Compression of Spherical Microphone Array Recordings using Eigenvelue Decomposition

Hironori Sato, Arif Wicaksono, Shuichi Sakamoto, Cesar Salvador, Jorge Trevino, Yôiti Suzuki

Graduate School of Information Sciences and Research Institute of Electrical Communication, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan
Phone:+81-22-217-5535
E-mail: {hironori@ais. arifhw@ais. saka@ais. salvador@ais. jorge@ais. yoh@}riec.tohoku.ac.jp

Abstract

A technique to capture high-definition 3D sound space is important for realizing highly realistic communication systems. We proposed a 3D sound system called SENZI. This system consists of a microphone array installed on a solid compact sphere, a digital signal processing unit, and a binaural auditory display. To transmit 3D sound to a distant place, the recorded signals must be transmitted via network. Therefore, since the number of channels of SENZI is very large, data size reduction is an important problem. In the present study, we propose to compress the signals recorded by the spherical microphone array by using the eigenvalue decomposition. Results of computer simulation showed that a compression ratio of 4% is obtained while preserving the accuracy of the re-synthesized 3D sound for most of the audible frequency range.

1. Introduction

Recording and reproduction of accurate 3D sound information is important for realizing highly realistic audio communications. As for recording, the real-time SENZI system shows promising results [1] [2]. This system relies on a 252-channel spherical microphone array as its main component to capture 3D sound information. The weighted sum of the recordings is used to synthesize a set of personalized binaural signals that match the end user's head orientation in real-time. For this reason, the 252 recorded signals must be transmitted to the listening location. However, transmission over a network can be difficult or entail considerable costs given the large amount of data resulting from the high channel count. Moreover, this problem is expected to become worse as the number of microphones is increased from the present 252-channel implementation in order to improve spatial resolution.

In the existing compression method that is represented by MPEG-1 Audio Layer-3 (mp3) , it compresses data based on characteristics of human's auditory sense (e.g. auditory masking) [3]. In these methods, spatial information will be lost because they do not consider spatial characteristics of the sound field. It is, therefore, important to develop a way to reduce the amount of data while preserving the spatial information that must be transmitted in implementations of the SENZI system.

This study considers a compression method based on the eigenvalue decomposition and applies it to the sound signals recorded by a 252-channel spherical microphone array. The algorithm is applied with the SENZI system, and its impact on the spatial accuracy of the resulting binaural signals is analyzed.

2. Compression Algorithm

The signals recorded by a spherical microphone array are expected to have high similarities because the microphones are located closely-spaced. This means that large redundancies exist across all channel signals. Removing these redundancies can significantly reduce the amount of data required to transmit the signals. To this end, the present research considers the Karhunen-Loéve transform (KLT) of the microphone recordings.

The block diagram of proposed algorithm is shown in Figure 1. We assume a set of N microphones. The signal recorded by the n-th microphone over a temporal frame of L samples is denoted by $s_n(\ell)$, with $\ell = 0, \dots L-1$. The first step of the proposed algorithm is to divide each signal by its standard deviation in order to normalize it according to the following equation:

$$m_n(\ell) = \frac{s_n(\ell)}{\mathrm{std}(s_n)}. \tag{1}$$

The normalized signals $m_n(\ell)$ are used to construct the following matrix:

$$\mathbf{M} = \begin{bmatrix} m_1(0) & m_2(0) & \dots & m_N(0) \\ m_1(1) & m_2(1) & \dots & m_N(1) \\ \vdots & \vdots & \ddots & \vdots \\ m_1(L-1) & m_2(L-1) & \dots & m_N(L-1) \end{bmatrix}_{L \times N}. \tag{2}$$

The covariance between all pairs of signals can then be calculated by the product $\mathbf{M}^T\mathbf{M}$; this is known as the covariance

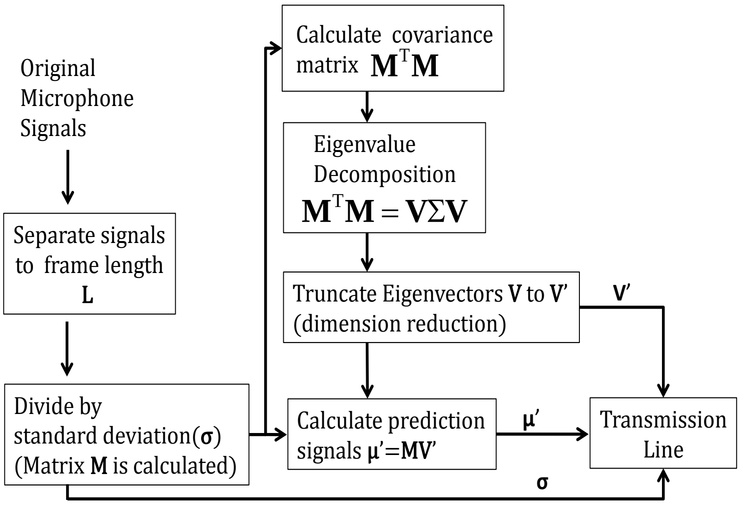

Original
Microphone
Signals
Separate signals to frame length L
Divide by standard deviation(σ) (Matrix M is calculated)
Calculate covariance matrix M^T M
Eigenvalue Decomposition M^T M = VΣV
Truncate Eigenvectors V to V′ (dimension reduction)
V′
Calculate prediction signals μ′=MV′
μ′
Transmission Line
σ

Figure 1: Block Diagram of Compression Algorithm

matrix [4]. This matrix is then split into factors using the eigenvalue decomposition (EVD) according to the following equation:

$$\mathbf{M}^{\boldsymbol{T}}\mathbf{M} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\boldsymbol{T}}. \tag{3}$$

Here, $\mathbf{V}$ is a square matrix containing the eigenvectors of the covariance matrix, $\mathbf{v}_n$, as its columns. $\boldsymbol{\Sigma}$ is a diagonal matrix formed by the eigenvalues λ_n of the covariance matrix.

The proposed method seeks to decrease the amount of data to be transmitted by reducing the dimension of eigenvectors matrix $\mathbf{V}$. The truncated matrix, $\mathbf{V}'$, can then be multiplied by the matrix of normalized signals $\mathbf{M}$ to obtain a set of independent components $\boldsymbol{\mu}'$ according to the following equation:

$$\boldsymbol{\mu}' = \mathbf{M}\mathbf{V}'. \tag{4}$$

Our proposal requires, then, to transmit only the independent components $\boldsymbol{\mu}'$, the eigenvectors matrix $\mathbf{V}'$, and the standard deviation values for each channel, $\boldsymbol{\sigma}$. The receiver can then approximate the normalized signals according to the following equation:

$$\mathbf{M}' = \boldsymbol{\mu}'\mathbf{V}'^{\boldsymbol{T}}. \tag{5}$$

Finally, original signals can be recovered by multiplying the columns of M' by their corresponding standard deviation values $\boldsymbol{\sigma}$.

3. Evaluation of the accuracy

3.1 Compression ratio

Compression ratio is evaluated by comparing the amount of data in compressed and uncompressed signals. The target sound sources consist of 1/3 octave band noise with a duration of 1 second at a sampling frequency of 48 kHz. The sound sources lie on the horizontal plane at steps of 1 degree. A total of 31 different bands were used to analyze the effect of compression for each frequency band, with center frequencies between 20 Hz and 20 kHz. Spherical microphone array recordings for these test sounds were generated through a computer simulation and used as input to the proposed compression algorithm. The cutoff level of eigenvector was used as a parameter to investigate the relationship between compression ratio and the accuracy of synthesized 3D sound.

We define the cumulative variance (CV) as follows:

$$\mathrm{CV}[\%] = \frac{\displaystyle\sum_{i=1}^{\ell} E_i}{\displaystyle\sum_{k=1}^{252} E_k} \times 100. \tag{6}$$

The variable E_i denotes eigenvalue that an index number is i. The denominator of this equation means a grand total of the eigenvalues, and the numerator indicates partial sum of the eigenvalues to index number ℓ. The matrix of eigenvectors $\mathbf{V}$ is truncated in such a way that the CV reaches the thresholds of 95%, 90%, 85%, 80%, 75%.

Figure 2 shows CV in the case of using various bandpass noises. This figure indicates that 3D sound can be reproduced by a fraction of eigenvectors. This means that there is a redundancy among the recording signals. When the frequency range of the sound source is increased, the total number of eigenvectors that are required to synthesize accurate 3D sound also increase.

On the above condition, compression rate (CR) is calculated using the following equation:

$$\mathrm{CR}[\%] = \frac{\mathrm{I}_{\mathrm{org}}}{\mathrm{I}_{\mathrm{comp}}} \times 100, \tag{7}$$

where $\mathrm{I}_{\mathrm{org}}$ denotes the amount of data of original signals. This is the total data which is recorded by 252-ch microphones at a sampling frequency of 48 kHz and quantization bit of 16 bit, for each 360 directions and each 31 octave bands. $\mathrm{I}_{\mathrm{comp}}$ denotes data amount of quantized $\mathbf{V}'$, $\boldsymbol{\mu}'$ and $\boldsymbol{\sigma}$ obtained by applying the proposed method to the original signals each sound source position and each frequency band. $\mathbf{V}'$ and $\boldsymbol{\mu}'$ are quantized in 16 bit, $\boldsymbol{\sigma}$ is quantized in 8 bit. Figure 3 shows the compression rate as a function of threshold to truncate the transmitted eigenvectors. The compression rate deteriorates with increasing of transmitted eigenvectors.

3.2 Spectral Difference

The impact of the compression method on the accuracy of spatial reproduction is analyzed by looking at the spectral difference (SD) . We calculate the average bandpass energy of binaural signals (sound pressure at the entrance to the listener's ear canal) which is synthesized by recorded signals with/without applying compression method. SD is defined as the absolute difference of average bandpass energy between

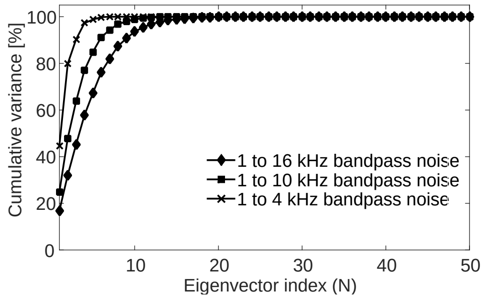


Figure 2: Cumulative variance (CV)

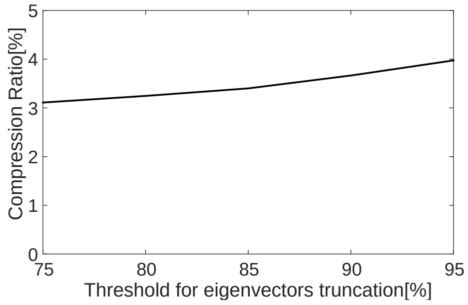


Figure 3: Compression Ratio in several thresholds

compressed and uncompressed signals using following equation:

$$P_{\theta,k} = \frac{\sum_{i=1}^{\ell} |A_{\theta,f_i}|}{\ell}, \tag{8}$$

$$SD_{\theta,k} = \left| 20 \log_{10} \left| \frac{P_{\text{comp } \theta,k}}{P_{\text{uncomp } \theta,k}} \right| \right|. \tag{9}$$

The variable $P_{\theta,k}$ denotes average bandpass energy of k-th center frequency of 1/3 octave noise at azimuth θ. The variables f_1 and f_l denote lower and upper limit frequency of k-th center frequency. The variable A_{θ,f_i} denotes spectral value of i-th frequency bin at azimuth of θ. The variable $P_{\text{comp } \theta,k}$ is average bandpass energy calculated from compressed signals and $P_{\text{uncomp } \theta,k}$ corresponds to uncompressed signals.

Figures 4 and 5 show the average bandpass energy of compressed and uncompressed sound signals for sound sources located on the horizontal plane, at different azimuth angles and frequency bands. Figure 6 shows the SD which is calculated by equation (9). Figure 6 indicates that the effect of the proposed method is very small on the horizontal plane. Figures 7 and 8 show SD in different thresholds for eigenvector truncation. In these figures, slightly large SD is observed at a frequency of around 315 Hz when threshold of truncation is decreased. This error is caused by the process of windowing. In the simulation, Hanning window at the length of 256 points is used to calculate covariance matrix with half-length

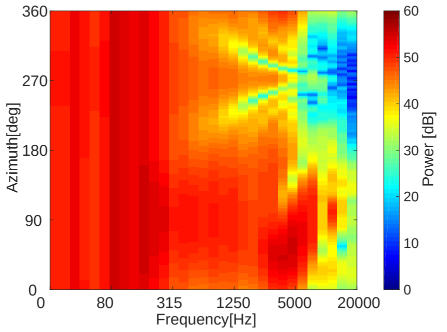


Figure 4: Average bandpass energy of synthesized 3D sound from uncompressed signals

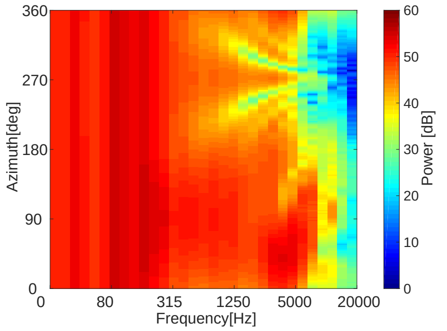


Figure 5: Average bandpass energy of synthesized 3D sound from compressed signals (Threshold of eigenvector truncation: 95%)

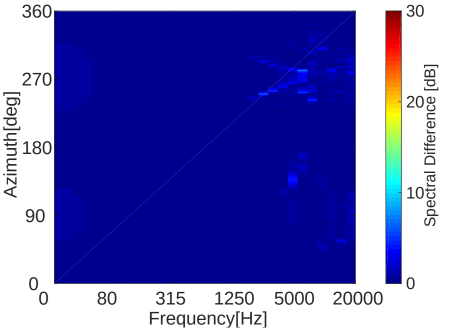
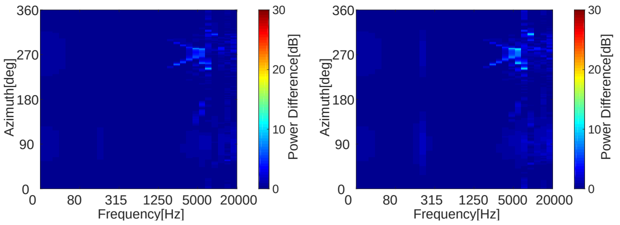


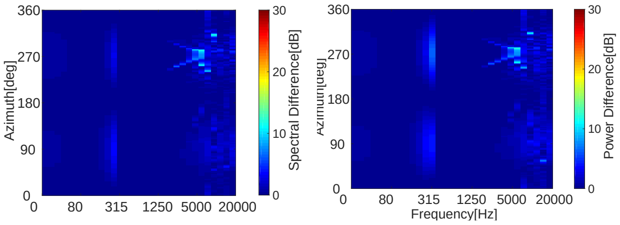
Figure 6: The effect of the compression on the synthesized 3D sound (Threshold for eigenvector truncation: 95%)



(a) Threshold for eigenvector truncation: 90%

(b) Threshold for eigenvector truncation: 85%

Figure 7: The effect of the compression on the synthesized 3D sound



(a) Threshold for eigenvector truncation: 80%

(b) Threshold for eigenvector truncation: 75%

Figure 8: The effect of the compression on the synthesized 3D sound

overlap. This 128 points are equivalent to 375 Hz. This discontinuity induces such error.

These results indicate that the proposed method can achieve a considerable compression ratio while preserving the accuracy of the synthesized 3D sound for most of the audible frequency range.

4. Conclusion

In present study, we proposed a compression method for spherical microphone array signals by using eigenvalue decomposition. In the proposed method, eigenvectors of covariance matrix are truncated to compress the total amount of recorded signals. The results of computer simulation indicate that the proposed method can achieve a considerable compression rate while preserving the accuracy of the synthesized 3D sound for most of the audible frequency range. In the future study, It is important to investigate whether this advantage is maintained when the proposed method is applied to the actual signals recorded by spherical microphone array.

Acknowledgment

We thank to Prof.Akinori Ito for his useful comments to simulate the environment. A part of this research was supported by a Grant-in-Aid for Scientific Research (A) (No. 24240016) from the Ministry of Education, Culture, Sports, Science and Technology, and the A3 Foresight Program for “Ultra-realistic acoustic interactive communication on next-generation Internet.”

References

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

[2] S. Sakamoto, *et al.*, ”3D Sound-Space Sensing Method Based on Numerous Symmetrically Arranged Microphones”, *IEICE trans.fundamentals*, vol.**E97-A** NO.9, 2014, pp.1893-1901

[3] J. Herre, ”From joint stereo to spatial audio coding recent progress and standardization”, *In 7th Int. Conf. on Digital Audio Effects (DAFx’04)*, 2004, pp.157-162.

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