World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:13, No:12, 2019

Slice Bispectrogram Analysis-Based Classification of Environmental Sounds Using Convolutional Neural Network

Katsumi Hirata

***Abstract*—**Certain systems can function well only if they recognize the sound environment as humans do. In this research, we focus on sound classification by adopting a convolutional neural network and aim to develop a method that automatically classifies various environmental sounds. Although the neural network is a powerful technique, the performance depends on the type of input data. Therefore, we propose an approach via a slice bispectrogram, which is a third-order spectrogram and is a slice version of the amplitude for the short-time bispectrum. This paper explains the slice bispectrogram and discusses the effectiveness of the derived method by evaluating the experimental results using the ESC-50 sound dataset. As a result, the proposed scheme gives high accuracy and stability. Furthermore, some relationship between the accuracy and non-Gaussianity of sound signals was confirmed.

***Keywords*—**Bispectrum, convolutional neural network, environmental sound, slice bispectrogram, spectrogram.

I. Introduction

THE development of systems with excellent functions to recognize the sound environment as humans do is an important problem. These systems include self-driving cars, autonomous robots, and the systems which support the hearing impaired. In this research, we focus on sound classification and aim to develop a procedure that automatically classifies various environmental sounds. Researchers have proposed various methods for the classification of environmental sounds [1]-[3]. One common approach is the use of a convolutional neural network, which is a deep learning technique [4]. It is essential to create input data that best describe the characteristics of the original data. This is because the performance of a method based on neural networks depends on it [5]. Time-frequency analysis, such as the spectrogram, is effective for time-varying sounds. Although the power spectrum used in the spectrogram cannot describe higher-order statistics, higher-order spectra allow us to analyze higher-order signal components and to extract useful characteristics. Therefore, we propose a mechanism that uses a third-order spectrogram based on bispectral analysis. In this paper, we explain the slice bispectrogram (SBS) and discuss the effectiveness of the method by evaluating the experimental results using the ESC-50 sound dataset [6], [7].

K. Hirata is with the Department of Innovative Electrical and Electronic Engineering, National Institute of Technology (KOSEN), Oyama College, Oyama, Tochigi 323-0806 Japan (e-mail: hirata@oyama-ct.ac.jp).

II. Bispectrogram

A. Bispectral Analysis

A power spectrum $P(k)$ of a random signal $x(n)$ is estimated from the expectation of the second-order periodogram using the Wiener–Khinchin theorem,

$$P(k) = E[|X(k)|^2], \tag{1}$$

where $X(k)$ is the Fourier spectrum of $x(n)$, and $E[\,]$ denotes the expectation. Similarly, the estimate of a bispectrum $B(k_1, k_2)$ is obtained as the following equation because the bispectrum is defined by the Fourier transform of the third-order auto-correlation function [8]:

$$B(k_1, k_2) = E[X(k_1)X(k_2)X^*(k_1 + k_2)] \tag{2}$$

where $X^*(k)$ denotes the complex conjugate of $X(k)$. It represents the physical linear dependency among the components on the three frequencies k_1, k_2 and $k_1 + k_2$.

B. SBS

Certain kinds of environmental sounds are time varying; they are nonstationary. For such signals, it is not suitable to apply the power spectral and bispectral analyses mentioned above because the spectra represent the averaged spectral density during the analyzed interval. We use the following short-time Fourier transform which is effective for nonstationary signals:

$$X(m, k) = \sum_{n=0}^{N-1} w(n - m)x(n) \exp(j2\pi kn/N), \tag{3}$$

where $w(n - m)$ is a window centered on the time $m + n/2$. Therefore, this equation represents a spectrum around each time $m + n/2$. Applying $X(m, k)$ to (1) instead of $X(k)$, we obtain the short-time power spectrum, i.e. *spectrogram*, as:

$$S_2(m, k) = E[|X(m, k)|^2]. \tag{4}$$

We refer to this spectrogram as the *power spectrogram* (PS) to distinguish this from another spectrogram mentioned subsequently. Similarly, applying (3) to (2), we obtain the short-time bispectrum as:

$$B(m, k_1, k_2) = E[X(m, k_1)X(m, k_2)X^*(m, k_1 + k_2)], \tag{5}$$

But it cannot be displayed as a two-dimensional image. This is the reason why we use the SBS

$$S_3(m,k) = |E[X(m,k)X(m,k)X^*(m,2k)]|, \quad (6)$$

which is the slice version of the amplitude for the short-time bispectrum at frequency $k_1 = k_2 = k$ [9], [10]. Fig. 1 illustrates how an SBS is obtained from the bispectra around each time frame. The SBS is formed in such a manner that the diagonal components of each bispectrum are arranged in the shape of a strip.

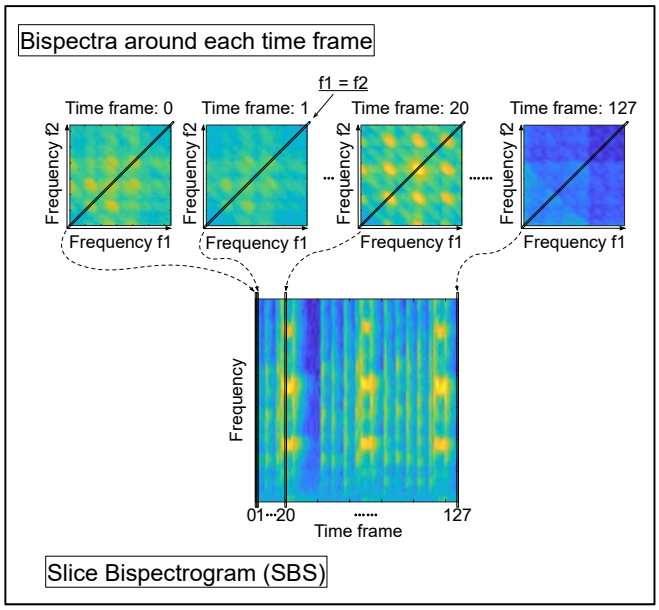

Fig. 1 SBS

III. Convolutional Neural Network

Our proposed method uses a convolutional neural network to classify environmental sounds. The layer structure of the neural network used in our experiments is illustrated in Fig. 2. The network consists of five convolutional layers, three pooling layers, and a fully connected layer. Each convolutional layer is followed by normalizing and applying the ReLU function to extract its characteristics effectively; a dropout operation is applied before the fully connected layer. After training the network using a large amount of spectrogram 2-D images generated from labeled environmental sounds, the images from the unknown class of sounds are input to the first convolutional layer, and a class ID corresponding to the sound signal is given as the output.

IV. Classification Experiments

We conducted experiments using either a PS or an SBS as the input for the neural network. We also discuss the effectiveness of our SBS method by comparing and evaluating the classification accuracies.

A. Conditions

An environmental sound dataset ESC-50 was employed as both the training and the test data. The dataset contained 2000 labeled sound excerpts of environmental sounds from 50 classes. The experiments were implemented using MATLAB. The main conditions of the experiment are shown in Table I. Every sound signal in the dataset was randomly allocated to one of the five folders. In the experiments, we conducted a 5-fold cross validation as follows: The network was trained on data from four of the five folders and tested on data from the remaining folders. This process was repeated five times (each time adopting a different set of four out of the five folders for the training and using the remaining set for testing). Subsequently, the classification performance was evaluated using average accuracies and standard deviations over all five tests for PS method and SBS method.

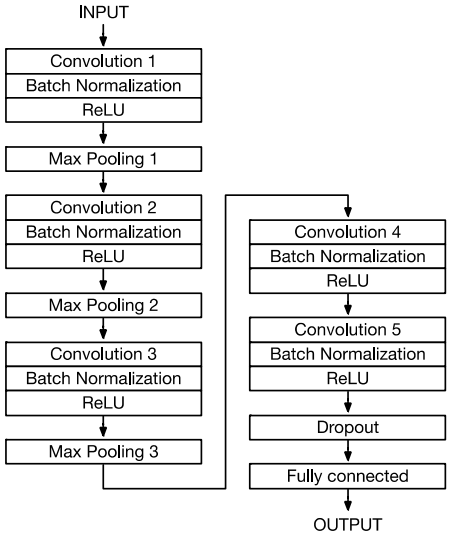

Fig. 2 Structure of a convolutional neural network

TABLE I
Conditions of the Experiment

Sampling frequency	8 kHz
Duration of each signal	5 sec
Number of points (pixels) of each spectrogram image	Time: 128 points Frequency: 128 points
Convolution layer 1	Filter size: 3×3, Number of filters = 12
Convolution layer 2	Filter size: 3×3, Number of filters = 24
Convolution layers 3, 4, 5	Filter size: 3×3, Number of filters = 48
Pooling layers 1, 2, 3	Size: 6×6
Dropout	70 %
Fully connected layer	50 output

B. Results

Examples of SBS images for each class of sound are shown in Fig. 3, where numbers under every image indicate the ID of class shown in Table II. The SBS images will be able to be used as input data to classify environmental sound signal because it seems that every SBS has a unique pattern. The accuracies of classification for each sound class and the methods PS, SBS are presented in Table II. Values in the "Accuracy" columns indicate what percentage of classifications was correct for each class of sounds. The SBS method was performed classifications with greater accuracy than the conventional PS for more than half of the classes (highlighted in yellow).

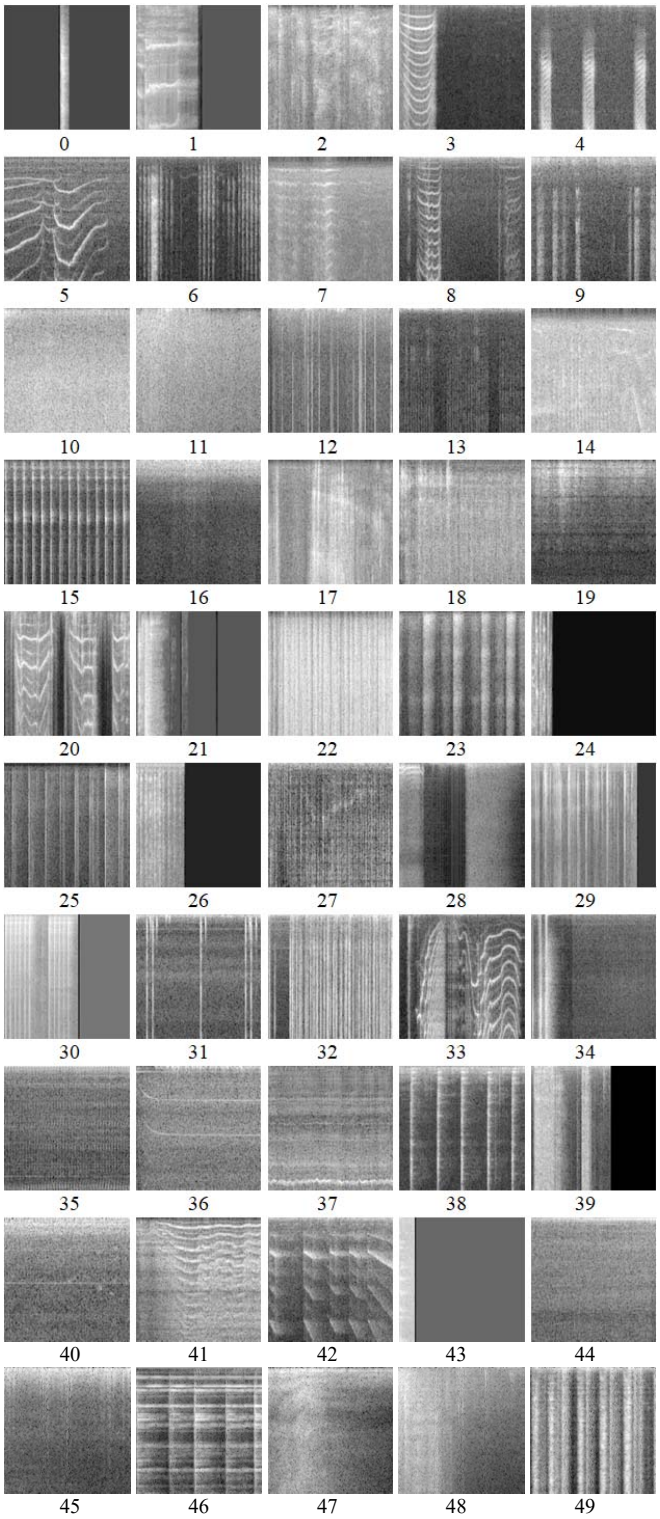

Fig. 3 Examples of SBS patterns

In particular, SBS method can obtain better accuracy for most of the classes in categories of animals, interior/domestic and exterior/urban sounds. The averaged accuracies and their standard deviations for each method are shown at the bottom of Table II. The average accuracies for both methods reached about 50%; the SBS method was achieved at a slightly higher accuracy and was more stable than the PS method.

TABLE II
ACCURACIES OF CLASSIFICATIONS FOR EACH CLASS

ID	Category	Class	Accuracy [%]	
			PS	SBS
0	Animals	Dog	45.0	52.5
1		Rooster	70.0	75.0
2		Pig	37.5	45.0
3		Cow	52.5	57.5
4		Frog	45.0	50.0
5		Cat	47.5	50.0
6		Hen	35.0	45.0
7		Insects (flying)	42.5	60.0
8	Natural soundscapes & water sounds	Sheep	45.0	40.0
9		Crow	60.0	40.0
10		Rain	50.0	37.5
11		Sea waves	45.0	65.0
12		Crackling fire	57.5	47.5
13		Crickets	20.0	17.5
14		Chirping birds	20.0	25.0
15		Water drops	22.5	20.0
16		Wind	50.0	35.0
17		Pouring water	60.0	50.0
18		Toilet flush	87.5	80.0
19		Thunderstorm	85.0	80.0
20	Human, non-speech sounds	Crying baby	57.5	77.5
21		Sneezing	62.5	70.0
22		Clapping	40.0	40.0
23		Breathing	27.5	25.0
24		Coughing	60.0	57.5
25		Footsteps	45.0	32.5
26		Laughing	37.5	35.0
27	Interior / domestic sounds	Brushing teeth	47.5	55.0
28		Snoring	70.0	57.5
29		Drinking, sipping	25.0	47.5
30		Door knock	60.0	75.0
31		Mouse click	30.0	25.0
32		Keyboard typing	67.5	67.5
33		Door, wood creaks	12.5	20.0
34		Can opening	47.5	62.5
35		Washing machine	22.5	27.5
36		Vacuum cleaner	47.5	57.5
37		Clock alarm	45.0	60.0
38		Clock tick	62.5	62.5
39		Glass breaking	47.5	47.5
40	Exterior / urban noises	Helicopter	17.5	30.0
41		Chainsaw	65.0	67.5
42		Siren	42.5	62.5
43		Car horn	37.5	55.0
44		Engine	25.0	30.0
45		Train	32.5	60.0
46		Church bells	90.0	95.0
47		Airplane	17.5	30.0
48		Fireworks	37.5	40.0
49		Hand saw	70.0	65.0
		Total average	**46.6 %**	**50.2 %**
		Standard deviation	**18.6 %**	**17.7 %**

The SBS is based on third-order statistics. Therefore, it can

represent the non-Gaussian characteristics that the conventional PS cannot describe. To clarify whether using SBS is effective for non-Gaussian signals, we examined the non-Gaussianity of signals and the effects of classification accuracy. Skewness is a measure of the asymmetry of a signal about the mean, and is calculated as:

$$s_x = \frac{E[\{x-m_x\}^3]}{\sigma_x^3}, \quad (7)$$

where x is the signal, m_x is its mean, and σ_x is its standard deviation. The skewness of the Gaussian signal is 0, and its absolute goes higher for more non-Gaussian signals. Then the relationship between the accuracies of SBS method and the averaged skewness was examined. Fig. 4 shows the relative accuracies of SBS method against PS method and the averaged skewness for every class. The relative accuracy trends correspond to the skewness, except for some classes.

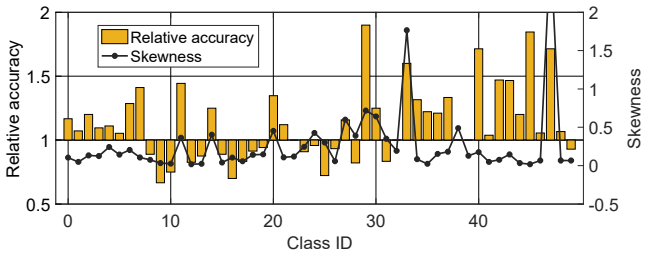

Fig. 3 Relationship between relative accuracies using SBS and the averaged skewness

Although the SBS had some effectiveness for classifying the non-Gaussian sounds, it was not effective despite its higher skewness values, such as ID21, ID25, ID28 or ID38. On the other hand, although the skewness of ID8, ID40, ID42 or ID45 is less, the classification accuracy of the SBS method was higher than PS method. There are some exceptions in the relationship between the relative accuracy of the SBS method and the skewness of the signal. However, selective use or combination of the SBS and PS may improve the classification accuracy. We will need to investigate this in our future studies.

V. Conclusion

Through this study aiming to realize stable and highly accurate environmental sound classification using convolutional neural networks, the following results were obtained:

- A method which uses SBS instead of the commonly used spectrogram as input data for convolutional neural network was proposed.
- Through fundamental experiments using ESC-50 data set, it was confirmed that the SBS method can classify with higher accuracy and stability than the PS method.
- The relationship between the relative accuracy of the SBS method and the skewness that represents the non-Gaussianity of the signal was evaluated, and it was made clear that there was some relationship.

Since higher-order spectral analysis is not easily affected by Gaussian noise, SBS is expected to be robust to the noise. We propose to further study the effectiveness of SBS in clarifying the properties of objects.

References

[1] S. Chu, S. Narayanan and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, 17-6, pp.1142-1158, Aug. 2009.

[2] F. Su, L. Yang, T. Lu and G. Wang, "Environmental sound classification for scene recognition using local discriminant bases and HMM," *Proceedings of the 19th ACM international conference on Multimedia*, pp.1389-1392, Nov. 2011.

[3] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp.1-9, Oct. 2013.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," *2015 IEEE international workshop on machine learning for signal processing*, Sept. 2015.

[5] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *ArXiv Prepr. ArXiv170607156*, 2017.

[6] "ESC-50: Dataset for Environmental Sound Classification", https://github.com/karoldvl/ESC-50 (Last accessed at Oct. 3, 2019).

[7] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp.1015-1018, Oct. 2015.

[8] C. L. Nikias and A. P. Petropulu, *Higher-order spectra analysis: a nonlinear signal processing framework*, Prentice Hall, 1993, pp.7-30

[9] V. Swarnkar, U. Abeyratne, and C. Hukins, "Objective measure of sleepiness and sleep latency via bispectrum analysis of EEG," *Medical and & biological engineering & computing*, 48, pp.1203-1213, Dec. 2010.

[10] K. Hirata, "Estimating 3D-Position of A Stationary Random Acoustic Source Using Bispectral Analysis of 4-Point Detected Signals," *International Journal of Computer and Information Engineering*, 8-6, pp.932-935, 2014.

Katsumi Hirata was born in Osaka, Japan, in 1975. He received the Ph.D. degree in engineering from University of Tsukuba in 2002. He is currently an associate professor at National Institute of Technology (KOSEN), Oyama College, Japan.

Dr. Hirata is a member of the IEEE, the Institute of Electronics, Informations and Communication Engineers of Japan and the Society of Instrument and Control Engineers of Japan.