Acoustic Scene Classification Based on Generative Model of Acoustic Spatial Words for Distributed Microphone Array

Keisuke Imoto*† and Nobutaka Ono‡*
*SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
†Ritsumeikan University, Shiga, Japan
‡National Institute of Informatics, Tokyo, Japan

***Abstract*—In this paper, we propose an acoustic scene classification method for a distributed microphone array based on a combination of spatial information of multiple sound events. In the proposed method, each acoustic scene is characterized by a spatial information representation based on a bag-of-words called the bag of acoustic spatial words. To calculate the bag-of-acoustic spatial words, spatial features extracted from multichannel observations are quantized and then aggregated over a sound clip, that is, each sound clip is regarded as a unit of a"document." Moreover, a supervised generative model relating acoustic scenes and bag-of-acoustic spatial words is also adapted, which enables robust acoustic scene classification. Experimental results using actual environmental sounds show that the proposed approach achieves more effective performance than the conventional acoustic scene classification approach not utilizing a combination of the spatial information of multiple sound events.**

I. Introduction

The classification of acoustic scenes (*cooking, vacuuming, watching TV, being on the bus, meeting*) or acoustic events (*footsteps, running water, voice*) has recently become important for many applications such as monitoring elderly people [1], [2], automatic surveillance [3]–[5], automatic classification of life-logging [6], [7], and multimedia retrieval [8]–[10].

To analyze acoustic scenes or acoustic events, many effective methods based on machine learning techniques have been proposed. For instance, Eronen *et al.* [6] and Mesaros *et al.* [11] have proposed methods based on the mel-frequency cepstral coefficients (MFCCs) for spectral feature extraction and hidden Markov models (HMM) for acoustic scene or event analysis. Cauchi *et al.* [12] have proposed an acoustic scene classification method that utilizes spectral bases captured by non-negative matrix factorization (NMF). As other methods of acoustic scene classifications, we can focus on the fact that acoustic scenes are characterized not by a single sound event but by a combination of multiple sound events. For instance, an acoustic scene "*cooking*" is characterized by a combination of multiple sound events including "*running water*," "*cutting ingredients*," and "*heating a skillet*." On the basis of this idea, Guo and Li [13], Kim *et al.* [14], and Imoto and coworkers [7], [15] proposed acoustic scene classification methods based on the bag-of-acoustic words, which quantize the spectral features into acoustic words and aggregates acoustic words into a histogram of them.

Meanwhile, the location of a sound source varies from acoustic scene to acoustic scene. For example, when considering an acoustic scene "cooking," sound sources are often in a kitchen, and when considering another acoustic scene "eating," sound sources are often in a dining area. Therefore, acoustic scenes can be characterized not only by using spectral information but also by using spatial information. Using many acoustic sensors simultaneously, such as smartphones, IoT devices, and surveillance cameras, spatial information can be extracted, and some researchers have proposed methods utilizing spatial information for acoustic scene classification or acoustic event detection [16]–[19]. One of the fundamental ways of extracting spatial information from a microphone array is to use a single-position information or the single direction of arrival (DOA) of a sound source based on sound source localization or DOA estimation. However, to apply these methods, the position information of microphones is needed and the microphones require synchronizing, and therefore, it is not easy to use them for a distributed microphone array. To address this problem, we previously proposed a method for extracting spatial information that does not need the position information of the microphone array and does not need precisely synchronized microphones if the microphones are positionally fixed [20].

On the other hand, acoustic scenes can be also characterized by a combination of sound events in the spatial approach as well as the spectral approach. For example, an acoustic scene "*cooking*" is characterized by a combination of sound events including "*running water from a faucet*," "*cutting ingredients on a cutting board*," and "*heating a skillet on a range*." Therefore, in this paper, as an acoustic scene classification method for a distributed microphone array that can consider a combination of the spatial information of multiple sound events in a long-term sound, we propose a bag-of-words-based approach for representing spatial information and an acoustic scene classification method based on a generative model of the bag-of-words-based spatial representation.

The remainder of this paper is structured as follows. In Section 2, we describe our proposed spatial feature extraction method based on the bag-of-words and a generative-model-based acoustic scene classification method. In Section 3, we present the results of acoustic scene classification experiments,

ISBN 978-0-9928626-7-1 © EURASIP 2017

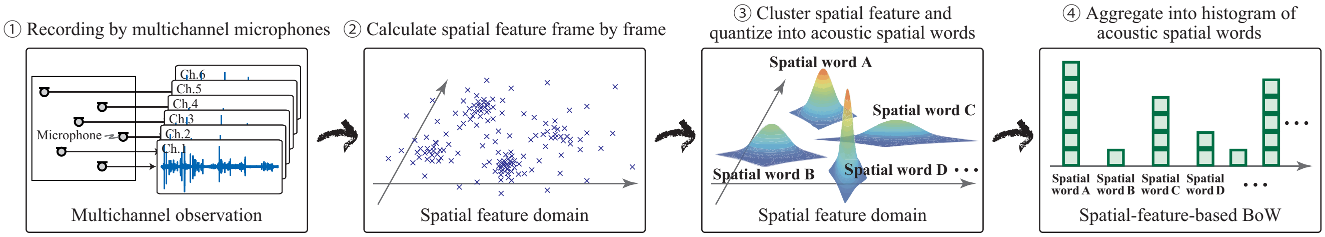

Fig. 1: Overview of spatial-feature-based BoW representation

and Section 4 concludes this paper.

II. Proposed spatial-information-based acoustic scene classification

A. Motivation and strategy of proposed method

In many situations, acoustic scenes are characterized by a combination of spatial information of multiple sound events in a long-term sound. To extract the combination of spatial information of multiple sound events, we focus on a histogram of the spatial information of sound events in short time frames during the long-term sound, which is based on the bag-of-words (BoW) representation [21], [22]. The BoW representation is a simple and effective means of representing acoustic scenes; however, it still has redundancy because there is partiality in the histogram of spatial information from acoustic scene to acoustic scene, for example; an acoustic scene “cooking” mostly occurs in the kitchen, that is, the histogram of spatial information has a sparse structure. Therefore, in this paper, we also apply a generative Bayesian model of spatial information for modeling and classifying acoustic scenes, which can reduce the redundancy of the spatial feature representation in acoustic scenes.

B. Spatial-feature-based BoW representation for multichannel observations

The BoW [21], [22], which characterizes the content of a document as a word histogram, is a simple feature representation of the document but an effective means of analyzing its content. Focusing on this idea, some researchers have employed the BoW representation in other research fields such as computer vision (bag-of-visual words) or acoustics (bag-of-acoustic words [7], [14], [19], bag-of-angle words [23]). Specifically, the bag-of-acoustic words is a discrete feature representation that quantizes the spectral features of sounds into acoustic words and aggregates acoustic words into a histogram of them. The bag-of-angle words is a discrete feature representation of the DOA of sound sources and it is also a simple means of representing spatial information. However, to utilize the bag-of-angle words, the DOA of sound sources must be estimated preliminarily, which premises that microphone positions are known. Therefore, it is not easy to apply the bag-of-angle words to acoustic scene classification using a distributed microphone array because in many cases, a distributed microphone array is used without location information.

Therefore, we here introduce a BoW representation that can be applied to any spatial feature of a distributed microphone array. Figure 1 shows an overview of the spatial-feature-based BoW representation. To calculate the spatial-feature-based BoW, spatial features are extracted from multichannel observations frame by frame and are quantized into acoustic spatial words. The acoustic spatial words in each sound clip are then aggregated into the BoW, that is, each sound clip is regarded as a unit of a “document.”

TABLE I: Definitions of symbols

Symbol	Definition
S	# of acoustic spatial word sequences (# sound clips)
A	# of classes of acoustic scenes
T	# of classes of acoustic spatial topics
M	# of classes of acoustic spatial words
$N_{\boldsymbol{w}_s}$	# acoustic spatial words in acoustic spatial word sequence $\boldsymbol{w}_s$
s	Index of acoustic spatial word sequence
a	Class index of acoustic scene
t	Class index of acoustic topic
m	Class index of acoustic spatial word
i	Order index of acoustic spatial word in each acoustic spatial word sequence
$\mathcal{W}$	Set of acoustic spatial word sequence
$\mathbf{a}_s$	Possible acoustic scenes in spatial word sequence s
a_s	Acoustic scene in spatial word sequence s
$\boldsymbol{z}$	Acoustic spatial topics
$\boldsymbol{w}_s$	sth spatial word sequence
$\boldsymbol{\theta}_a$	Acoustic spatial topic distribution of acoustic scene a
$\theta_{a,t}$	Occurrence probability of acoustic spatial topic t in acoustic scene a
$\boldsymbol{\phi}_t$	Acoustic spatial word distribution of acoustic spatial topic t
$\phi_{t,m}$	Occurrence probability of acoustic spatial word m in acoustic topic t
α, β	Hyperparameters for Dirichlet distribution
n_t^a, n_m^t	# of acoustic spatial words assigned to acoustic spatial topic t in acoustic scene a, etc.
$\backslash s, i$	Exclude ith acoustic spatial word in $\boldsymbol{w}_s$

C. Generative model of BoW for acoustic scene classification

There is partiality in the distribution of the spatial-feature-based BoW, that is, the spatial-feature-based BoW has a sparse structure. Therefore, we also apply a generative Bayesian model of the spatial-feature-based BoW for modeling and classifying acoustic scenes, which enables the sparse modeling of acoustic scenes. Such a generative model has been proposed for the bag-of-acoustic words, which is called the supervised acoustic topic model (sATM) [7], and thus, we adapt this model for the spatial-feature-based BoW.

In this model, the process generating acoustic spatial words can be represented by a hierarchical process including the

ISBN 978-0-9928626-7-1 © EURASIP 2017

acoustic scenes, acoustic spatial words, and acoustic spatial topics, where an acoustic spatial topic represents the latent structure appearing in acoustic spatial words. Additionally, to explicitly model the relationship between acoustic scenes and the BoW and to utilize the proposed model for acoustic scene classification, we here propose a supervised generative model of acoustic spatial words.

The specific generative process is as follows. As preprocessing of the generative model, a continuous acoustic signal of, for example, 1h length, is divided into sound clips of 10s length. Each sound clip is then represented by a sequence of acoustic spatial words (bag of acoustic spatial words), which is converted from an acoustic signal into an acoustic spatial word frame by frame. As the generative model, we assume that possible acoustic scene labels are preliminarily given to each acoustic spatial word sequence explicitly, and that an acoustic scene is generated randomly from them in its generative process. Then, we assume that each acoustic scene has a different acoustic spatial topic distribution $\boldsymbol{\theta}_a$, and an acoustic spatial topic is then generated from its distribution frame by frame. After that, an acoustic spatial word is generated from the distribution of acoustic spatial words $\boldsymbol{\phi}_t$ frame by frame, which depends on the acoustic topic. Note that these distributions $\boldsymbol{\theta}_a$ and $\boldsymbol{\phi}_t$ have Dirichlet priors, that is, hyperparameters α and β control the sparseness of the acoustic spatial topic and spatial word distributions, with which we can avoid overfitting of the model to the given data. Thus, the generative process of the acoustic spatial words is represented as follows, where the symbols used in this paper are defined in Table I.

```
A set of possible acoustic scenes $\mathbf{a}_s$ is given,
for $a = 1$ to $A$ do
    Choose $\boldsymbol{\theta}_a$                          $\sim$ Dirichlet($\alpha$)
end for
for $t = 1$ to $T$ do
    Choose $\boldsymbol{\phi}_t$                          $\sim$ Dirichlet($\beta$)
end for
for $s = 1$ to $S$ do
    Choose $a_s$                           $\sim$ Uniform($\mathbf{a}_s$)
    for $i = 1$ to $N_{\boldsymbol{e}_s}$ do
        Choose $z_{s,i} \mid \boldsymbol{\theta}_{a_s}, a_s$        $\sim$ Categorical($\boldsymbol{\theta}_{a_s}$)
        Choose $w_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}}, z_{s,i}$      $\sim$ Categorical($\boldsymbol{\phi}_{z_{s,i}}$)
    end for
end for
```

Additionally, the generative probability of all acoustic spatial words $\mathcal{W}$ can be represented as follows.

$$\begin{aligned}
&p(\mathcal{W}|\alpha, \beta, \gamma, \mathbf{a}_s) \\
&= \prod_{s=1}^{S} \prod_{i=1}^{N_{\boldsymbol{w}_s}} \sum_{a=1}^{A} \sum_{t=1}^{T} \sum_{m=1}^{M} p(w_{s,i} = m | z_{s,i} = t, \alpha, \beta, a_s = a) \\
&\quad \cdot p(z_{s,i} = t | a_s = a, \alpha) p(a_s = a | \mathbf{a}_s) \\
&= \frac{1}{A} \prod_{s=1}^{S} \Bigg[\int \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \\
&\quad \cdot \prod_{i=1}^{N_{\boldsymbol{w}_s}} \Bigg\{ \prod_{t=1}^{T} \theta_{a,t}^{\alpha-1+n_t^a} \int \frac{\Gamma(M\beta)}{\Gamma(\beta)^M} \prod_{m=1}^{M} \phi_{t,m}^{\beta-1+n_m^t} d\boldsymbol{\phi}_t \Bigg\} d\boldsymbol{\theta}_a \Bigg] \quad (1)
\end{aligned}$$

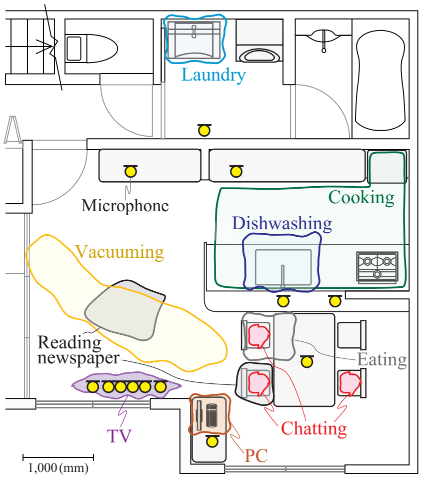

Fig. 2: Microphone arrangements and locations of sound sources in each acoustic scene

To estimate the model parameters, we here introduce Bayesian inference based on collapsed Gibbs sampling (CGS) [24]. CGS iteratively samples latent variables corresponding to acoustic scenes and spatial topics in accordance with the conditional posterior probability of given acoustic spatial words as follows, which do not include updated acoustic scenes and spatial topics, respectively.

$$p(z_{s,i}|\boldsymbol{w}, \boldsymbol{z}_{\backslash s,i}, \boldsymbol{a}, \alpha, \beta) \propto (n^a_{(\backslash s,i),t} + \alpha) \cdot \frac{n^t_{(\backslash s,i),m} + \beta}{n^t_{(\backslash s,i),\cdot} + M\beta} \quad (2)$$

$$p(a_s|\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \alpha, \beta) \propto \tfrac{n^a_{(\backslash s),t} + \alpha}{n^a_{(\backslash s),\cdot} + T\alpha} \quad (3)$$

This sampling is repeated until the iterative update converges, then the posterior distributions of the acoustic spatial topics and spatial words are calculated from the estimated latent variables.

When classifying acoustic scenes using the proposed model, we first estimate the distributions of acoustic spatial topics $\boldsymbol{\theta}_a$ and acoustic spatial words $\boldsymbol{\phi}_t$ using a training dataset, and then we estimate an acoustic scene in test data by selecting the acoustic scene with highest posterior probability as follows.

$$\arg\max_{a} p(a|\boldsymbol{\theta}_a, \boldsymbol{\phi}_t, \boldsymbol{w}_s, \alpha, \beta) \quad (4)$$

D. Spatial and spectral integrated bag-of-word representation and integrated generative model

Considering the resemblance of the spatial-feature-based BoW and the bag-of-acoustic words, we can introduce an integrated BoW representation utilizing the spatial and spectral information as well as a generative model of the integrated BoW representation.

To calculate the integrated BoW representation, concatenated vectors of the spatial and spectral features are quantized into acoustic spatial and spectral words frame by frame, then they are aggregated into the BoW. In a generative model of the BoW and the acoustic scene classification using the model, we can introduce posterior probabilities of acoustic scenes by replacing acoustic spatial word sequence $\boldsymbol{w}_s$ with the acoustic spatial-spectral word sequence in Section II-C.

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE II: Typical sounds in each acoustic scene

Acoustic scene	Typical sounds
Vacuuming	whine of cleaner, footsteps
Cooking	cutting, sizzling, running water, clattering dishes
Dishwashing	running water, clattering dishes
Eating	clattering dishes, voices, coughing
Newspaper	flipping newspaper, footsteps
PC	clicking mouse, clacking keyboard, fan noise
Chatting	voices, coughing
TV	voices, music, sound effects, cheering
Laundry	running water, rinsing sound, notification sound

TABLE III: Experimental conditions

# of distributed microphones	13
Sampling rate/Quantization bit rate	48 kHz/16 bits
Reverberation time of living room	0.31 s
Average SNR	25.2 dB
Sound clip length	8 s
Frame length/FFT points	20 ms/2,048
# of frequency bins (GFSC)	8

III. Experiments

A. Environmental sound recordings

To evaluate the scene classification performance of the proposed method, we conducted an experiment using an actual environmental sound dataset. Nine acoustic scenes that frequently occur in a living room were chosen and the sound dataset was recorded using 13 synchronized microphones as shown in Fig. 2. In Fig. 2, the locations of sound sources related to each acoustic scene are also shown. Each acoustic scene typically included acoustic events listed in Table II. The sound dataset has 257.1 min of sounds and it was separated into 7,712 clips of the sounds, where none of the acoustic scenes overlapped with each other in all the sound clips. The other recording conditions and experimental conditions are listed in Table III.

B. Spatial-feature-based BoW calculation and acoustic scene classification

Figure 3 shows the acoustic scene classification process using the spatial-feature-based BoW representation and the generative model of acoustic spatial words. To extract the spatial features, we utilized the spatial cepstrum (SC) proposed in [20]. Similarly, the generalized-frequency spatial cepstrum (GFSC) is used for the integrated feature including spatial and spectral information [20]. The SC and GFSC can extract the spatial information efficiently and robustly without using locations of the microphones. Specifically, the SC is calculated by principal component analysis (PCA) of the channel-based log-amplitude vector

$$\mathbf{q}_\tau = \begin{pmatrix} \log b_{\tau,1} \\ \log b_{\tau,2} \\ \vdots \\ \log b_{\tau,n} \\ \vdots \\ \log b_{\tau,N} \end{pmatrix}, \tag{5}$$

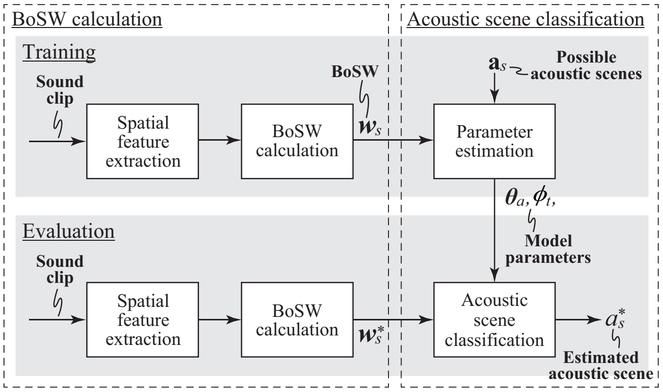

Fig. 3: Process of acoustic scene classification with spatial-feature-based BoW and sASTM

TABLE IV: Acoustic feature, number of feature dimensions, method of acoustic scene modeling, and average estimation accuracy in proposed and conventional methods

Acoustic feature	Feature dimension	Scene modeling	Average F-score
MFCCs	12	GMM	42.4%
SC	13	GMM	46.8%
GFSC	25	GMM	50.4%
BoW (MFCCs)	512	GMM	38.6%
BoW (SC)	512	GMM	47.4%
BoW (GFSC)	512	GMM	57.7%
BoW (MFCCs)	512	sATM	52.7%
BoW (SC)	512	sATM	53.0%
BoW (GFSC)	512	sATM	**64.3%**
BoW (MFCCs)	512	Classifier stacking [19]	45.1%

where τ, n, and $b_{\tau,n}$ are the time frame index, microphone channel index, and multichannel power observation at each time frame. After calculating the SC and GFSC, they were quantized by using a Gaussian mixture model (GMM), and a spatial-feature-based BoW was calculated sound clip by sound clip. Here, the SC and GFSC were classified by using the GMM in an unsupervised manner, and then each Gaussian component was defined as a single acoustic spatial word or acoustic spatial-spectral word. In the acoustic scene classification, the parameters of the proposed generative models were estimated using acoustic scene labels and BoWs in the training dataset. Then, acoustic scenes of the test dataset were classified by the maximum a posteriori (MAP) estimation.

C. Comparative approaches for acoustic scene classification

To compare the acoustic scene classification performance, we evaluated a conventional GMM-based approach, which extract acoustic features and calculate likelihoods for acoustic scenes using GMM frame by frame, and then, product the likelihoods over the number of frames in each sound clip. For this approach, we utilize the SC and GFSC as the spatial feature. We also evaluated other methods of acoustic scene classification utilizing spectral information. In these methods, we first aggregated acoustic signals recorded by multichannel microphones to a central node and averaged them over the channels. We then extracted MFCCs as spectral acoustic features frame by frame or the bag-of-acoustic words sound clip by sound clip. As the acoustic scene classifiers, the GMM and the supervised acoustic topic model (sATM) [7] were used. As another method for acoustic scene classification utilizing the distributed microphone array, we also evaluated a classifier based on the late fusion-based classification method [19]. In this method, the bag-of-acoustic words was first extracted in each channel and acoustic scenes were classified channel by channel. Then, the acoustic scene classifier was learned by a combination strategy from the training data [19].

D. Experimental results

Table IV shows the classification accuracy of acoustic scenes in terms of the average F-score. These results indicate that the spatial information extracted by the spatial-feature-based BoW representation enables acoustic scenes to be classified effectively as well as when using the bag-of-acoustic words. Moreover, the proposed approach achieves more effective performance than the conventional approaches not utilizing combinations of the spatial information of multiple observations. Additionally, the BoW representation combining spectral information and spatial information has a higher performance than that utilizing either spectral or spatial information. Thus, when using the BoW (GFSC) and the generative model of acoustic spatial and spectral words, the classification accuracy achieved its highest performance (64.3%).

IV. Conclusion

We proposed a spatial-information-based method for acoustic scene analysis, which utilizes a generative model of acoustic spatial words. In the proposed method, each acoustic scene is characterized by the spatial-feature-based BoW representation. To calculate the spatial-feature-based BoW, spatial features extracted from multichannel observations are quantized and then aggregated over a sound clip, that is, each sound clip is regarded as a unit of a "document." Moreover, a supervised generative model relating acoustic scenes and spatial-feature-based BoW is also proposed, which enables robust acoustic scene classification. Experimental results conducted using real-life environmental sounds indicated that the proposed method is more efficient for acoustic scene classification than the conventional acoustic scene classification approach utilizing only single-spatial information or spectral information. Additionally, experimental results also indicated that the integrated BoW representation of spectral information and spatial information has a higher performance than that utilizing either spectral or spatial information.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (A) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 16H01735).

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] 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.

[5] 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*, pp. 158–161, 2005.

[6] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio, Speech, Language Process.*, pp. 321–329, 2006.

[7] 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, 2016.

[8] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," *Proc. INTERSPEECH*, 2012.

[9] T. Zhang and C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 9, no. 4, pp. 441–457, 2001.

[10] 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.

[11] 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.

[12] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," *Master's Thesis, ATIAM, Université Pierve et Marie Curie*, 2011.

[13] G. Guo and S. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, pp. 209–215, 2003.

[14] 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*, pp. 37–40, 2009.

[15] 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.

[16] 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. 18th European Signal Processing Conference* (*EUSIPCO*), pp. 1271–1275, 2015.

[17] 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*, pp. 1–5, 2015.

[18] 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.

[19] 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.

[20] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using distributed microphone array for acoustic scene analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1335–1343, 2017.

[21] Z. S. Harris, "Distributional structure," *Word*, vol. 10, pp. 146–162, 1954.

[22] T. Joachims, "Learning to classify text using support vector machines: Methods, theory, and algorithms," *J. Comput. Linguist.*, vol. 29, pp. 655–664, 2003.

[23] K. Ishiguro, K. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic speaker diarization with bag-of-words representations of speaker angle information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 447–460, 2012.

[24] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," *Department of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1*, 1993.