Acoustic Scene Analysis Using Partially Connected Microphones Based on Graph Cepstrum

Keisuke Imoto

Ritsumeikan University, Japan

Abstract—In this paper, we propose an effective and robust method for acoustic scene analysis based on spatial information extracted from partially synchronized and/or closely located distributed microphones. In the proposed method, to extract spatial information from distributed microphones while taking into account whether any pairs of microphones are synchronized and/or closely located, we derive a new cepstrum feature utilizing a graph-based basis transformation. Specifically, in the proposed graph-based cepstrum, the logarithm of the amplitude in a multichannel observation is converted to a feature vector by an inverse graph Fourier transform, which can consider whether any pair of microphones is connected. Our experimental results indicate that the proposed graph-based cepstrum effectively extracts spatial information with consideration of the microphone connections. Moreover, the results show that the proposed method more robustly classifies acoustic scenes than conventional spatial features when the observed sounds have a large synchronization mismatch between partially synchronized microphone groups.

I. INTRODUCTION

Acoustic scene analysis (ASA), which analyzes scenes in which sounds are produced, is now a very active research area in acoustics, and it is expected that ASA will enable many useful applications such as systems monitoring elderly people or infants [1], [2], automatic surveillance systems [3]–[6], automatic file-logging systems [7]–[9], and advanced multimedia retrieval [10]–[13].

To analyze scenes from an acoustic signal, many approaches based on machine learning techniques have been proposed. For instance, Eronen *et al.* [7] and Mesaros *et al.* [14] have proposed spectral feature-based methods such as mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture models (GMMs). Han *et al.* [15] and Jallet *et al.* [16] have proposed methods using the mel-spectrogram as input features and the convolutional neural network (CNN) or recurrent convolutional neural network (RCNN) as classifiers. Guo and Li [17], Kim *et al.* [18], and Imoto and co-workers [8], [19] have investigated ASA utilizing intermediate feature representations based on acoustic event histograms.

ASA based on spatial information extracted from a microphone array composed of smartphones, smart speakers, and IoT devices has also been proposed [20]–[22]. Many of these methods extract spatial information based on observed time differences or sound power ratios between channels, and therefore, they require that the microphones are synchronized and the microphone locations and array geometry are known. However, since the distributed microphones in multiple smartphones, smart speakers, or IoT devices are often unsynchronized and the microphone locations and array geometry are
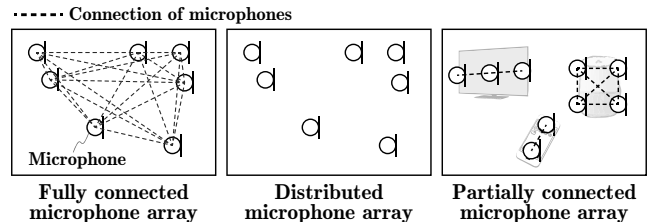

Fig. 1. Example of microphone connections

unknown, conventional methods cannot be applied to such distributed microphone arrays. To extract spatial information using unsynchronized distributed microphones whose locations and array geometry are unknown, Imoto and Ono have proposed a spatial cepstrum that can be applied under these conditions [23]. In this approach, log-amplitudes obtained by multiple microphones are converted to a feature vector similarly to when using the cepstrum, which is based on principal component analysis (PCA) of the feature vector.

On the other hand, the numbers of smartphones, smart speakers, or IoT devices that have multiple microphones have been increasing. A microphone array composed of these microphones are often partially synchronized or closely located as shown in Fig. 1; we refer to these synchronized or closely located microphones collectively as *connected* microphones. The time delay or sound power ratio between channels is a significant cue for extracting spatial information even when the microphones are partially connected; however, the conventional spatial cepstrum does not consider whether some of the microphones are partially connected.

In this paper, we propose a novel spatial feature extraction method for a distributed microphone array that can take into account whether or not microphones are partially connected. To consider whether any pairs of microphones are connected, we utilize a graph representation of the microphone connections, where the power observations and microphone connections are represented by the weights of the nodes and edges, respectively. Then, the proposed method introduces a graph Fourier transform, which enables spatial feature extraction considering the connections between microphones.

This paper is organized as follows. In section 2, the spatial cepstrum used in conventional spatial feature extraction for a distributed microphone array is introduced. In section 3, the proposed method of extracting a spatial feature for partially connected distributed microphones and the similarity of the proposed method to the conventional cepstrum and spatial

cepstrum are discussed. In section 4, experiments performed to evaluate the proposed method are reported. In section 5, we conclude this paper.

II. CONVENTIONAL SPATIAL FEATURE EXTRACTION FOR DISTRIBUTED MICROPHONES

To extract spatial information from unsynchronized distributed microphones whose locations and array geometry are unknown, the spatial cepstrum, which is a similar technique to the *cepstrum* feature, has been proposed [23].

Suppose that a multichannel observation is recorded by N microphones and $\bar{a}_{\tau,n}$ denotes the power observed for microphone n at time frame τ. In the case of unsynchronized distributed microphones, synchronization over channels is still a challenging problem and phase information may be unreliable. Therefore, the spatial cepstrum utilizes only the log-amplitude vector

$$\mathbf{q}_\tau = \begin{pmatrix} \log \bar{a}_{\tau,1} \\ \log \bar{a}_{\tau,2} \\ \vdots \\ \log \bar{a}_{\tau,n} \\ \vdots \\ \log \bar{a}_{\tau,N} \end{pmatrix}, \tag{1}$$

which is relatively robust to a synchronization mismatch. Considering that the distributed microphones may be non-uniformly located, PCA is then applied for the basis transformation of the spatial cepstrum instead of the inverse discrete Fourier transform (IDFT). Suppose that $\mathbf{R}_q$ is the covariance matrix of $\mathbf{q}_\tau$ and given by

$$\mathbf{R}_q = \frac{1}{T} \sum_\tau \mathbf{q}_\tau \mathbf{q}_\tau^\mathsf{T}, \tag{2}$$

where T is the number of time frames. Since $\mathbf{R}_q$ is a symmetric matrix, the eigendecomposition of $\mathbf{R}_q$ can be represented as

$$\mathbf{R}_q = \mathbf{E}\mathbf{D}\mathbf{E}^\mathsf{T}, \tag{3}$$

where $\mathbf{E}$ and $\mathbf{D}$ are the eigenvector matrix and the diagonal matrix whose diagonal elements are equal to the eigenvalues in descending order, respectively. Using this eigenvector matrix $\mathbf{E}$, the spatial cepstrum is defined as

$$\mathbf{d}_\tau = \mathbf{E}^\mathsf{T}\mathbf{q}_\tau. \tag{4}$$

The spatial cepstrum can extract spatial information without microphone locations or the array geometry, although it requires training sounds to estimate the eigenvector matrix $\mathbf{E}$ by PCA. Moreover, since the spatial cepstrum does not consider whether or not the microphones are connected, observed time differences or sound power ratios between channels cannot be utilized for spatial feature extraction.
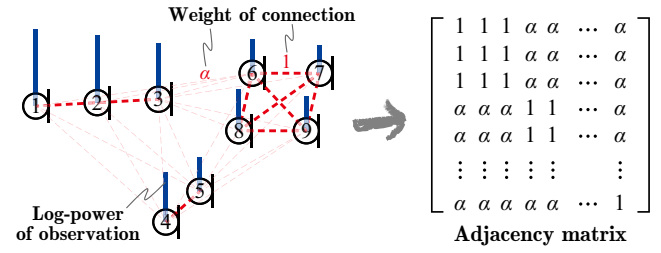

Fig. 2. Example of observations on graph and relationship between microphone connections and adjacency matrix

III. SPATIAL FEATURE EXTRACTION BASED ON GRAPH CEPSTRUM

A. Graph Cepstrum

We consider the situation that a microphone array is composed of multiple generic acoustic sensors mounted on smartphones, smart speakers, or IoT devices, where some of the microphones mounted on each device are connected. To extract spatial information while considering microphone connections, we here propose a novel spatial feature extraction method that utilizes a graph representation of the multichannel observations and microphone connections. Specifically, to extract spatial information, the proposed method performs the graph Fourier transform [24] instead of PCA in the spatial cepstrum. This makes it possible to take into account which pairs of microphones are connected.

Consider the logarithm powers of the observations on the graph shown in Fig. 2, where the power observations and microphone connections are represented by the weights of the nodes and edges, respectively. Here, the $N \times N$ adjacency matrix is defined as

$$\mathbf{A}(m,n) = \begin{cases} 1 & \text{if channels } m \text{ and } n \text{ are connected} \\ 0 \text{ or } \alpha & otherwise, \end{cases} \tag{5}$$

where α is an arbitrary weight of the connection within the range of 0.0–1.0. We also assume the $N \times N$ degree matrix $\mathbf{D}$, which is a diagonal matrix whose diagonal elements are represented as

$$\mathbf{D}(m,m) = \sum_n \mathbf{A}(m,n). \tag{6}$$

The degree matrix indicates the number of microphones connected with microphone m. Then, the unweighted graph Laplacian is written as

$$\mathbf{L} \triangleq \mathbf{D} - \mathbf{A}, \tag{7}$$

where $\mathbf{L}$ is also a symmetric matrix since both $\mathbf{D}$ and $\mathbf{A}$ are symmetric matrices. Thus, eigendecomposition of $\mathbf{L}$ can be expressed as

$$\mathbf{L} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}, \tag{8}$$

where $\mathbf{U}$ and $\boldsymbol{\Lambda}$ are the eigenvector matrix and the diagonal matrix whose diagonal elements λ_m are equal to the eigenvalues in ascending order, respectively. The eigenvector
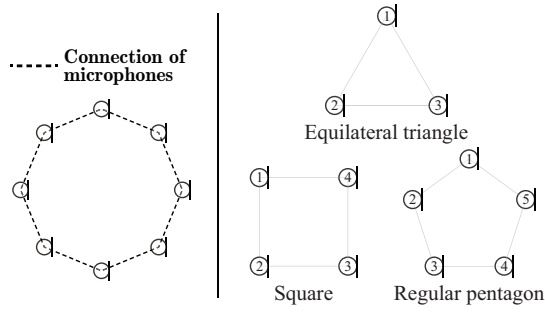
Fig. 3. Examples of ring graph condition (left) and circularly symmetric microphone arrangements (right)

matrix $\mathbf{U}$ and its transpose $\mathbf{U}^T$ are the graph Fourier transform (GFT) matrix and the inverse graph Fourier transform (IGFT) matrix, respectively, which enable the basis transformations considering the connections between microphones.

Thus, the proposed spatial feature, which can consider the connections between microphones, is defined in terms of the IGFT of the log-amplitude vector $\mathbf{q}_\tau$ as

$$\mathbf{e}_\tau = \mathbf{U}^\mathsf{T}\mathbf{q}_\tau. \tag{9}$$

Because this proposed spatial feature also resembles the conventional cepstrum as well as the spatial cepstrum, we call it the graph cepstrum (GC).

B. Graph Cepstrum on Ring Graph

Let us consider a circular connected condition, namely the ring graph condition shown in Fig. 3. For this condition, a graph Laplacian is represented as the circulant matrix

$$\mathbf{L}_{sym} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & -1 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ -1 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}. \tag{10}$$

On the basis of the fact that a circulant matrix is diagonalized by an IDFT matrix $\mathbf{Z}_N$ [25] defined by

$$\mathbf{Z}_N = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & \zeta^1 & \zeta^2 & \cdots & \zeta^{N-2} & \zeta^{N-1} \\ 1 & \zeta^2 & \zeta^4 & \cdots & \zeta^{2(N-2)} & \zeta^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \zeta^{N-2} & \zeta^{2(N-2)} & \cdots & \zeta^{(N-2)^2} & \zeta^{(N-1)(N-2)} \\ 1 & \zeta^{N-1} & \zeta^{2(N-1)} & \cdots & \zeta^{(N-2)(N-1)} & \zeta^{(N-1)^2} \end{bmatrix} \tag{11}$$

$$\zeta = e^{j2\pi/N}, \tag{12}$$

the IGFT is identical to the IDFT.

Thus, in the case of a ring graph, the GC is identical to the definition of the cepstrum. Moreover, it is also identical to the definition of the spatial cepstrum of circular symmetric
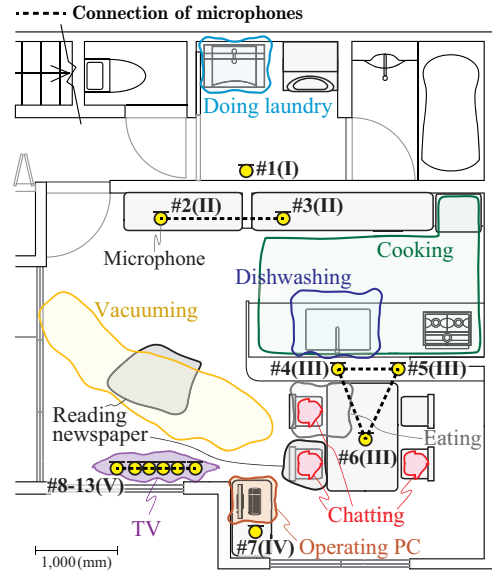

Fig. 4. Microphone arrangement and sound source locations. Channel indices (1–13) and group indices of synchronized microphones (I–V) are also indicated.

microphones in an isotropic sound field [23]. This means that the ring connection in the GC domain corresponds to the circular symmetric arrangement of microphones in an isotropic sound field in the acoustic spatial condition.

IV. EXPERIMENTS

A. Experimental Conditions

To evaluate the effectiveness of the proposed method for partially synchronized microphones, we conducted classification experiments on acoustic scenes in a living room. Since most of the public datasets for acoustic scene analysis including TUT Acoustic Scenes 2017 [26] and AudioSet [27] are provided in single or stereo channels, we recorded a multichannel sound dataset with 13 synchronized microphones in a real environment. The sound dataset includes nine acoustic scenes, "vacuuming," "cooking," "dishwashing," "eating," "reading a newspaper," "operating a PC," "chatting," "watching TV," and "doing the laundry," which happen frequently around the living room. The microphone arrangement and the locations of the sound sources are shown in Fig. 4. The recorded sounds consisted of 257.1 min. of recordings, which were randomly separated into 5,180 sound clips for model training and 2,532 sound clips for classification evaluation, where no acoustic scene overlapped with another scene in all the sound clips. To evaluate the scene classification performance with synchronization mismatch among the microphone groups, the recorded sounds for classification evaluation were misaligned with various error times among the microphone groups shown in Fig. 4. The error times were randomly sampled from a Gaussian distribution with $\mu = 0$ and various variances σ^2. The other recording conditions and experimental conditions are listed in Table I.

TABLE I
EXPERIMENTAL CONDITIONS

Sampling rate	48 kHz
Quantization bit rate	16 bits
Sound clip length	8 s
Frame length/FFT point	20 ms/2,048
Connection weight α	0.01
Network structure of CNN	3 conv. & 3 dense layers
Pooling in CNN layers	2×2 max pooling
Activation function	ReLU, softmax (output layer)
# channels of CNN	32, 24, 16
# units of dense layers	128, 64, 32
Optimizer	Adam

Fig. 5. IGFT matrix $\mathbf{U}^\mathsf{T}$ in red-blue color map representation

B. Spatial Information Extracted by Graph Cepstrum

To clarify how the GC extracts spatial information, we show the IGFT matrix $\mathbf{U}^\mathsf{T}$ in Fig. 5. The nth-row vector of $\mathbf{U}^\mathsf{T}$ corresponds to the nth-eigenvector of the graph Laplacian $\mathbf{L}$. The nth-order GC is calculated using the nth-row vector of $\mathbf{U}^\mathsf{T}$ as follows:

$$e_{\tau,n} = \mathbf{u}_n^\mathsf{T}\mathbf{q}_\tau = \sum_{m=1}^{N} u_{n,m} q_{\tau,m}, \qquad (13)$$

where $e_{\tau,n}$, $\mathbf{u}_n^\mathsf{T}$, $u_{n,m}$, and $q_{\tau,m}$ are the nth-order GC, the nth-row vector of $\mathbf{U}^\mathsf{T}$, the (n,m) entry of $\mathbf{U}^\mathsf{T}$, and the mth element of $\mathbf{q}_\tau$, respectively. This indicates that the nth-order GC is obtained by a linear combination of log-amplitudes $q_{\tau,m}$, where $u_{n,m}$ is the weight of the linear combination. From Fig. 5, it can be interpreted that the first-order GC represents the average sound level in the whole space because all the weights $\mathbf{u}_n^\mathsf{T}$ are positive. For the middle-order eigenvectors, the signs of the weights between connected microphones are similar. This indicates that the GC can capture spatial information while taking the connections of microphones into account. For the higher-order eigenvectors, the weights of only part of the connected microphone group are active and the signs of the weights differ. These eigenvectors capture spatial information of the sound sources close to the microphone groups because if the sound sources are far from the microphone groups, the linear combination of the microphone groups is canceled in Eq. (13).

C. Acoustic Scene Classification

Acoustic scenes were then modeled and classified for each sound clip using a Gaussian mixture model (GMM), a supervised acoustic topic model (sATM) [8], [28], and a convolutional neural network (CNN). Specifically, the GMM was applied to acoustic feature vectors $\mathbf{e}_\tau$ and $\mathbf{d}_\tau$ for each acoustic scene x. After that, acoustic scene x of sound clip c was estimated by calculating the product of the likelihoods over the sound clip as follows:

$$x_c = \arg \max_x \prod_{\tau=1}^{T_c} p_\tau(\mathbf{f}_\tau|x), \qquad (14)$$

where T_c, $\mathbf{f}_\tau$, and $p_\tau(\mathbf{f}_\tau|x)$ are the number of frames in sound clip c, an acoustic feature vector calculated frame by frame,
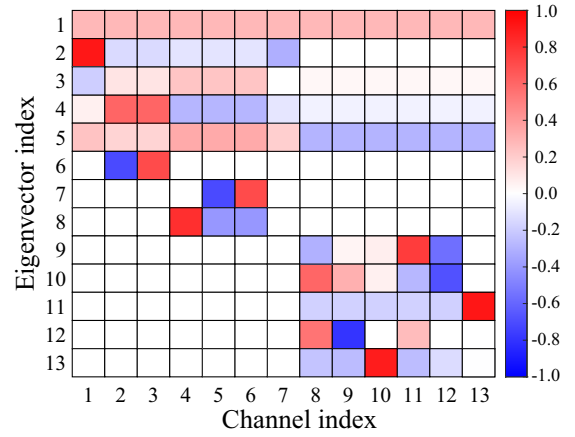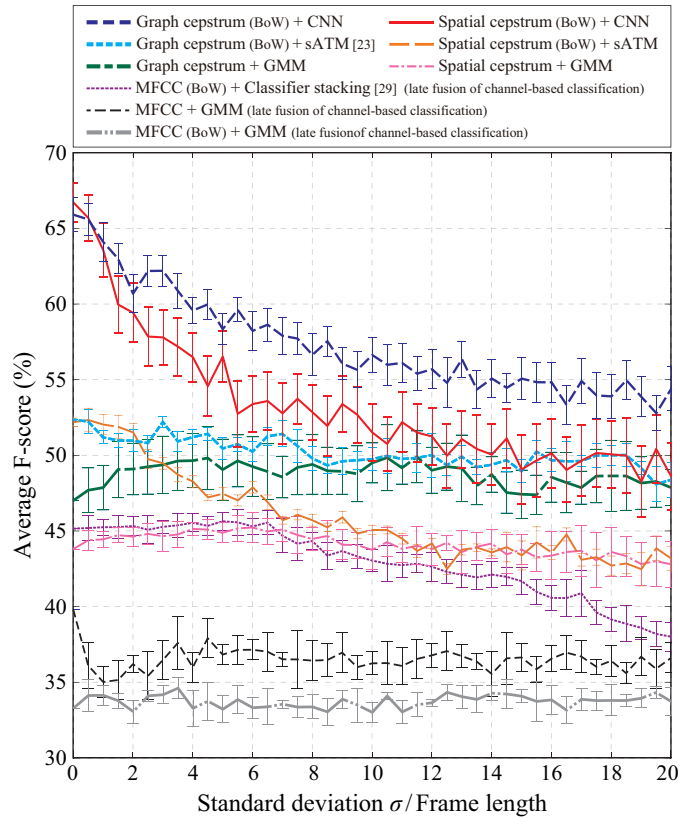

Fig. 6. Acoustic scene classification accuracy with various synchronization error times between connected microphone groups

such as $\mathbf{d}_\tau$ or $\mathbf{e}_\tau$, and the likelihood of acoustic scene x at time frame τ, respectively. As other methods for acoustic scene classification utilizing a distributed microphone array, we also evaluated classifiers based on late fusion-based classification methods [29].

D. Experimental Results

The classification performance of acoustic scenes is shown in Fig. 6. For each experimental condition, the acoustic scene modeling and classification were conducted ten times with

various synchronization error times sampled randomly. These results show that when the synchronization error between microphone groups is small, the GC and conventional spatial cepstrum effectively classify acoustic scenes. When the synchronization error between microphone groups increases, the scene classification performance for the GC slightly decreases. In contrast, the classification accuracy decreases rapidly when using conventional methods. This indicates that the proposed GC is more robust against synchronization error than conventional methods.

V. CONCLUSION

In this paper, we proposed an effective spatial feature extraction method for acoustic scene analysis using partially synchronized or closely located distributed microphones. In the proposed method, we derived the graph cepstrum (GC), which is defined as the inverse graph Fourier transform of the logarithm power of a multichannel observation. We then demonstrated that the GC in a ring graph is identical to the conventional cepstrum and spatial cepstrum in a circularly symmetric microphone arrangement with an isotropic sound field. Our experimental results using real environmental sounds showed that the GC more robustly classifies acoustic scenes than conventional spatial features even when the synchronization mismatch between partially synchronized microphone groups is large.

ACKNOWLEDGMENTS

REFERENCES

[1] Y. Peng, C. Lin, M. Sun, and K. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1218–1221, 2009.
[2] P. Guyot, J. Pinquier, and R. André-Obrecht, "Water sound recognition based on physical models," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 793–797, 2013.
[3] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2005.
[4] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 158–161, 2005.
[5] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 165–168, 2009.
[6] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 376–380, 2017.
[7] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, 2006.
[8] K. Imoto and S. Shimauchi, "Acoustic scene analysis based on hierarchical generative model of acoustic event sequence," *IEICE Trans. Inf. Syst.*, vol. E99-D, no. 10, pp. 2539–2549, October 2016.
[9] J. Schröder, J. Anemüller, and S. Goetze, "Classification of human cough signals using spectro-temporal Gabor filterbank features," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6455–6459, 2016.

[10] T. Zhang and C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 9, no. 4, pp. 441–457, 2001.
[11] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," *Proc. INTERSPEECH*, 2012.
[12] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino, "Bayesian semi-supervised audio event transcription based on Markov Indian buffet process," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3163–3167, 2013.
[13] J. Liang, L. Jiang, and A. Hauptmann, "Temporal localization of audio events for conflict monitoring in social media," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1597–1601, 2017.
[14] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," *Proc. 18th European Signal Processing Conference (EUSIPCO)*, pp. 1267–1271, 2010.
[15] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.
[16] H. Jallet, E. Çakır, and T. Virtanen, "Acoustic scene classification using convolutional recurrent neural networks," *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.
[17] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 209–215, 2003.
[18] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," *Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 37–40, 2009.
[19] K. Imoto, Y. Ohishi, H. Uematsu, and H. Ohmuro, "Acoustic scene analysis based on latent acoustic topic and event allocation," *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013.
[20] H. Kwon, H. Krishnamoorthi, V. Berisha, and A. Spanias, "A sensor network for real-time acoustic scene analysis," *Proc. IEEE International Symposium on Circuits and Systems*, pp. 169–172, 2009.
[21] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins, "A multichannel fusion framework for audio event detection," *Proc. 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015.
[22] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," *Proc. 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1271–1275, 2015.
[23] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using distributed microphone array for acoustic scene analysis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1335–1343, June 2017.
[24] A. Ribeiro, A. G. Marques, and S. Segarra, "Graph signal processing: Fundamentals and applications to diffusion processes," *Proc. 24th European Signal Processing Conference (EUSIPCO)*, 2016.
[25] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1996.
[26] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 85–92, 2017.
[27] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.
[28] K. Imoto and N. Ono, "Acoustic scene classification based on generative model of acoustic spatial words for distributed microphone array," *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 2343–2347, 2017.
[29] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-features acoustic event detection for sensor networks," *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 55–59, September 2016.