SPATIAL CODING BASED ON THE EXTRACTION OF MOVING SOUND SOURCES IN WAVEFIELD SYNTHESIS

Toshiyuki Kimura[*1] *Kazuhiko Kakehi*[2] *Kazuya Takeda*[1] *Fumitada Itakura*[3]

[1]Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
[2]School of Computer and Cognitive Sciences, Chukyo University
101, Tokodachi, Kaizu-cho, Toyota-shi, Aichi, 470-0393, Japan
[3]Faculty of Science and Technology, Meijo University
1-501, Shiogamaguchi, Tenpaku-ku, Nagoya, 468-8502, Japan

ABSTRACT

Since a sound field reproduction system based on wavefield synthesis usually needs a great number of channel signals, the amount of data transmitted should be reduced. This paper therefore proposes a spatial coding method that is based on the extraction of moving sound sources and reduces the amount of data transmitted from an amount proportional to the number of channels to an amount proportional to the number of sound sources. A coding experiment was performed for a reverberant sound field which was simulated with image method. The effect of the proposed method on the perceptual quality was evaluated by the subjective assessment.

1. INTRODUCTION

Wavefield synthesis [1, 2] is a sound field reproduction technique synthesizing wave fronts at the boundary of a reproduction area by playing channel signals recorded at the boundary of an original area somewhere else. This technique, unlike binaural [3] and transaural [4] techniques, enables people in the reproduction area to experience the original area without any constraint on their movements.

Since the number of channel signals needed to reproduce a sound field is very large, the amount of data transmitted in teleconferencing and tele-ensemble applications needs to be reduced by coding the signals. The amount of data transmitted is proportional to the number of channel signals even when conventional coding methods such as AC-3 [5] and MPEG2 AAC [6] are used. What is needed is a spatial coding method reducing the number of channel signals transmitted.

We developed a spatial coding method based on the extraction of sound sources and showed that the effect of the proposed method on perceptual quality was acceptable when the amount of data transmitted was reduced from an amount proportional to the number of channel signals (24) to an amount proportional to the number of sound sources (5) [7]. That method assumes that sound sources are stationary, however, whereas the sources of reproduced sound fields are often moving. This paper therefore describes a new spatial coding method that can be used with moving sound sources. The algorithm of the method is explained in section 2, and a coding experiment for a reverberant sound field simulated with an image method is described in section 3. The results obtained when the effect of the proposed method on perceptual quality was evaluated by subjective assessment are reported in section 4.

*kimura@sp.m.is.nagoya-u.ac.jp

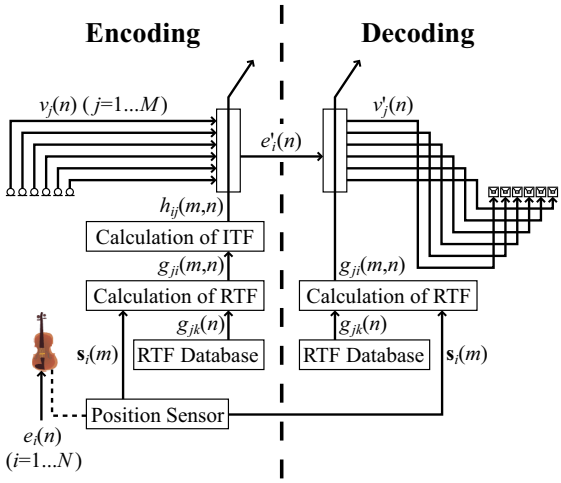

Fig. 1. Block diagram of the proposed method.

2. ALGORITHM

2.1. Encoding(Extraction of moving sound sources)

A block diagram of the proposed method is shown in Fig.1, where $e_i(n)$ and $v_j(n)$ are respectively the ith source signal of N moving sound sources and the jth channel signal recorded by M microphones. Let the sampling frequency of both signals be F_s[Hz], and let $\mathbf{s}_i(m)$ be the position information of sound sources that is recorded by the position sensor with the sampling frequency F_p[Hz]. m and n respectively denote the discrete times in periods $1/F_p$ and $1/F_s$. Thus, it is considered in the coding that the position of sound sources is changing every $P_{sw}(=F_s/F_p)$ samples.

At an encoding site, $g_{ji}(m,n)$, the room transfer function from ith source signal to jth channel signal in time m, is first calculated from the $\mathbf{s}_i(m)$ as follows:

$$g_{ji}(m,n) = g_{jk}(n), \tag{1}$$

where $g_{jk}(n)$ is the room transfer function from the sound source whose position is $\mathbf{s}_k[=\mathbf{s}_i(m)]$ to jth channel signal. It is assumed that $g_{jk}(n)$ is known at the encoding and decoding sites.

Then $h_{ij}(m,n)$, the inverse transfer function from jth channel signal to the ith source signal in time m, is calculated from $g_{ji}(m,n)$ as follows [4]:

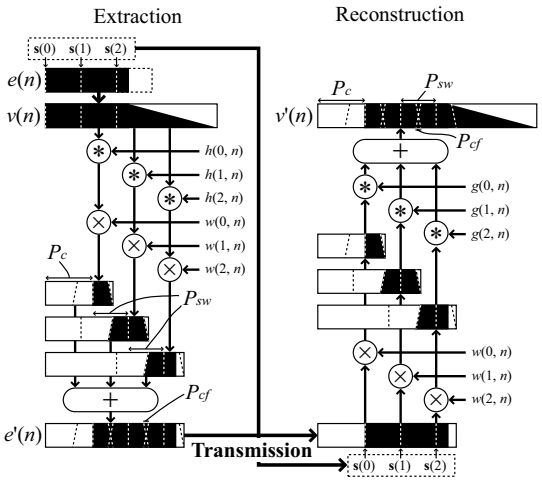

Fig. 2. Circuit diagram of extraction and reconstruction.

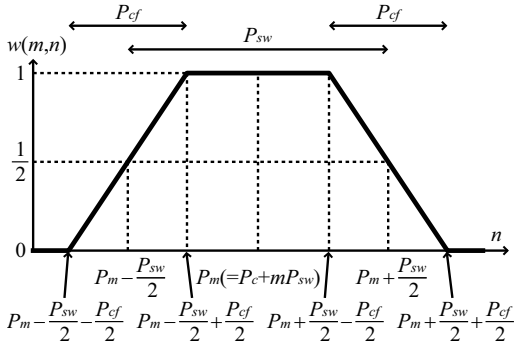

Fig. 3. Shape of window function.

$$\mathbf{H}(m,\omega) = \mathbf{G}^{+}(m,\omega)\mathbf{D}(\omega), \tag{2}$$

where $\mathbf{G}^{+}(m,\omega)$ is the Moore-Penrose pseudo inverse matrix of $\mathbf{G}(m,\omega)$. The matrices $\mathbf{H}(m,\omega)$, $\mathbf{G}(m,\omega)$ and $\mathbf{D}(\omega)$ are defined as follows:

$$\begin{aligned}
\mathbf{H}(m,\omega) &= \begin{pmatrix} H_{11}(m,\omega) & \dots & H_{N1}(m,\omega) \\ \vdots & \ddots & \vdots \\ H_{1M}(m,\omega) & \dots & H_{NM}(m,\omega) \end{pmatrix} \\
\mathbf{G}(m,\omega) &= \begin{pmatrix} G_{11}(m,\omega) & \dots & G_{M1}(m,\omega) \\ \vdots & \ddots & \vdots \\ G_{1N}(m,\omega) & \dots & G_{MN}(m,\omega) \end{pmatrix} \\
\mathbf{D}(\omega) &= \begin{pmatrix} e^{-j\omega T_c} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{-j\omega T_c} \end{pmatrix},
\end{aligned} \tag{3}$$

where $H_{ij}(m,\omega)$ and $G_{ji}(m,\omega)$ are the Fourier transforms of $h_{ij}(m,n)$ and $g_{ji}(m,n)$. T_c(=P_c/F_s) is the delay time needed to calculate inverse transfer functions satisfying causality.

Finally, the source signals $e_i'(n)$ are extracted by convolving $h_{ij}(m,n)$ to $v_j(n)$. The circuit diagram is shown in the left of Fig. 2. $w(m,n)$, which is shown in Fig. 3, is the window function used to smooth the waveform of the extracted source signal.

2.2. Decoding(Reconstruction of Channel Signals)

At a decoding site, $g_{ji}(m,n)$ is first calculated from the received position information $\mathbf{s}_i(m)$ according to Eq.(1). $v_j'(n)$ is reconstructed by convolving $g_{ji}(m,n)$ to the received source signal $e_i'(n)$ as shown in the right of Fig. 2. As a result of the proposed method, the amount of data transmitted is reduced from an amount proportional to the number of channel signals (M) to an amount proportional to the number of sound sources (N). The proposed method is very efficient when the number of sound sources is much less than that of channel signals ($N \ll M$).

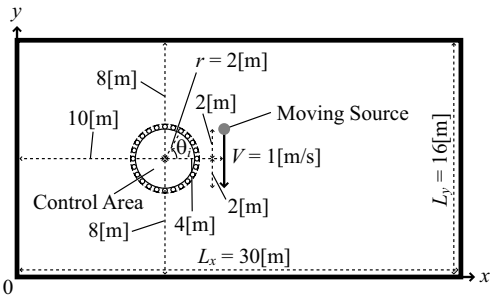

Fig. 4. Original sound field used in coding experiment.

3. CODING EXPERIMENT

3.1. Synthesis of Channel Signals

Channel signals for a reverberant sound field were simulated with an image method [8]. The positions of the one moving sound source and 24 microphones in the 2-dimensional original sound field used in the coding experiment are shown in Fig. 4. Channel signals were synthesized by the image method as follows:

$$\begin{aligned}
v_j(n) = \sum_{i=1}^{N} \sum_{p_x=0}^{1} \sum_{p_y=0}^{1} \sum_{q_x=-\infty}^{\infty} \sum_{q_y=-\infty}^{\infty} D[\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)] \\
\beta^{|q_x-p_x|+|q_x|+|q_y-p_y|+|q_y|} \frac{e_i[n+a-\frac{F_s}{c}|\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)|]}{4\pi|\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)|},
\end{aligned} \tag{4}$$

where c is the speed of sound (340 m/s) and

$$\begin{aligned}
\mathbf{d}_{ji}^{\mathbf{pq}}(n+a) &= \mathbf{s}_i^{\mathbf{pq}}(n+a) - \mathbf{r}_j \\
&= \begin{bmatrix} (1-2p_x)s_{ix}(n+a)+2q_xL_x \\ (1-2p_y)s_{iy}(n+a)+2q_yL_y \end{bmatrix} - \begin{bmatrix} r_{jx} \\ r_{jy} \end{bmatrix}.
\end{aligned} \tag{5}$$

$\mathbf{s}_i^{\mathbf{pq}}(n+a)$ and $\mathbf{r}_j$ are the position vectors of the ith mirrored sound source and the jth microphone. $D[\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)]$, the directivity function of the jth microphone, is defined as follows:

$$D[\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)] = \frac{1}{2}\left(1+\frac{\mathbf{u}_j \cdot \mathbf{d}_{ji}^{\mathbf{pq}}(n+a)}{|\mathbf{u}_j||\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)|}\right), \tag{6}$$

where $\mathbf{u}_j = (u_{jx}, u_{jy})^T$ is the directivity vector of the jth microphone. We can see from Eq. (4) that the channel signal is outputted in time $n+a+\frac{F_s}{c}|\mathbf{d}_{ji}^{\mathbf{pq}}(n+a)|$ if the source signal is inputted in time $n+a$. Thus the input signal $e_i(n+a)$ is used after interpolating as follows:

$$e_i(n+a) = (1-a)e_i(n) + ae_i(n+1). \tag{7}$$

a($0 \le a < 1$) is calculated in each n according to following quadratic equation:

$$\begin{aligned}
\{|\Delta\mathbf{d}_{ji}^{\mathbf{pq}}(n)|^2 - \frac{c^2}{F_s^2}\}a^2 + 2\{\mathbf{d}_{ji}^{\mathbf{pq}}(n)\cdot\Delta\mathbf{d}_{ji}^{\mathbf{pq}}(n) \\
+ \frac{c^2}{F_s^2}(K-n)\}a + \{|\mathbf{d}_{ji}^{\mathbf{pq}}(n)|^2 - \frac{c^2}{F_s^2}(K-n)^2\} = 0,
\end{aligned} \tag{8}$$

where $\Delta\mathbf{d}_{ji}^{\mathbf{pq}}(n) = \mathbf{d}_{ji}^{\mathbf{pq}}(n+1) - \mathbf{d}_{ji}^{\mathbf{pq}}(n)$, K is an arbitrary integer satisfying following condition:

$$\text{ceil}(n+\frac{F_s}{c}|\mathbf{d}_{ji}^{\mathbf{pq}}(n)|) \le K \le \text{floor}(n+1+\frac{F_s}{c}|\mathbf{d}_{ji}^{\mathbf{pq}}(n+1)|). \tag{9}$$

Synthesis conditions are listed in Table 1. These conditions assume that a person is walking as he speaks or plays the flute. Speech was recorded at an anechoic room and flute sounds were synthesized by a MIDI module.

Table 1. Synthetic conditions of channel signals

Dry source	Speech	Flute
F_s(Sampling frequency)	48[kHz]	
Duration of sound source	4[second]	
β(Reflection Coefficient)	0.5	0.7
R(Maximum Reflection Order)	6	10
Reverberation time[second]	0.6	1.0
V(Velocity of sound source)	1[m/s]=3.6[km/h]	

Table 2. Calculation conditions of inverse transfer functions

Reverberation time	0.6[second]	1.0[second]
FFT frame length[sample]	65536	131072
T_c(Coding delay time) =P_c(Coding delay samples)	20[ms] =960[sample]	
FIR filter length[sample]	28800	48000

3.2. Calculation of Room Transfer Functions

Room transfer functions from $\mathbf{s}_k(k$=1...481) to $\mathbf{r}_j(j$=1...24) were calculated as follows:

$$g_{jk}(n) = \sum_{p_x=0}^{1}\sum_{p_y=0}^{1}\sum_{q_x=-\infty}^{\infty}\sum_{q_y=-\infty}^{\infty} D(\mathbf{d}_{jk}^{\mathbf{pq}}) \beta^{|q_x-p_x|+|q_x|+|q_y-p_y|+|q_y|} \frac{\delta[n-\mathrm{round}(\frac{F_s}{c}|\mathbf{d}_{jk}^{\mathbf{pq}}|)]}{4\pi|\mathbf{d}_{jk}^{\mathbf{pq}}|}, \quad (10)$$

where $\delta(n)$ is Dirac's delta function. The reflection coefficients and the maximum reflection orders were the ones listed in Table 1. The $\mathbf{s}_k$, $\mathbf{u}_j$, and $\mathbf{r}_j$ used to calculate $D(\mathbf{d}_{jk}^{\mathbf{pq}})$ were set as follows:

$$\mathbf{s}_k = \begin{pmatrix} 14 \\ 10-\frac{k-1}{120} \end{pmatrix}, \mathbf{u}_j = \begin{pmatrix} \cos\frac{\pi(j-12)}{12} \\ \sin\frac{\pi(j-12)}{12} \end{pmatrix}, \mathbf{r}_j = 2\mathbf{u}_j + \begin{pmatrix} 10 \\ 8 \end{pmatrix}. \quad (11)$$

3.3. Convolution of Transfer Functions

First, $g_{ji}(m,n)$ was assigned from $g_{jk}(n)$ according to Eq. (1) where $m = \frac{F_p(k-1)}{120}$ and F_p was set to 30, 60, and 120 Hz. Then $h_{ij}(m,n)$ was calculated from $g_{ji}(m,n)$ according to Eq. (2). The calculation conditions are listed in Table 2. $e_i'(n)$ was extracted by convolving $h_{ij}(m,n)$ to $v_j(n)$ where F_p(=F_s/P_{sw}) was the same as described below and T_{cf} (=P_{cf}/F_s) was set to 1 and 4 ms. $v_j'(n)$ was reconstructed by convolving $g_{ji}(m,n)$ to $e_i'(n)$.

4. SUBJECTIVE ASSESSMENT

4.1. Experimental Environment

Subjective assessment was conducted in the low-reverberation room that had a reverberation time of about 80 ms. The arrangement of loudspeakers and the listening position of a subject are shown in Fig. 5. When 24 channel signals were fed into the loudspeaker

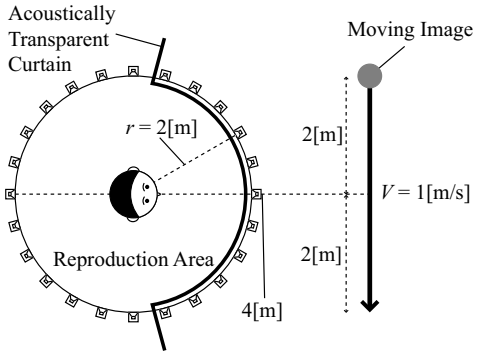

Fig. 5. Positions of the loudspeaker array and the subject used in the subjective assessment.

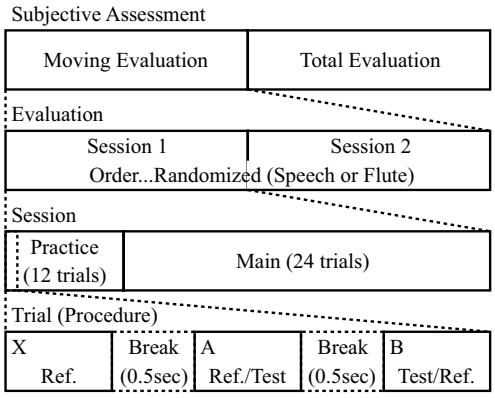

Fig. 6. Flowchart of the subjective assessment

Table 3. Series of the coding sound

	1	2	3	4	5	6
F_p	30Hz	30Hz	60Hz	60Hz	120Hz	120Hz
T_{cf}	1ms	4ms	1ms	4ms	1ms	4ms

array, the sound field was reproduced and the subject felt that the sound image is moving as shown in Fig. 5. The background noise level of the room was 25.0 dB(A). The presented sound level was set to about 70dB(A) at the position of the subject. To avoid the effect of visual perception, the light in the room was dimmed and the loudspeakers were covered by an acoustically transparent curtain.

4.2. Experimental Design

The subjects were 8 male students. A "Double-blind triple-stimulus with hidden reference" protocol was used according to ITU-R recommendations [9]. A flowchart of the subjective assessment is shown in Fig. 6. Two types of channel signals (Speech or Flute) were presented in the two sessions of each evaluation. The order of sessions was randomized in each subject. In the trial, three 4-second stimuli were presented in order. The first stimulus "X" was always the original sound, and either second stimulus "A" or third stimulus "B" was each one of the six types of coding sounds listed in Table 3. The order of trials was randomized in each subject. Twenty-four main trials [=6 (Types of coding sound)×2 (Either "A" or "B")×2 (Repetition)] were performed after 12 practice trials [=6 (Types of coding sound)×2 (Either "A" or "B")].

4.3. Experimental Procedure

The subjective assessment was divided into two parts: a "Moving evaluation" and a "Total evaluation." In the moving evaluation the subjects were asked to judge whether "A" or "B" is the same movement of sound as "X." Then the subjects were instructed to grade the perceptual impairment of the sound movement for the stimulus, which they noticed different from "X", with the one decimal scale of 1.0 to 4.9 under the condition that the grade of the other stimulus is 5.0 as reference (see Table 4).

In the total evaluation the subjects were asked to judge whether "A" or "B" is the same as "X." Then the subjects were instructed to grade the sound quality of the stimulus as in the same procedure in the first part. The subjects were allowed to rotate their heads

Table 4. Scale table of impairment

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

Table 5. Discrimination result of each subject in the subjective assessment.

Subject	Sample	Moving evaluation		Total evaluation	
		Speech	Flute	Speech	Flute
A	24	**15**	**23**	**15**	**22**
B	24	7	**23**	**16**	**24**
C	24	7	**24**	14	**24**
D	24	**14**	21	11	21
E	24	**14**	14	**15**	17
F	24	11	22	9	22
G	24	11	19	9	20
H	24	9	18	10	17

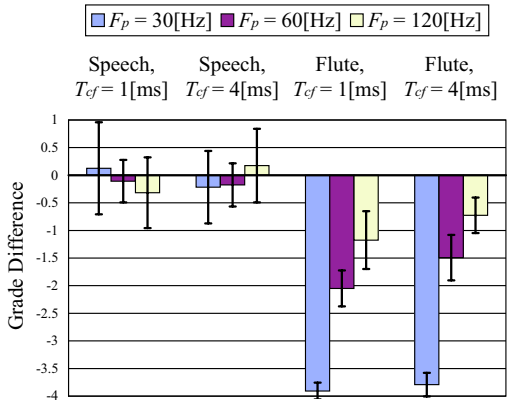

Fig. 7. Results of moving evaluation in the subjective assessment.

during the subjective assessment and to repeatedly listen to three stimuli during the trials.

4.4. Experimental Result & Discussion

The reliability of the subjects' responses was checked by the number of the correct responses, the grading of the stimulus assigned the original sound as 5.0. The data of subjects who answered by guess were excluded from the data set that was analyzed. The number of the correct responses for each subject is listed in Table 5. Further analysis was conducted for the data of the top three subjects of each session (shown in Table 5 by the bold font).

The average grade difference between the original and the coding sounds are shown for each sound source in Figs. 7 and 8, where also shown are the 95% confidence intervals for the averages calculated from 12 data values. If there were no perceptual distortion for the coding sound, the average grade difference should be close to 0. For the flute conditions the grade differences were lower than -2 when F_p was less than 60 Hz because of the moving distortion of the proposed method. Thus, if a tele-ensemble system for the music is actually made, F_p should be set to more than 120 Hz in order to preserve the coding quality. On the other hand, the grade differences were almost 0 under the speech conditions. It is thus considered that the perceptual quality of the proposed method is adequate for a teleconference system even when F_p is set to 30 Hz.

5. CONCLUSION

To reduce the number of channel signals to be transmitted, spatial coding method based on the extraction of moving sound sources was proposed in this paper. A coding experiment with a reverberant sound field synthesized by an image method was performed. It was confirmed from the subjective assessment that the perceptual quality obtained with the proposed method is acceptable when

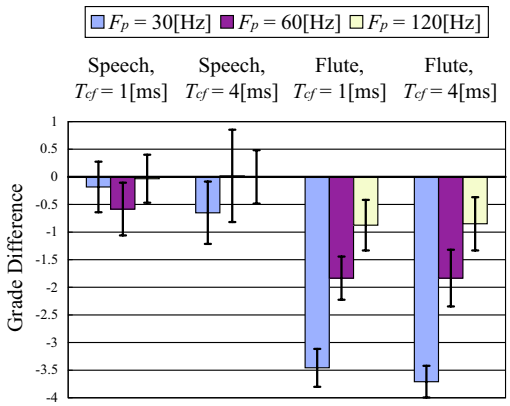

Fig. 8. Results of total evaluation in the subjective assessment.

appropriate parameters for moving sound sources are applied according to the type of sound sources.

It the future the proposed method needs to be evaluated by recording channel signals in a real environment. In this paper, it was assumed that the room transfer functions are known at the encoding and decoding sites. The room transfer functions therefore need to be estimated so that more practical telecommunication systems can be built.

6. REFERENCES

[1] M. Camras, "Approach to recreating a sound field," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1425–1431, Nov. 1968.

[2] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[3] S. Takane, Y. Suzuki, T. Miyajima, and T. Sone, "A new theory for high definition virtual acoustic display named ADVISE," *Acoust. Sci. and Tech.*, vol. 24, no. 5, pp. 276–283, Sep. 2003.

[4] J. Bauck and D. H. Cooper, "Generalized transaural stereo and applications," *J. Audio Eng. Soc.*, vol. 44, no. 9, pp. 683–705, Sep. 1996.

[5] S. Vernon, "Design and implementation of AC-3 coders," *IEEE Trans. CE*, vol. 41, no. 3, pp. 754–759, Aug. 1995.

[6] ISO/IEC 13818-7, *Information Technology - Genetic Coding of Moving Pictures and Associated Audio Information - Part 7 Advanced Audio Coding*.

[7] T. Kimura, K. Kakehi, K. Takeda, and F. Itakura, "Spatial compression of multi-channel audio signals using inverse filters," in *Proc. ICASSP*, 2002.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr 1979.

[9] ITU-R Recommendation BS.1116-1, *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*.