Compensation of unmasking noise for scalable transmission of three-dimensional multichannel sound

Akio Ando\textsuperscript{1,*} and Takafumi Fujita\textsuperscript{2,†}

\textsuperscript{1}University of Toyama, 3190 Gofuku, Toyama, 930–8555 Japan
\textsuperscript{2}Japan Display Inc., 2–9–50 Chuo, Ebina, 243–0432 Japan

(Received 4 October 2017, Accepted for publication 18 April 2018)

Abstract: Even if three-dimensional multichannel sound becomes available in broadcasting, backward compatibility to conventional sound systems will be necessary. There are two transmission formats that can achieve this requirement. One is the simulcast format, and the other is the channel-scalable format. Although the channel-scalable format is advantageous over the simulcast format in terms of the required data rate, the unmasking artifact cannot be avoided when matrix operations are used to realize scalability. To solve this problem, this paper proposes a novel approach that models the quantization noise signal with a polynomial expansion of a decoded signal and removes it from the decoded signal. A subjective evaluation revealed that the proposed method can alleviate the unmasking artifact in the scalable coding of 8.1- and 22.2-channel audio signals.

Keywords: Three-dimensional audio, Audio coding, Transmission format, Quantization noise

PACS number: 43.60.Ek [doi:10.1250/ast.39.327]

1. INTRODUCTION

Over the past few decades, a considerable number of studies have been carried out on three-dimensional (3D) sound production and reproduction [1–3]. However, so far the study of the transmission of 3D sound signals has been superficial. From 2016, Nippon Hoso Kyokai (Japan Broadcasting Corporation, hereinafter referred to as “NHK”) started 8K ultra high-definition test broadcasting with 22.2-channel sound. When such 3D multichannel sound becomes available in regular broadcasting, the difficulty of setting up a 3D loudspeaker arrangement in a typical home environment should be considered. To solve this problem, we previously developed a method of converting the original sound signal into one for an alternative sound system with fewer loudspeakers [4]. This study showed that, as the result of subjective experiments on six-, eight- and ten-loudspeaker reproduction, the sound impressions of the 22.2 multichannel sound without Low Frequency Effect (LFE) channels can be reproduced by eight loudspeakers. Such an alternative sound reproduction should be considered in designing the transmission format of 3D sound.

\begin{itemize}
  \item\textsuperscript{*}e-mail: andio@eng.u-toyama.ac.jp
  \item\textsuperscript{†}e-mail: takafumi.fujita.pg@j-display.com
\end{itemize}

There are two transmission formats that can cope with this issue. One is the “simulcast format” and the other is the “scalable” format [5]. If 3D sound has \(N\) channels and the alternative reproduction requires \(M\) channels, the number of channels used for transmission becomes \(N + M\) in the simulcast format. In this format, the number of channels used for transmission increases if the compatibility with another sound system can be maintained. On the other hand, in the scalable format, \(N\)-channel 3D audio signals and other sound signals with the fewer channels can be transmitted by only \(N\)-channel signals by the matrix operation. The largest problem for scalable transmission is that the unmasking artifact occurs when a perceptual audio coder is used. Usually, a perceptual audio coder uses the masking properties of the auditory system to reduce the bit rate in such a way that the quantization noise generated by the bit rate reduction can be masked by the signal component in each sub-band [6]. However, if the decoded signal is processed by the matrix operation, the masked noise becomes perceptible [7].

MPEG surround, which was standardized in the Working Group 11 of ISO/IEC JTC1/SC29 (so-called Moving Picture Expert Group) as a multichannel audio transmission method, is known as a method to realize channel scalability while avoiding the unmasking artifact [8]. It realizes a backward compatibility to a stereo sound by transmitting a
two-channel downmixed signal with the additional spatial parameters of the sound. Since it was designed for low-bit-rate coding, the basic MPEG surround is not suitable for high-quality coding. Although MPEG surround is also applicable to high-quality coding by transmitting the predictive residual signal with the downmixed signal [9], it does not tolerate the desirable downmix for the home reproduction of 3D sound because it requires a specific downmix matrix.

We previously developed a method that transmits downmixed signals with additional signals that represented the spatial difference in the reproduced sound between the original and downmixed sound [10]. This method, however, could not avoid the unmasking artifact. This artifact was observed in the case that one channel signal had much lower energy than the other channels. The quantization noise caused by the other channel signals might leak into the low-energy channel via the inverse matrix $W^{-1}$. Such problems are described in Sect. 2 with the details of the method.

In this paper, we propose a novel method that models the quantization noise signal with a polynomial expansion of the decoded signal and removes it from the decoded signal. To simplify the problem, this paper treats the scalable transmission of 8 channels to provide enhanced spatial impressions. Although loudspeakers for such a sound system can be arranged optimally in a theater, they are difficult to set up in a typical home environment. “Downmixing” is a widely known method of reducing the number of channels in multichannel audio. Regarding the transmission of such 3D sound signals, it is desirable that the downmixed sound signals can be extracted easily from the transmitted sound signals, while the original multichannel signals must be restored. Figure 1 shows a block diagram of a transmission method satisfying such a requirement [10].

Let us denote a matrix converting an $N$-channel signal $s$ into an $M$-channel basic signal $y_m$ as $W_m$, and that converting $s$ into an $(N-M)$-channel additional signal $s$ as $W_s$, namely,

$$y_m(t) = W_ms(t),$$

$$y_s(t) = W_s(t).$$

An $N$-by-$N$ matrix $W$ is then defined as

$$W = \begin{pmatrix} W_m & W_s \end{pmatrix}.$$

### 2. DESIGN OF CONVERSION MATRIX

#### 2.1. Transmission of 3D Multichannel Sound

The State-of-the-art 3D multichannel sound has many channels to provide enhanced spatial impressions. Although loudspeakers for such a sound system can be arranged optimally in a theater, they are difficult to set up in a typical home environment. “Downmixing” is a widely known method of reducing the number of channels in multichannel audio. Regarding the transmission of such 3D sound signals, it is desirable that the downmixed sound signals can be extracted easily from the transmitted sound signals, while the original multichannel signals must be restored. Figure 1 shows a block diagram of a transmission method satisfying such a requirement [10].

Let us denote a matrix converting an $N$-channel signal $s$ into an $M$-channel basic signal $y_m$ as $W_m$, and that converting $s$ into an $(N-M)$-channel additional signal $y_s$ as $W_s$, namely,

#### 2.2. Design of Initial Conversion Matrix

In our scenario, the submatrix $W_m$ of matrix $W$ is determined by the loudspeaker arrangement of an 8-channel alternative sound system [4]. Thus, under the condition that $W_m$ is fixed, we propose a design method for $W_s$. In the first step, $W_s$ is given such that $y_s$ represents the...
The quantization noise will be amplified by the inverse matrix $W^{-1}$ because $W$ will approach a singular matrix in such a case.

To improve the matrix $W$ defined in the previous section, we previously proposed a method that reduces the power of the additional signal while keeping the condition number of $W$ small [10], on the basis of the fact that the condition number indicates the degree of singularity of the matrix. In this method, the objective function was defined as a product of the condition number of $W$ and the maximum energy of the additional signal:

$$F(W) = (\text{cond}(W))^{\gamma} \cdot \left( \max(E[y_{s,l}^2]) \right),$$

where $\text{cond}(W)$ is the condition number of $W$, $y_{s,l}$ is the $l$th-channel component of $y_s$, and $E[ \ ]$ is the expectation operator. The objective function was minimized by simulated annealing [12]. To control the balance between the two factors, we used the $\gamma$th power of the condition number of $W$ instead of the condition number itself in the objective function. As the result, when using a larger $\gamma$, we can obtain a matrix with the smaller condition number.

### 3. Modeling of Quantization Noise

In this section, we propose a new method that models the quantization noise signal to solve the unmasking artifact problem [5]. Figure 2 shows a block diagram of the proposed unmasking noise compensation method. The noise signal is defined as the difference between the original and decoded signals. Such a noise signal could be a nonlinear function of the decoded signal. Usually, a nonlinear function can be approximated by a polynomial with higher exponents of the variable. Therefore, the noise signal is modeled as a polynomial of the decoded signal in our study. The coefficients of the polynomial are estimated by minimizing the square error between the noise signal and its polynomial expansion.
3.1. Polynomial Expansion of Quantization Noise

In this paper, a certain time interval $\Gamma$ is defined by $[0, T]$. Let $\xi : \Gamma \rightarrow \mathbb{R}^N$ and $x : \Gamma \rightarrow \mathbb{R}^N$ be the quantization noise signal and the decoded signal, respectively. The quantization noise

$$\xi(t) = x(t) - s(t)$$

is modeled as a nonlinear function $\Phi$ as

$$\xi(t) = \Phi(x(t)),$$

where

$$\begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \vdots \\ \xi_N(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_N(t) \end{bmatrix}$$

and

$$\Phi(x(t)) = \begin{bmatrix} \Phi_1(x_1(t)) \\ \Phi_2(x_2(t)) \\ \vdots \\ \Phi_N(x_N(t)) \end{bmatrix}.$$ 

The $l$th element of the function $\Phi$, namely $\Phi_l$, can be approximated by the following polynomial expansion of $x_l$:

$$\Phi_l(x_l) \approx a_{l,0} + a_{l,1}x_l + a_{l,2}x_l^2 + \cdots + a_{l,P}x_l^P.$$ (8)

Thus, the quantization noise signal $\xi_l(t)$ for the $l$th channel can be estimated as

$$\hat{\xi}_l(t) = \sum_{m=0}^{P} a_{l,m}x_l^m(t).$$ (9)

In such a case, the coefficient $a_{l,m}$ can be solved by the least-squares method. The error function for the $l$th channel $E_l$ is defined as

$$E_l = \sum_{t=0}^{T} (\xi_l(t) - \hat{\xi}_l(t))^2 = \sum_{t=0}^{T} (\xi_l(t) - \sum_{m=0}^{P} a_{l,m}x_l^m(t))^2.$$ (10)

Differentiating this error function with respect to the coefficient $a_{l,k}$ and equating the result to zero, we obtain

$$\sum_{m=0}^{P} \left( \sum_{t=0}^{T} x_l^{m+k}(t) \right) a_{l,k} = \sum_{t=0}^{T} \xi_l(t)x_l^k(t), \quad k = 0, 1, \cdots, P.$$ (11)

The coefficients $a_{l,k}$, $k = 0, \cdots, P$ can be obtained by solving the matrix equation

$$X_l a_l = z_l,$$ (12)

where

$$a_l = (a_{l,0}, \ldots, a_{l,P})^T.$$

3.2. Recursive Estimation of Quantization Function

The error function $E_l$ ($l = 1, \ldots, N$) is minimized recursively for each $a_{l,m}$, $m = 0, \ldots, P$ in this section. In the first step, we solve the coefficient $a_{l,0}$ that minimizes the error function

$$E_{l,0} = \sum_{t=0}^{T} (\xi_l(t) - a_{l,0}^2).$$

The solution is

$$\hat{a}_{l,0} = \frac{1}{T+1} \sum_{t=0}^{T} \xi_l(t),$$

and the minimum error

$$\hat{E}_{l,0} = \sum_{t=0}^{T} (\xi_l(t) - \hat{a}_{l,0})^2$$

is obtained.

In the $(n+1)$th step, the coefficient $a_n$ is solved so as to minimize the error function

$$E_{l,n} = \sum_{t=0}^{T} \left( \xi_l(t) - \sum_{m=0}^{n-1} \hat{a}_{l,m}x_l^m(t) - a_nx_l^n(t) \right)^2.$$ (14)

As a result, the following solution and minimum error can be obtained:

$$\hat{a}_{l,n} = \frac{\sum_{t=0}^{T} (\xi_l(t) - \sum_{m=0}^{n-1} \hat{a}_{l,m}x_l^m(t)) x_l^n(t)}{\sum_{t=0}^{T} x_l^{2n}(t)}$$ (15)
In general, the estimation accuracy of the recursive method is lower than that of the method described in Sect. 3.1. However, the recursive estimation is stable even for higher \( P \) values.

### 3.3. Restoration of Original Signal

According to Eq. (6), the original \( N \)-channel signal \( s(t) \) can be estimated from the decoded signal \( x(t) \) and the estimated quantization noise signal \( \hat{\xi}(t) \) as

\[
\hat{s}(t) = x(t) - \hat{\xi}(t).
\]

Therefore, the \( l \)-th element of \( \hat{s}(t) \), namely \( \hat{s}_l(t) \), can be calculated using

\[
\hat{s}_l(t) = x_l(t) - \sum_{m=0}^{P} \hat{a}_{lm} x_l^m(t).
\]  

### 4. EVALUATION USING 22.2-CHANNEL SOUND

#### 4.1. Sound Material

As 3D multichannel sound materials, we used the 22.2-channel sounds shown in Table 1. The sounds Church, Organ and SL were natural 22.2-channel recordings. The other sounds were extracted from 8K ultra high-definition programs. The sampling frequency and bit depth were 48 kHz and 24 bits, respectively.

#### 4.2. Experimental Setup

As described in Sect. 1, the downmixing matrix \( W_m \) was selected so as to convert the 22-channel sound materials into 8-channel signals. Therefore, the numbers of channels \( N \) and \( M \) shown in Fig. 1 were 22 and 8, respectively. Figure 3 shows the loudspeaker arrangements of the 22- and 8-channel systems. In the calculation of the initial conversion matrix, we used 14 directional functions, whose directions of observation were set to \((0,0), (0,30), (0,60), (0,90), (0,120), (0,150), (0,180), (0,270), (45,45), (45,90), (45,135), (45,215), (45,315), \) and \((90,0) \) \[deg\] \[10\].

In Fig. 2, we used the AAC encoder and decoder developed by Fraunhofer Institute for Integrated Circuits. The quantization noise modeling and compensation were carried out in each frequency band of 24 sub-bands, each of which had the same bandwidth of 1 kHz. In the modeling, we set the time interval \( T \) as the total length of each sound material.

#### 4.3. Selection of Conversion Matrix and Bit Allocation

To investigate the relationship between the condition number of the conversion matrix and the bit allocation to

---

**Table 1** 22.2-channel sound materials used in the experiment.

<table>
<thead>
<tr>
<th>Material name</th>
<th>Length [s]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMFB</td>
<td>10</td>
<td>Sound in American football game</td>
</tr>
<tr>
<td>Birds</td>
<td>6</td>
<td>Sound of birds singing with light breeze</td>
</tr>
<tr>
<td>Church</td>
<td>6</td>
<td>Church bell with sound of footsteps on snow</td>
</tr>
<tr>
<td>Dam</td>
<td>12</td>
<td>Sound of water falling from large dam mixed with background music</td>
</tr>
<tr>
<td>Harmonica</td>
<td>6</td>
<td>Music including sound of harmonica</td>
</tr>
<tr>
<td>Jet</td>
<td>6</td>
<td>Sound of jet planes flying overhead</td>
</tr>
<tr>
<td>Orchestra</td>
<td>6</td>
<td>Orchestra sound mixed with birds singing</td>
</tr>
<tr>
<td>Organ</td>
<td>6</td>
<td>Pipe organ sound playing in cathedral</td>
</tr>
<tr>
<td>Rocket</td>
<td>14</td>
<td>Sound of rocket launched into outer space</td>
</tr>
<tr>
<td>SL</td>
<td>6</td>
<td>Sound of steam locomotive passing from left to right</td>
</tr>
<tr>
<td>Temple</td>
<td>6</td>
<td>Quiet environment with sound of water in specific direction</td>
</tr>
</tbody>
</table>

---

![Fig. 3](image-url) Loudspeaker arrangements for conversion of multichannel sound signal. \( \bullet \): loudspeaker, \( \circ \): listening position.
the basic and additional signals, six conversion matrices having different condition numbers were generated. The condition numbers were 8, 15, 20, 24, 32, and 40, which were obtained by controlling the parameter \( \gamma \) in Eq. (5). In this generation, the sound material “Birds” was used. The converted signals were coded by the AAC encoder with three types of bit allocation. Namely, the bitrates for the basic and additional signals were 800 kbps and 400 kbps, 720 kbps and 480 kbps, and 640 kbps and 560 kbps, respectively, as shown in Table 2. Since the AAC system at 320 kbps per five channels is known to fulfill the ITU-R requirements for indistinguishable quality [13], 640 kbps should be enough bit rate for the basic signal. In this experiment, the total bitrate of the 22.2-channel signal was defined as 1.2 Mbps, which is known as a bitrate satisfying broadcast quality [14].

For the evaluation of matrices and bitrates, we defined the SN ratio as

\[
\text{SNR} = 10 \log_{10} \frac{E[(x(t), x(t))]}{E[(\xi(t), \xi(t))]},
\]

where \(( , )\) denotes the inner product of vectors. Figure 4 shows typical results using the materials “Church,” “Orchestra,” and “Organ.” As shown in Fig. 4, the SN ratio increases with increasing condition number of the conversion matrix. Moreover, a gap of the SN ratio between condition numbers of 15 and 20 was found. Since a lower condition number is desirable in general, we selected the matrix whose condition number was 20 for use in subsequent experiments.

In terms of the bitrate, the results were almost the same in nine sound materials except for “Organ” and “Orchestra.” Figure 4(a) shows one such example. In these results, although a type II obtained the best result, the difference between types II and III was not significant. The type II also showed the best result for “Organ” but the worst result for “Orchestra.” According to the results, the type III seems to be the most appropriate bit allocation for the basic and additional signals.

### 4.4. Subjective Evaluation

We conducted a subjective evaluation experiment to examine the effectiveness of the proposed method of compensating the quantization noise. Seven sound materials, namely, “Birds,” “Church,” “Harmonica,” “Jet,” “Orchestra,” “Organ” and “Temple,” were selected from Table 1, considering the burden on the subject, and used in the evaluation. The maximum order \( P \) in Eq. (8) was set to three in consideration of our previous result, in which a polynomial approximation up to the third order was found to fully represent the quantization noise [5].

Since LFE channels require only a small bit rate, we ignored such channels in the description of our method. However, we used LFE signals in the subjective evaluation because the LFE channel effects the impression of the
reproduced sound. After generating the sum and difference signals from two LFE signals, we encoded these signals with the additional signals. Note that the decoded sum signal can be used as the LFE signal of an 8.1-channel signal.

4.4.1. Subjective evaluation method

In the subjective evaluation, the double-blind triple-stimulus with hidden reference method standardized in ITU-R BS.1116 [15] was used. Figure 5 shows the experimental procedure, where stimulus R denotes the reference sound, namely, the PCM sound, and stimuli A and B denote the sounds used for evaluation. Either stimulus A or B is the same as R, which is referred to as the hidden reference. The other stimulus is referred to as the object. The object stimulus was referred to as the object. The object stimuli were signals coded using the AAC at various bit rates as shown in Table 3, where the total bit rate did not include the additional bits for the polynomial expansion coefficients. After listening to a set of the three sounds twice, the subject assessed the impairment of A and B compared with R using the continuous five-grade impairment scale shown in Table 4. The impairment was assessed on the basis of any differences between the reference and evaluation sounds, including those of timbre, localization, and envelopment.

The subjects were 24 people in their twenties and thirties with experience of playing musical instruments.

4.4.2. Evaluation results

The average difference in the evaluated grade for all sound materials is shown in Fig. 6. Without noise compensation, no significant differences among bit rates of 1.2 Mbps (BH), 1.0 Mbps (BM), and 0.8 Mbps (BL) were found, suggesting that an unmasking noise caused a certain amount of deterioration of the sound regardless of the bit rate under the experimental conditions. The noise compensation method reduced the amount of deterioration, particularly when the bit rate was 1.2 Mbps. The difference between AH and BH was significant at a 5% significant level.

5. DISCUSSION

According to the experimental results, the compensation of the unmasking artifact by the combination of a perceptual audio coder and matrix operations is possible. Moreover, we confirmed the possibility of 1.2 Mbps scalable coding of 22.2 multichannel sound, although further experiments using critical materials are necessary.

According to our previous study, only odd orders of polynomial expansion coefficients are necessary [5]. If we send the first- and third-order coefficients, the number of additional bits required for the coefficients is 33 kbits (2 parameters × 32 bits × 22 channels × 24 bands) for 10 s of 22-channel signals. This amount of data is very small compared with that of sound signals.

In the experiments discussed in Sect. 4, so-called batch processing was used. In a real situation, we should use real-time processing, where the sound signal is divided into short time frames. In such a case, the estimation may
become easier because the sound signal can be regarded as stationary, while the polynomial expansion coefficients should be sent every frame. A study on such real-time processing will be carried out in the future.

In the experiment, we found a problem with evaluating 3D sound. Even if the sound was coded at a 1.2 Mbps bit rate and compensated (AH), some degradation was perceptible in the single-channel reproduction of the channel in which marked unmasking was observed before compensation. However, this artifact in the channel was imperceptible in all 22.2-channel reproduction. More discussion is necessary on how to resolve this problem.

6. CONCLUSION

We proposed a method of transmitting \( N \)-channel 3D sound signals as \( M \)-channel basic and \( (N - M) \)-channel additional signals in a scalable manner. We also proposed a method of compensating the unmasking artifact caused by matrix operations. A combination of these two methods may realize the scalable transmission of 22.2-channel sound with a 1.2 Mbps bit rate with a sound quality that is almost indistinguishable from the original, according to our subjective evaluation of the sound quality.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 26420348. Part of this study was conducted when one of the authors belonged to NHK. We thank NHK for its support of this study including the provision of 22.2-channel sound materials. We also thank Fraunhofer Institute for Integrated Circuits for providing us with the AAC encoder and decoder.

REFERENCES