worse because listeners do not always listen to objects with increased bitrate. It is difficult to use "Proposal" for encoding without knowing the objects that are mainly listened to.

As for "General," the 95% confidence intervals of "General" for the three test items did not cross 5.0. According to BS.1116-3, there are statistically significant differences between "General" and the original sound. The actual sound qualities of "General" and the original sound, however, may not be significantly different than each other because the 95% confidence intervals of Reference did not cross 5.0.

## Conclusion

In this e-brief, we considered improving the sound quality of object-based audio encoder by compensating the equal loudness curve, which utilizes the position information and gain information (loudness) of the metadata. We conducted a preliminary experiment to confirm that the bit-allocation of "Proposal" changed from "General," as we had intended. We conducted a subjective listening test for three test items to confirm whether the subjective sound quality changed from "General" to "Proposal" by using 95% confidence intervals and mean values of test methodology of BS.1116-3. As for the 95% confidence intervals, the upper values of "General" did not cross 5.0 for all the three test items. However, the upper values of "Proposal" did cross 5.0 for two test items. As for the mean values, the mean of "Proposal" was higher than that of "General" for all the three test items.

There still remain two concerns in this e-brief. The first is that we did not apply post-screening in our listening test because almost all the subjects were excluded. The second is that the 95% confidence intervals of Reference did not cross 5.0. We will continue our study to resolve these concerns and seek different subjective listening test methodologies for obtaining a high-quality audio codec.

## References

[1] A. Nakai, M. Tsuji, T. Chinen "Directional Dependency of Subjective Sound Pressure Perception on Three-Dimensional Sound," in *148th AES Convention*, (2020).

[2] 360 Reality Audio Official Web Site. https://www.sony.com/electronics/360-reality-audio

[3] J. D. Johnston, "Estimation of Perceptual Entropy Using Noise Masking Criteria," in Proc. ICASSP, pp. 2524–2527, (1988).

[4] ISO/IEC 23008-3:2019, "High Efficiency Coding and Media Delivery in Heterogeneous Environments–Part 3: 3D audio,".

[5] Recommendation ITU-R BS.1116-3 (02/2015), "Methods for the Subjective Assessment of Small Impairments in Audio Systems."