A Compression Method for Spherical Microphone Array Recordings using Principal Component Analysis

Hironori Sato[1], Arif Wicaksono[1], Shuichi Sakamoto[1], Cesar Salvador[1], Jorge Trevino[1],
Yôiti Suzuki[1], and Akinori Ito[2]

[1] Grad. Sch. Info. Sci., Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, Miyagi, Japan
Phone: +81-22-217-5535
E-mail: {hironori@ais, arifhw@ais, saka@ais,
salvador@ais, jorge@ais, yoh@}riec.tohoku.ac.jp

[2] Grad. Sch. of Engineering, Tohoku University
6-6, Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi, Japan
Phone: +81-22-795-5805
E-mail: aito@spcom.ecei.tohoku.ac.jp

Abstract

It is essential to capture the auditory spatial information in the development of ultra-realistic telecommunication systems. The microphone array with numerous microphones is a core device to record accurate three-dimensional (3D) sound. However, the network usage costs are a serious problem to transmit recorded signals by the super multichannel microphone arrays when accurate 3D sound information is reproduced at the distant place. In this study, we propose a compression method for spherical microphone array recordings using principal component analysis (PCA). While the PCA was applied in the time domain in the previously study, the technique is applied in the frequency domain by using the discrete cosine transform (MDCT). The Numerical experiments showed that the bitrate of the compressed data can be reduced to around one third of that of the original, uncompressed recordings without degrading the accuracy of spatial information.

1. Introduction

Capturing and reproduction of accurate 3D sound information is a core part of development of the ultra-realistic audio telecommunication systems. To capture such a sound space information, the microphone arrays are mainly used[1, 2, 3, 4]. To record sound space information with high spatial resolution, it is important to set numerous microphones densely. However, the transmission cost of such super multichannel signals over the network would be serious problem. Therefore, the efficient compression method for multi-channel microphone array recordings are highly required. The microphone array signals recorded by a large number of channels arranged densely are expected to have large inter-channel redundancies. Therefore, reducing such redundancies results to compress the recordings.

Previously, we proposed a novel compression algorithm to reduce the data size of recorded signals by spherical microphone array [5]. In this method, PCA was applied to the time domain signals. However, the algorithm was very time consuming because the PCA was applied sample by sample. Moreover, there was not high inter-channel redundancies between the samples of each microphones. Therefore, in this study, the PCA was applied in the frequency domain by using MDCT.

In this study, we propose a compression method by removing the inter-channel redundancies using PCA, and evaluate its performance. Proposed method was applied to the recordings by the 252-ch spherical microphone array that we constructed our previous study [4]. Afterwards, the impact of compression on the spatial accuracy of the output signal is analyzed.

2. Compression Method

As mentioned previously, it is effective to use spherical microphone arrays with numerous microphones desely mounted on the rigid sphere in order to recorde accurate sound space information. In such microphone arrays, the distance between each microphone is short enough to reduce the redundancy caused by the similarity of the signals. The PCA is a statistical technique to remove such redundancies.

In this study, the PCA was applied in the frequency domain bu using MDCT (modified discrete cosine transform). The MDCT is a time-frequency conversion tool using cosine bases and overlap-add method. This technique allows an efficient processing of long recordings [6]. The block diagram of our compression method is shown in Fig.1

First, the recorded signals of all microphones are windowed by a sine window with the length L. Assuming an array of N microphones, the windowed uncompressed signals in the time domain, represented by $s_n(\ell)$, are defined as follows:

$$\mathbf{S} = [s_{\ell n}]_{L\times N}, \;\; s_{\ell n} = s_n(\ell), \tag{1}$$

where $n = 1, 2, \ldots N$ is the microphone index, and $\ell = 1, 2, \ldots L$ indicates the time index. Next, the recorded signals are converted to the frequency domain by mean of the MDCT as follows:

$$\begin{aligned}
\tilde{\mathbf{S}} &= \mathbf{C}\,\mathbf{S},\ \ \tilde{\mathbf{S}} = [\tilde{s}_{kn}], \quad \tilde{s}_{kn} = \tilde{s}_n(k), \qquad (2)\\
\mathbf{C} &= [c_{k\ell}], \quad c_{k\ell} = \cos\left\{\frac{\pi}{2L}\left(2\ell - 1 + \frac{L}{2}\right)(2k-1)\right\},\\
k &= 1, 2, \ldots L/2,\\
\ell &= 1, 2, \ldots L.
\end{aligned}$$

Here, k indicates the frequency index.

To normalize the each colums of the matrix $\tilde{\mathbf{S}}$, the columns of matrix $\tilde{\mathbf{S}}$ are divided by their own standard deviations as follows:

$$\mathbf{M} = \tilde{\mathbf{S}}\ \text{diag}\left(\left[\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots \frac{1}{\sigma_N}\right]\right), \qquad (3)$$
$$\text{where} \quad \sigma_n = \text{std}(\tilde{\mathbf{S}}_n).$$

The correlation $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ [7] is then split into factors using the eigenvalue decomposition according to the following equation:

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}. \qquad (4)$$

Here, $\mathbf{V}$ is a square matrix containing the eigenvectors of the covariance matrix,$\mathbf{v_n}$, along its columns. Moreover, $\boldsymbol{\Sigma}$ is a diagonal matrix formed by the eigenvalues λ_n of the correlation matrix.

The proposed method seeks to decrease the amount of data to be transmitted by reducing the dimension of eigenvector matrix $\mathbf{V}$. The number of columns to be transmitted is determined by selecting a target cumulative variance (CV), defined as follows:

$$\text{CV} = 10\log_{10}\left(\frac{1}{\text{tr}(\boldsymbol{\Sigma})}\sum_{i=1}^{N'}\lambda_i\right). \qquad (5)$$

Here, the trace of Σ is the sum of the diagonal components of eigenvalue matrix $\boldsymbol{\Sigma}$. The CV value indicates the contribution that the eigenvalues up to the N'-th one have on their total sum as a logarithm. If the target CV is chosen to be high, a large number of eigenvectors are preserved and the total compression ratio will be low. However, choosing a small target CV results in the excessive loss of information and may reduce the accuracy of auditory spatial information. Figure 2 shows that the behavior how CV in relation to the number of preserved eingenvectors for various types of band-pass noise. The figure indicates the existence of redundancy in the recorded signals. In this step, the number of eigenvector is truncated N to N'.

The truncated matrix, $\mathbf{V}'$, can then be multiplied by the matrix of normalized signals $\mathbf{M}$ to obtain a set of independent components $\boldsymbol{\mu}'$. Explicitly, we perform the following operation:

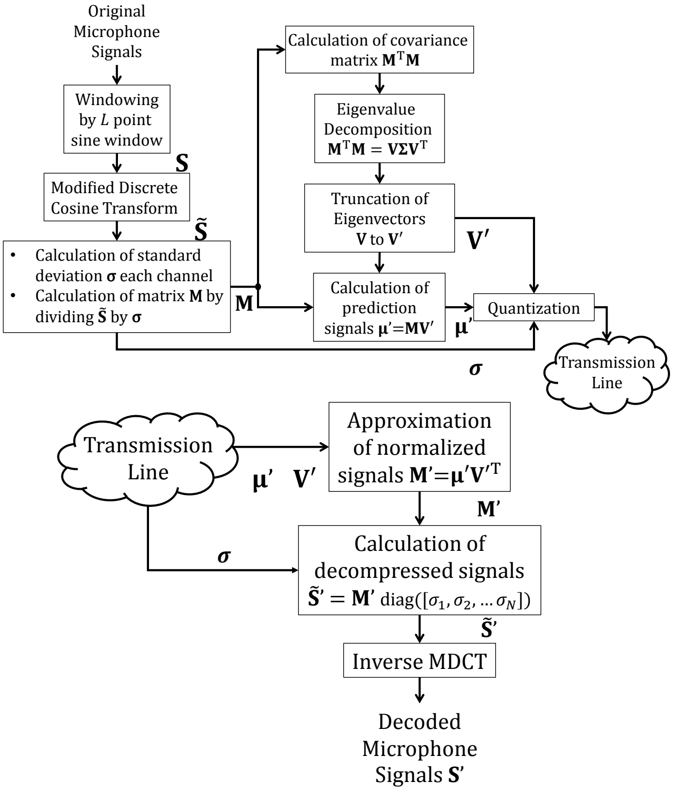

Original Microphone Signals
Windowing by L point sine window
$\mathbf{S}$
Modified Discrete Cosine Transform
$\tilde{\mathbf{S}}$
• Calculation of standard deviation $\boldsymbol{\sigma}$ each channel
• Calculation of matrix $\mathbf{M}$ by dividing $\tilde{\mathbf{S}}$ by $\boldsymbol{\sigma}$
$\mathbf{M}$
Calculation of covariance matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$
Eigenvalue Decomposition $\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$
Truncation of Eigenvectors $\mathbf{V}$ to $\mathbf{V}'$
$\mathbf{V}'$
Calculation of prediction signals $\boldsymbol{\mu}'$=$\mathbf{M}\mathbf{V}'$
$\boldsymbol{\mu}'$
Quantization
$\boldsymbol{\sigma}$
Transmission Line
Transmission Line
$\boldsymbol{\mu}'$ $\mathbf{V}'$
Approximation of normalized signals $\mathbf{M}'$=$\boldsymbol{\mu}'\mathbf{V}'^{\mathrm{T}}$
$\mathbf{M}'$
$\boldsymbol{\sigma}$
Calculation of decompressed signals $\tilde{\mathbf{S}}' = \mathbf{M}'$ diag$([\sigma_1, \sigma_2, \ldots \sigma_N])$
$\tilde{\mathbf{S}}'$
Inverse MDCT
Decoded Microphone Signals $\mathbf{S}'$

Figure 1: The block diagram of the method

$$\boldsymbol{\mu}' = \mathbf{M}\mathbf{V}'. \qquad (6)$$

By the proposed method, recorded signals are expressed by using only the independent components $\boldsymbol{\mu}'$, the eigenvector matrix $\mathbf{V}'$, and the standard deviation values of all channels, $\boldsymbol{\sigma} = [\sigma_1,\ \sigma_2,\ \ldots\sigma_N]$. These $\boldsymbol{\mu}'$, $\mathbf{V}'$, and $\boldsymbol{\sigma}$ indicates the compressed signals to be transmitted.

On the receiver side, the normalized signals are approximated according to the following equation:

$$\mathbf{M}' = \boldsymbol{\mu}'\mathbf{V}'^{\mathrm{T}}. \qquad (7)$$

Finally, multiplication of the columns of $\mathbf{M}'$ and their corresponding standard deviation values $\boldsymbol{\sigma}$ gives the decompressed signals $\tilde{\mathbf{S}}'$ in the frequency domain in such a way that:

$$\tilde{\mathbf{S}}' = \mathbf{M}'\ \text{diag}\ ([\sigma_1, \sigma_2, \ldots \sigma_N]). \qquad (8)$$

The reconstructed signals in the time domain $\mathbf{S}'$ are obtained via the inverse MDCT.

3. Evaluation

3.1 Numerical experiment condition

In the evaluation of the proposed method, the point source was set on the horizontal plane at the distance of 1.5 m.

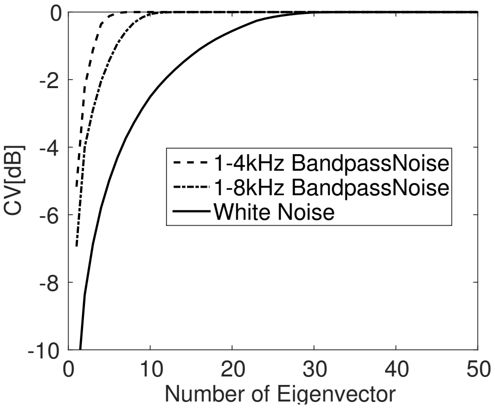


Figure 2: The behavior of CV value when some bandpass noises are input.

The angular position was changed at the interval of 1 degree. Here, 1-s white noise was presented from the point source and recorded by a 252-ch microphone array [4] with sampling frequency of 48 kHz, and quantization of 16-bits. Since the directional resolution was 1 degree, sound signal was recorded 360 times.

The proposed method was applied to the recordings to compute the compressed signals $\boldsymbol{\mu}'$, $\mathbf{V}'$, and $\boldsymbol{\sigma}$, all of which have the same resolution of 16-bits. The number of eigenvectors to be transmitted was set to $N' = 40$. As a comparison, two conventional compression methods, and simple reduction of the number of microphone were used. The conventional ones are the common audio compression algorithm, MPEG-1 audio layer 3 (mp3) at the bitrate of 320-kbps , and a lossless compression method, roshal archive (RAR). The simple reduction of the number of microphone array was composed by picking up 72 microphones from 252-ch spherical microphone array in order to keep the same bitrate of compressed signals. The 72 microphones were selected as evenly distributed as possible. In summary, we compared the uncompressed original recordings, our present proposal, mp3, RAR, and simple reduction of the number of microphone (RNM) recordings in order to evaluate the compression ratio and the accuracy of sound space information.

3.2 The accuracy of spatial information

In this study, the decompressed recordings $\mathbf{S}'$ and the original recordings $\mathbf{S}$ were transformed to a set of binaural signals which have spatial information by the SENZI algorithm [4]. The impact of the proposed method on the spatial accuracy of the binaural signals was evaluated in terms of a metric known as the spectral distortion (SD). The left ear signal in the frequency domain is defined as $P_L(\theta, k)$, where θ is the azimuth of sound source direction, k is the frequency index. Then the SD value between two synthesized signals is defined as follows:

$$\mathrm{SD}(\theta, k) = 20 \log_{10} \left| \frac{P'_L(\theta, k)}{P_L(\theta, k)} \right|, \tag{9}$$

where the $P_L(\theta, k)$ indicates the left ear signal calculated from uncompressed microphone recordings, and $P'_L(\theta, k)$ denotes that calculated from the compressed recordings.

3.3 Results

This SD value was computed for all azimuth on the horizontal plane (360 directions). The SD values of the resynthesized 3D sound space in each condition are shown in Fig. 3. Because RAR is a lossless compression method, the SD value was not analyzed. It is clearly seen that SD value of our proposal is quite small and outperfomes the order two compression methods. This figures indicate that 3D sound space information is accurately reproduced by using our proposal.

The performance of signal compression was evaluated by comparing the bitrate. The transmission rates of all condition are shown in Fig. 4. This figure indicates that our proposal can reduce the data size of recordings around 1/3.

4. Conclusions

In this study, we proposed a compression method based on the inter-channel redundancy removal for super multichannel spherical microphone array recordings. The PCA was applied to the correlation matrix of MDCT coefficients. The numerical experiment demonstrated that our proposal can achieve a good compression performance while preserving the accuracy of the re-synthesized sound space for most audible range.

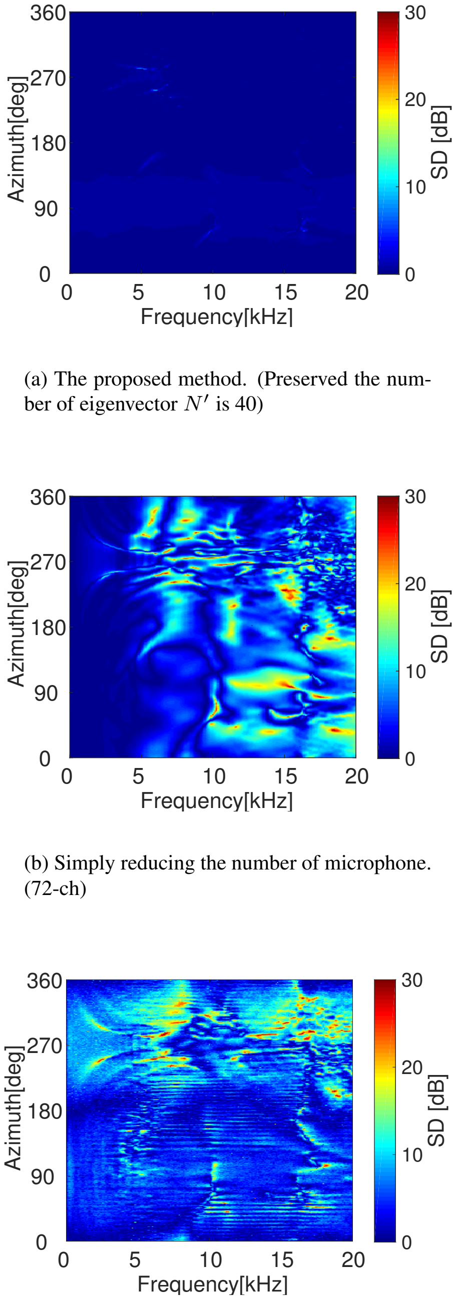
As future works, we will introduce the efficient quantization scheme based on the charactaristics of the human auditory system to achieve higher compression rate. For example, the psychoacoustic model which is used in mp3 and advanced audio coding (AAC) can be applied to quantize the parameters to be transmitted in our proposal. The psychoacoustic model gives a quantization strategy to hide the quantization noise into inaudible area hence the bitrate can be reduced without degrading perceptual quality.

Acknowledgment

Parts of this research were supported by a Grant-in-Aid for Scientific Research (A) (No.16H01736).

References

[1] A.J. Berkhout, ”A holographic approach to acoustic control,” *J. Audio Eng. Soc.*, **36**, pp. 977-995, 1988.



(a) The proposed method. (Preserved the number of eigenvector N' is 40)

(b) Simply reducing the number of microphone. (72-ch)

(c) MPEG-1 layer III (mp3 320-kbps).

Figure 3: The SD values of re-synthesized sound space on the horizontal plane in each condition.

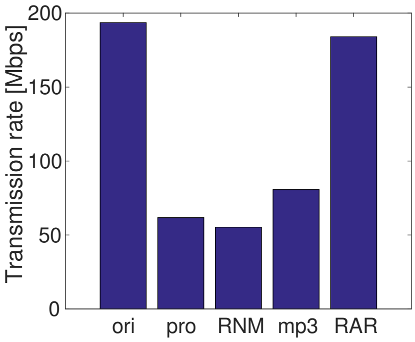


Figure 4: Transmission rate of each condition. The labels ”ori”, ”pro”, ”RNM”, ”mp3”, ”RAR” indicate original recordings (uncompressed), compressed by our proposal, simply reducing the number of microphone, mp3 compression with 320 kbps, roshal archive, respectively.

[2] M.A. Poletti, ”Three-dimensional surround systems based on spherical harmonics,” *J. Audio Eng. Soc.*, **53**, pp. 1004-1025, 2005.

[3] S. Ise, ”A Principle of Sound Field Control based on the Kirchhoff-Helmholtz Integral Equation and the Theory of Inverse Systems”, *Acustica*, Vol. 85, pp. 78–87, 1999.

[4] S. Sakamoto, *et al.*, ”Sound-space recording and binaural presentation system based on a 252-channel microphone array” *Acoust. Sci. &. Tech.*, **36** (6), pp. 516-526, 2015.

[5] H. Sato, *et al.*, ”Compression of Spherical Microphone Array Recordings using Eigenvalue Decomposition”, *RISP International Workshop on NCSP*, 2016

[6] A. Spanias, *et al.*, *Audio Signal Processing and Coding* , Wiley-Interscience, 2007.

[7] M. Loeve, *Probability Theory II (Graduate Texts in Mathematics)*, Springer-Verlag 1978.