Experimental investigation of robustness of spatial cepstrum features under various recording conditions

Taiga Kawamura* and Ryoichi Miyazaki* and Keisuke Imoto† and Nobutaka Ono‡
* National Institute of Technology, Tokuyama College, Yamaguchi, Japan
E-mail: i15kawamura@tokuyama.kosen-ac.jp
† Doshisha University, Kyoto, Japan
‡ Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—In recent years, devices that can easily record sounds, such as smartphones and tablets, have become widespread. Also, speech enhancement and acoustic scene analysis have been researched using a distributed microphone array consisting of these devices. The use of the spatial cepstrum has been proposed as a method for acquiring spatial information from a distributed microphone array. However, the behavior of the spatial cepstrum has been shown in the minimal experiment where the sound source is only white Gaussian noise, the microphone position is not changed, and the experiment is conducted under a specific reverberation condition. Therefore, we experimentally investigate the robustness of the spatial cepstrum under conditions closer to the real environment. The experimental results show that changes in the experimental conditions only affect the high-dimensional spatial cepstrum; thus, the low-dimensional spatial cepstrum is a robust feature that is not easily affected by disturbances.**

I. Introduction

In recent years, devices that can easily record sounds, such as smartphones and tablets, have become widespread. Also, speech enhancement and acoustic scene analysis have been researched using a distributed microphone array consisting of these devices [1], [2], [3]. In distributed microphone arrays, the microphone position and array geometry are unknown, and the microphones are asynchronous. Some researchers are working on solving these problems [4], [5], [6], [7], [8]. Other researchers have tried to apply spatial information extracted from the distributed microphone arrays for acoustic scene analysis [9], [10], [11], [12], [13], [14].

In the research of acoustic scene analysis, the cepstrum and mel-frequency cepstrum are generally used as features. However, we cannot utilize the advantage of the microphone array because these features do not include spatial information. Moreover, when acquiring spatial information in a distributed microphone array, there are problems such as mismatching of sampling frequencies and/or loss of arrival time of the signal awing to synchronization. It is known that such problems are difficult to solve.

Imoto and Ono proposed a new feature called the spatial cepstrum [15]. The spatial cepstrum has been proposed for extracting spatial information using a distributed microphone array that does not require the positions of microphones or a precise time synchronization among microphones. The cepstrum be obtained by performing the discrete Fourier transform (DFT) on the spectrum. In contrast, we obtain the spatial cepstrum by principal component analysis (PCA) instead of DFT. Imoto and Ono argue that the spatial cepstrum is robust against changes in recording conditions. However, the behavior of the spatial cepstrum has been shown in the minimal experiment where the sound source is only white Gaussian noise, the microphone position is not changed, and the experiment is conducted under a specific reverberation condition. In this study, we conduct the following three experiments to confirm the behavior of the spatial cepsturm in detail: (i) simulation with sound sources close to real-environment sound sources, (ii) simulation with the various positions of some microphones in a distributed microphone array, and (iii) the experiment in a real environment with various positions of some microphones in a distributed microphone array experiment. The experimental results show that changes in the experimental condition only affect the high-dimensional spatial cepstrum; thus, the low-dimensional spatial cepstrum is a robust feature that is not easily affected by disturbances.

II. Cepstrum and spatial cepstrum

A. Cepstrum

We denote N as the number of microphones and $s_{\omega,\tau,n}$ as the short-time Fourier transform (STFT) representations of the n-th channel observation with ω, τ, and n representing the frequency bin, time frame, and channel indices, respectively. Additionally, we define $a_{\omega,\tau,n}$ as the amplitude of $s_{\omega,\tau,n}$, indicated by $a_{\omega,\tau,n} = |s_{\omega,\tau,n}|$. To extract spectral information, we consider the frequency-based log-amplitude vector

$$\boldsymbol{p}_\tau = (p_{1,\tau}, p_{2,\tau}, \cdots, p_{\Omega,\tau})^{\mathsf{T}}, \tag{1}$$

$$p_{\omega,\tau} = \log \sqrt{\frac{1}{N}\sum_n a^2_{\omega,\tau,n}}, \tag{2}$$

where Ω is the number of frequency bins and $^{\mathsf{T}}$ is the transpose of the vector and matrix. We can obtain a cepstrum as

$$\boldsymbol{c}_\tau = \boldsymbol{Z}_\Omega \boldsymbol{p}_\tau, \tag{3}$$

where $\boldsymbol{Z}_\Omega$ is the $\Omega \times \Omega$ DFT matrix. In the case of a cepstrum, spatial information is lost because it is the average among channels in (1).

B. Spatial Cepsrum

In analogy with the definition of the cepstrum, we consider the channel-based log-amplitude vector

$$\boldsymbol{q}_\tau = (q_{\tau,1}, q_{\tau,2}, \cdots, q_{\tau,N})^\mathsf{T}, \tag{4}$$

$$q_{\tau,n} = \log \sqrt{\frac{1}{\Omega} \sum_\omega a^2_{\omega,\tau,n}}. \tag{5}$$

We consider the orthogonal transformation of $\boldsymbol{q}_\tau$ in the same way as in the case of the cepstrum. DFT, one of the orthogonal transforms, requires that the subbands are uniformly spaced on the linear axis. DFT can be applied as in (3) because the frequency-based log-amplitude vector $\boldsymbol{p}_\tau$ is uniformly spaced on the linear frequency axis. By contrast, we need to place the microphones uniformly to apply DFT to $\boldsymbol{q}_\tau$ because the channel-based log-amplitude vector $q_{\tau,n}$ is placed on the channel axis. However, we cannot apply DFT to $\boldsymbol{q}_\tau$ because the microphones in the distributed microphone array most likely be nonuniformly placed. Therefore, we perform orthogonal transform by PCA instead of DFT. We consider the covariance matrix $\boldsymbol{R}_q$ of $\boldsymbol{q}_\tau$ to perform PCA:

$$\boldsymbol{R}_q = \frac{1}{M} \sum_\tau \boldsymbol{q}_\tau \boldsymbol{q}_\tau^\mathsf{T}, \tag{6}$$

where M is the number of time frame. We obtain the transformation matrix by performing the eigendecomposition on the obtained $\boldsymbol{R}_q$ as follows:

$$\boldsymbol{R}_q = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\mathsf{T}, \tag{7}$$

where $\boldsymbol{E}$ is the eigenvector matrix and $\boldsymbol{D}$ is the diagonal matrix in which the diagonal elements are eigenvalues in descending order. $\boldsymbol{E}$ becomes the DFT matrix $\boldsymbol{Z}_\Omega$ when the covariance matrix $\boldsymbol{R}_q$ is a circular matrix. We can perform orthogonal transformation on $\boldsymbol{q}_\tau$ in the same way as in the case of the cepstrum by using $\boldsymbol{E}$. The spatial cepstrum is defined using $\boldsymbol{E}$ as

$$\boldsymbol{d}_\tau = \boldsymbol{E}^\mathsf{T} \boldsymbol{q}_\tau. \tag{8}$$

Imoto *et al.* assert that relative spatial information can be obtained by using the spatial cepstrum [15].

III. Experiments

In previous studies, it was not tested in detail whether the spatial cepstrum is robust to changing conditions. Therefore, we perform two simulations with the conventional experimental conditions changes as shown in Table I and Fig. 1. We use the audio signal processing software Pyroomacoustics for the simulation [16]. Additionally, experiment (iii) is similar to experiment (ii) but the move of microphones in the real environment. Throughout the three experiments, the window function is a Hamming window, the window length and frame length are both 512, the amount of frameshift is 128, and the sampling frequency is 16000 Hz. In the experiment, to confirm the behavior of the spatial cepstrum, we compare the components of the spatial cepstrum for each microphone and the dimensions of the spatial cepstrum. $\boldsymbol{d}_\tau$ does not have the components of the spatial cepstrum for each microphone. Therefore, we define a vector $\boldsymbol{d}_k$ that sums the components of the spatial cepstrum of each microphone on the k order, as

$$\boldsymbol{d}_k = \frac{1}{M} \sum_\tau \boldsymbol{e}_k^\mathsf{T} \circ \boldsymbol{q}_\tau \tag{9}$$

where $\boldsymbol{e}_k^\mathsf{T}$ is the k-th row vector of $\boldsymbol{E}^\mathsf{T}$ and $\circ$ is the Hadamard product. Imoto *et al.* assert that $\boldsymbol{d}_1$ indicates the average sound level of the whole space [15]. Therefore, $\boldsymbol{d}_1$ is not consider in this study.

TABLE I
Experimental conditions in conventional study.

Arrangement	Fig. 1
Sound source	White Gaussian noise
Overlap	–
Playback order	S1 to S6 in Fig. 1
Reverberation	–

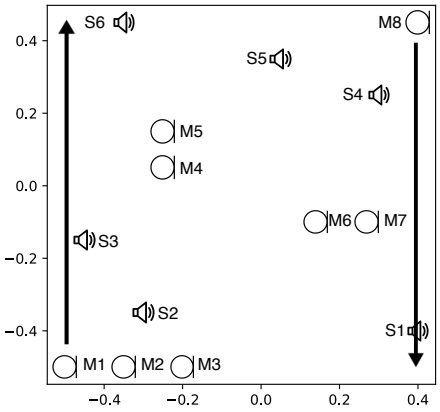

Fig. 1. Arrangement of microphones and speakers in simulation experiment.

A. Robustness of Spatial Cepstrum against Sound Source Changes

In experiment (i), we use white noise and other sound sources. We confirm the robustness of the spatial cepstrum to changing sound sources on the basis of the correlation coefficient between $\boldsymbol{d}_k$ obtained using white noise and that obtained using other sound sources.We add two types of wood hitting sound like impulsive sounds and two types of human voice like wave-like sounds, which have not been investigated in [15]. The other experimental conditions are the same as in Table I. We use wood hitting sounds in the Real World Computing Partnership (RWCP) [17] and human voices in the Japanese Newspaper Article Sentences Read Speech Corpus (JNAS) [18]. Table II shows the absolute value of the correlation coefficient between $\boldsymbol{d}_k$ obtained using

TABLE II
ABSOLUTE VALUE OF THE CORRELATION COEFFICIENT BETWEEN $\boldsymbol{d}_k$ OBTAINED USING WHITE NOISE AND THAT OBTAINED USING OTHER SOUND SOURCES.

Dimensions	2	3	4	5	6	7	8
wood hitting1	**0.99**	**0.88**	**0.86**	0.69	0.44	0.20	0.34
wood hitting2	**0.99**	**0.81**	**0.76**	0.36	**0.96**	0.42	0.32
human voice1	**0.98**	**0.97**	**0.90**	0.05	0.16	0.46	0.18
human voice2	**0.99**	**0.87**	**0.83**	**0.83**	**0.82**	0.62	0.63

white noise and that obtained using other sound sources. The horizontal axis is the number of dimensions used to calculate the correlation. From Table II, we can see that we have a high correlation of 0.7 or more in 2 to 4 dimensions. On the other hand, correlations are lower in 5 to 8 dimensions than in lower dimensions. Therefore, we find that there is little difference in the behavior of the low-dimensional $\boldsymbol{d}_k$ and that the low-dimensional spatial cepstrum is robust against sound source changes. The high-dimensional spatial cepstrum is less robust against sound source changes.

B. Robustness of Spatial Cepstrum against Movement of Microphones

In experiment (ii), we move two microphones, the non-isolated microphone M1 and the isolated microphone M8. The non-isolated microphone M1 is close to other microphones, while the isolated microphone M8 is far from other microphones in Fig. 1. We confirm the robustness of the spatial cepstrum to the movement of a microphone on the basis of the correlation coefficient between $\boldsymbol{d}_k$ before and after moving the microphones. We move the two microphones as shown by the arrows in Fig. 1. Other experimental conditions are the same as those in Table I. First, we move the non-isolated microphone M1 upwards in increments of 10 mm to a total of 950 mm. Figure 2 shows the absolute value of the correlation coefficient between $\boldsymbol{d}_k$ obtained at the initial position and that at the moved position of microphone M1. The horizontal axis is the displacement from the initial position of microphone M1, and the vertical axis is the absolute value of the correlation coefficient. From Fig. 2, we can see that the correlation decreases with increasing displacement of microphone M1 from the initial position. However, the correlation is low when the distance from the initial position of microphone M1 is between 20 cm and 50 cm. We speculate that the cause is microphone M1 being close to sound source S3.

Next, we change the position of isolated microphone M8 downwards in increments of 1 cm to a total of 95 cm. Figure 3 shows the absolute value of the correlation coefficient between $\boldsymbol{d}_k$ obtained at the initial position and at the moved position of microphone M8. From Fig. 3, we can see that there is little difference in the spatial cepstrum behavior in lower dimensions. Also, we can see that the correlation is low at around the distance of 80 cm where microphone M8 is close to sound source S1. In both case, if the displacement is about 20 cm or less, there is almost no effect on the low-dimensional spatial cepstrum. Therefore, do not have to meet the restriction

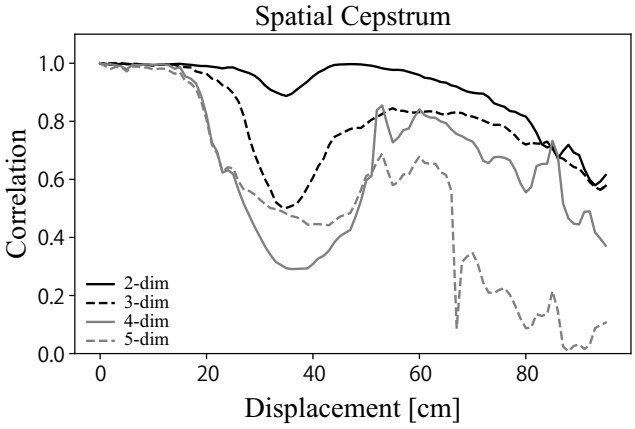

Fig. 2. Absolute value of the correlation coefficient between $\boldsymbol{d}_k$ obtained at initial position and that at moved position of microphone M1.

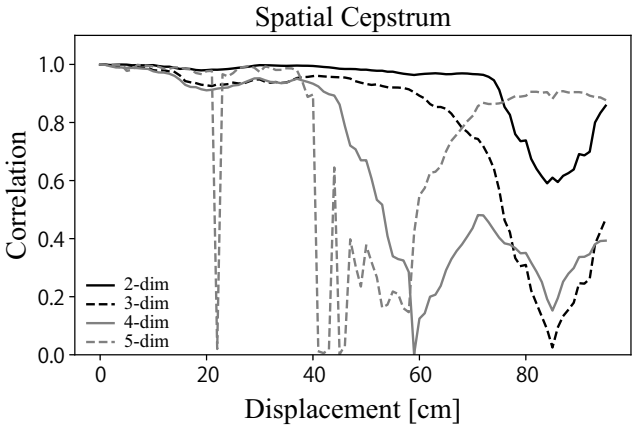

Fig. 3. Absolute value of the correlation coefficient between $\boldsymbol{d}_k$ obtained at initial position and that at moved position of microphone M8.

of strict placement of the microphone when actually recording. From these results, we find that the low-dimensional spatial cepstrum is robust against the movement of a microphone.

C. Evaluation of Robustness of Spatial Cepstrum in Real Environment

In experiment (iii), we carry out the experiment similarly to experiment (ii) but the move the of microphones in the real environment. We use part of the SINS database, which is a continuous recording of one person living in a vacation home over a period of one week, as real-environment sound sources [19]. The layout is shown in Fig. 4. Each node in Fig. 4 is a four-channel linear microphone array, and the distance between microphones is 5 cm. We cannot use Node 5 because the node is problematic and has not been published. We change the microphone positions in only the remaining nodes. We obtain recordings in one channel of Node 8 (double circle in Fig. 4) and three channels of each of the other nodes (black circles in Fig. 4) for a total of nineteen channels. We

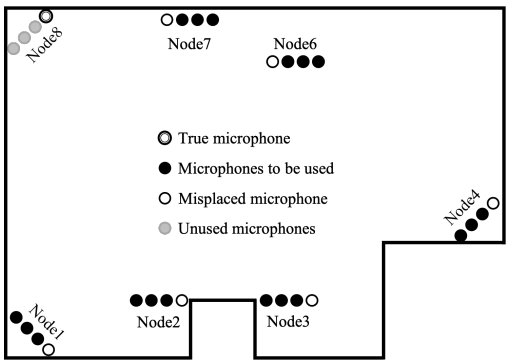

Fig. 4. Arrangement of microphones in real environment.

TABLE III
ABSOLUTE VALUE OF THE CORRELATION COEFFICIENT BETWEEN $\boldsymbol{d}_k$ OBTAINED BY TRUE MICROPHONE AND THAT OBTAINED BY MISTAKEN MICROPHONES.

Dimensions	2	3	4	5	6	7	8
Node 1	**1.00**	**1.00**	**1.00**	**0.99**	**0.84**	**0.81**	0.22
Node 2	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	0.68	0.65
Node 3	**0.90**	**0.87**	**0.84**	**0.97**	**0.97**	0.66	0.64
Node 4	**1.00**	**1.00**	**0.98**	**0.99**	0.68	0.59	0.67
Node 6	**1.00**	**0.95**	**0.89**	**1.00**	**0.98**	0.63	0.69
Node 7	**1.00**	**0.98**	**0.95**	**0.89**	**0.95**	**0.74**	**0.70**

place the channel of Node 8 incorrectly in the position of the fourth channel of another node (white circle in Fig. 4) and check whether the spatial cepstrum is robust against this misplacement. Similarly to previous experiments, we evaluate robustness on the basis of the correlation coefficient of $\boldsymbol{d}_k$ obtained for true and mistaken microphones. Table III shows the experimental results. The horizontal axis is the number of dimensions used to calculate the correlation and the vertical axis is the misplaced node. As a result, we found a high correlation in low dimensions and a low correlation in high dimensions. These results are similar to those in section III-B, and we conclude that the low-dimensional spatial cepstrum is robust even in a real environment.

IV. CONCLUSION

In this study, we confirmed the robustness of the spatial cepstrum. We predicted that the spatial cepstrum is robust against changes in sound sources and the movement of a microphone because it is a feature of amplitude. We changed the sound source and moved the microphone by simulation. From these simulations, we confirmed that spatial information can be robustly acquired using only low dimensions of the spatial cepstrum because elements related to changes in conditions, such as changes in sound source and microphone positions, are concentrated in high dimensions. We found that even for real-environment sound sources, spatial information could be robustly acquired. In the conventional method, the placement of microphones was strict, but this is not the case when using spatial cepstrum, which is a great practical advantage.

V. ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for Scientific Research (A) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 20H00613).

REFERENCES

[1] A. Bertrand, "Applications and Trends in Wireless Acoustic Sensor Networks: A Signal Processing Perspective," *Proc. IEEE Symposium on Communications and Vehicular Technology in the Benelux*, pp. 1–6, 2011.

[2] S. Araki *et al.*, "Meeting Recognition with Asynchronous Distributed Microphone Array," *Proc. Automatic Speech Recognition and Understanding Workshop*, pp. 32–39, 2017.

[3] M. Souden *et al.*, "An Integration of Source Location Cues for Speech Clustering in Distributed Microphone Arrays," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 111–115, 2013.

[4] N.Ono *et al.*, "Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 161–164, 2009.

[5] E. R. Arnuncio *et al.*, "On Dealing with Aampling Rate Mismatches in Blind Source Separation and Acoustic Echo Cancellation," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 34–37, 2007.

[6] Z. Liu, "Sound Source Separation with Distributed Microphone Arrays in the Presence of Clock Synchronization Errors," *Proc. International Workshop for Acoustic Echo and Noise Control*, 2008.

[7] R.Sakanashi *et al.*, "Speech Enhancement with Ad-hoc Microphone Array Using Single Source Activity," *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–6, 2013.

[8] S. Markovich-Golan *et al.*, "Blind Sampling Rate Offset Estimation and Compensation in Wireless Acoustic Sensor Networks with Application to Beamforming," *Proc. International Workshop for Acoustic Echo and Noise Control*, pp. 1–4, 2012.

[9] J. P. Bello *et al.*, "Sound Analysis in Smart Cities," *Computational Analysis of Sound Scenes and Events*, pp. 373–397, 2018.

[10] R. Radhakrishnan *et al.*, "Audio Analysis for Surveillance Applications," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 158–161, 2005.

[11] E. Wold *et al.*, "Content-based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[12] Q. Jin *et al.*, "Event-based Video Retrieval Using Audio," *Proc. Interspeech*, 2012.

[13] L. Lin *et al.*, "Guided Learning Convolution System for DCASE 2019 TASK 4," *Tech. Rep. DCASE 2019 Challenge Task4*, 2019.

[14] L. D. Poulat, and C. Plapous, "Mean Teacher with Data Augmentation for DCASE 2019 TASK 4," *Tech. Rep. DCASE 2019 Challenge Task4*, 2019.

[15] K. Imoto, and N. Ono, "Spatial Cepstrum as a Spatial Feature Using a Distributed Microphone Array for Acoustic Scene Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, 2017.

[16] R. Scheibler *et al.*, "Pyroomacoustics: A Python Package for Audio Room Simulations and Array Processing Algorithms," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 351–355, 2018.

[17] S. Nakamura *et al.*, "Sound Scene Data Collection in Real Acoustical Environments," *Journal of the Acoustical Society of Japan*, vol. 20, pp. 225–231, 1999.

[18] K. Ito *et al.*, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," *The Journal of Acoustical Society of Japan*, vol. 20, pp. 196–206, 1999.

[19] G. Dekkers *et al.*, "The SINS Database for Detection of Daily Activities in a Home Environment Using an Acoustic Sensor Network," *Proc. Detection and Classification of Acoustic Scenes and Events*, pp. 32–36, 2017.