Hindawi Publishing Corporation
Applied Computational Intelligence and Soft Computing
Volume 2012, Article ID 650818, 6 pages
doi:10.1155/2012/650818

Research Article

Environmental Sound Recognition Using Time-Frequency Intersection Patterns

Xuan Guo,[1] Yoshiyuki Toyoda,[1] Huankang Li,[2] Jie Huang,[1] Shuxue Ding,[1] and Yong Liu[1]

[1] *Graduate Department of Computer and Information Systems, Graduate School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu 965-8580, Japan*
[2] *Department of Computer Science and Engineering, Shanghai Jiaotong University, 200240 Shanghai, China*

Correspondence should be addressed to Jie Huang, j-huang@u-aizu.ac.jp

Received 13 January 2012; Accepted 27 February 2012

Academic Editor: Zhishun She

Environmental sound recognition is an important function of robots and intelligent computer systems. In this research, we use a multistage perceptron neural network system for environmental sound recognition. The input data is a combination of time-variance pattern of instantaneous powers and frequency-variance pattern with instantaneous spectrum at the power peak, referred to as a time-frequency intersection pattern. Spectra of many environmental sounds change more slowly than those of speech or voice, so the intersectional time-frequency pattern will preserve the major features of environmental sounds but with drastically reduced data requirements. Two experiments were conducted using an original database and an open database created by the RWCP project. The recognition rate for 20 kinds of environmental sounds was 92%. The recognition rate of the new method was about 12% higher than methods using only an instantaneous spectrum. The results are also comparable with HMM-based methods, although those methods need to treat the time variance of an input vector series with more complicated computations.

1. Introduction

Understanding environmental sounds is an essential function of human hearing. For example, people can recognize the beginning of a rain shower by the rain sound, be cautious when they hear footsteps coming from behind at night, and open the door to welcome visitors after the sound of the door-knocking. Environmental sound recognition is also important for intelligent robots and computer systems. An intelligent robot can be aware of the environments by the audition and use its hearing function to complement its vision [1].

In recent years, environmental sound recognition has received increasing attention, and we have seen some pioneering research in this field. An environmental sound database (RWCP-DB) has been created for research use [2]. The sounds in the database were recorded in an anechoic environment with durations of 250 to 500 ms. In total, there are 105 instances, with each instance including 100 samples. We reclassified this database into 12 types and 45 kinds as listed in Table 1. For many sounds, there are multiple instances with similar but different materials.

An environmental sound recognition method using the instantaneous spectrum at the power peak was proposed [3]. It was reported that the rate of recognition was about 80% for 20 instances of environmental sounds. In this research, the target sounds are limited to impact sounds that have a single power peak followed by exponential attenuation. The instantaneous spectrum $S_p(\omega_m)$ was calculated at the power peak, where ω_m $(m = 1, 2, \ldots, M)$ is the frequency. Since the input information was only based on the peak spectrum without time variance, it was not able to capture the environmental sounds and thus the recognition rate was low.

It is natural to consider using existing methods that have proven useful for speech recognition, for example, the hidden Markov Model (HMM) method and the time delay neural network (TDNN) method [4–6], since those methods deal with time variations of an input vector series. Miki and others achieved recognition rate of 95.4% using HMM method for 90 instances of RWCP-DB [5], and Sasou, and

Table 1: The RWCP environmental sound database.

Sound type	Kind of materials	Instances
Impact sound	Wood plates	12
	Metal cans, boxes, and so forth	10
	Plastic cases	3
	Glass cups, bottles, and so forth	8
	Bundle of paper	1
	Handclap/handclaps	4
Falling pieces	Grains	2
	Coin/coins	7
	Dice	3
Air jet	Small air pump	1
	Spray	1
	Firecracker	1
	Air bubbles	1
	Dryer	1
Friction sound	File	1
	Sand paper	2
	Saw	2
Musical instruments	Castanets	1
	Cymbals	1
	Drum	1
	Horn	1
	Kara	1
	Maracas	1
	Ring	1
	String	1
	Whistle	3
	Tambourine	1
Phone, buzzer	Buzzer	1
	Clock alarm	2
	Phone	4
	Toys	3
Open	Cap	2
Broken	Chopsticks	1
	Tearing paper	1
	Crumpling paper	1
Release	Clip	2
Shaking	Metal bell/bells	7
Rotation	Coffee mill	1
Others	Doorlock	1
	Leaf through a book	2
	Mech bell	1
	Padlock	1
	Punch	1
	Shaver	1
	Stapler	1

others reported the recognition rate for 59 instances of RWCP-DB using AR-HMM method was 83.0% [6].

The recognition rate of the HMM method was greater than that of the peak-spectrum method. Because the HMM method uses a time series of frequency-feature vectors $[S_n(\omega_m)]$ that includes the time-frequency variance of the signals, where ω_m $(m = 1, 2, \ldots, M)$ is the frequency and $S_n(\omega_m)$ indicates the spectrum (or cepstrum) for time frame n $(n = 1, 2, \ldots, N)$. However, HMM-based methods may not be the best choice for environmental sound recognition

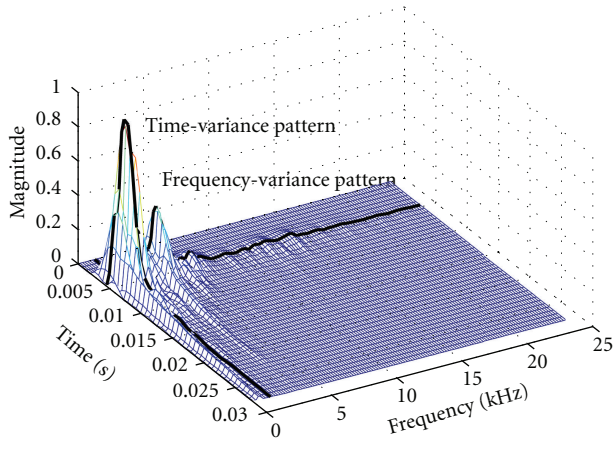

Figure 1: The time-frequency intersection pattern refers to the combination of the time-variance patter containing instantaneous powers (or their square roots) for all time frames and the frequency-variance patter with the instantaneous spectrum at power peak. (The time-variance pattern is illustrated as the line along with the spectrum-peaks.)

because environmental sounds differ from human speech. The frequency characteristics of most environmental sounds do not significantly change over time, and therefore it is not necessary to deal with state-transferring in many cases, as the HMM methods for speech signals require.

We can use a simpler method using the combination of a time-variance pattern containing the instantaneous powers (or their square roots) calculated by the sum-of-squares method for all time frames and a frequency-variance pattern with the instantaneous spectrum at the power peak as illustrated in Figure 1. Since this combination contains both time-variance and frequency variance of the signal, it incorporated almost the information needed for environmental sound recognition. We call this input data type a time frequency intersection pattern and refer to the time-variance patter of power as power-variance pattern. Thus, the information can be represented as $[S_p(\omega_m), P(t_n)]$, where ω_m $(m = 1, 2, \ldots, M)$ is the frequency, $S_p(\omega)$ indicates the spectrum at the time frame of power peak, and $P(t_n)$ indicates the power of sound for time frame t_n $(n = 1, 2, \ldots, N)$. The total information includes two vectors with sizes M and N (total $M + N$), which is less than that of HMM-based methods ($M \times N$ in total). This method can drastically reduce the input data while preserving the main time-frequency characteristics of environmental sounds.

We use perceptron NNs for environmental sound recognition. A multistage classification-recognition strategy is adopted to cover environment sounds with different time lengths. The first stage is the classification part, which classifies environmental sounds into three categories, single bursts, repeated sounds, and continuous sounds, based on their long-term power-variance patterns. The second stage is the recognition part, for individual recognition of each sound. In this stage, three different NN groups are used for different categories of environmental sounds. Two experiments were conducted using an original environment sound database recorded in an ordinary room and the RWCP database recorded in an anechoic chamber to verify the proposed new method.

2. Environmental Sound Database and Preprocessing

Since this research is concerned with a project that aims to develop a security patrol and home-helper robot capable of understanding environmental sounds, the target environmental sounds are chosen to be important for the robot to achieve its tasks. As seen in Table 3, 10 kinds of environmental sounds were selected and recorded in an ordinary room environment, with 30 samples of each kind. The original sampling frequency was 44.1 kHz.

For comparison with the previous methods, we selected 10 kinds of sounds and a total of 45 instances from the RWCP-DB as seen in Table 4.

Since there are unlimited kinds of environmental sounds, no database can cover all of them. Therefore, no system will be able to recognize all environmental sounds. Instead, for a practical system, the target sounds must be limited according to the practical environment and the purpose of tasks. That is, environmental sound recognition is task dependent.

At the preprocessing stage, the environmental sound data were downsampled to 8 kHz. The instantaneous power was calculated for each time frame of 128-point length. While the long-term power-variance patter contains the power data of 48 frames, the short-term power-variance patter is of 16 frames. The peak spectrum was calculated around power peak with a time frame of 64 points. All data were normalized to have a maximum value of one.

3. System Construction

In many cases, environmental sounds can be mainly classified into collision sounds, friction sounds, vibration sounds, electric sound, and other noises. Based on their power-variance patterns, environmental sounds can be roughly classified into single bursts, repeated bursts, continuous sounds, and other noises. It is reasonable to first classify the environmental sounds into different categories based on their long-term power-variance patterns in the classification stage. Recognition based on the combination of short-term power-variance patterns and frequency-variance patterns at the power peak will be performed in the second stage.

The data flow of the environmental sound recognition system is presented in Figure 2. The system consists of a classification part and a recognition part.

A three-layer perceptron NN is used for sound classification and recognition. The construction of the NN is described in Table 2.

3.1. Classification by Long-Term Power-Variance Patterns. The data needed for classification is the long-term power-variance patterns for each input sound. An example of the

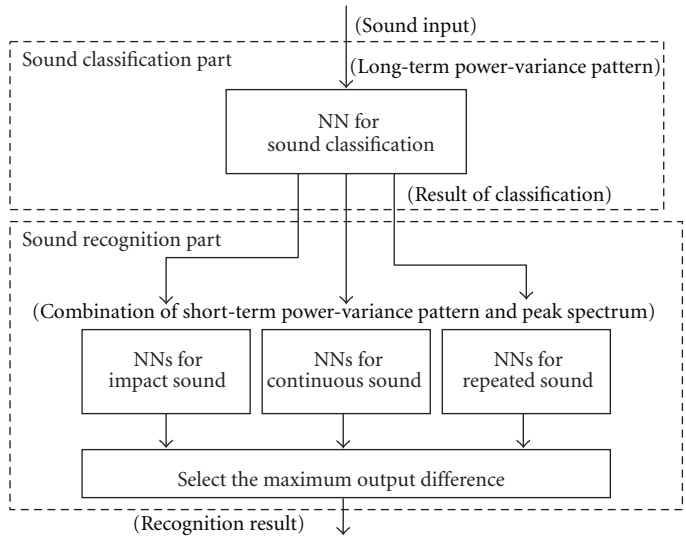

Figure 2: System data flow.

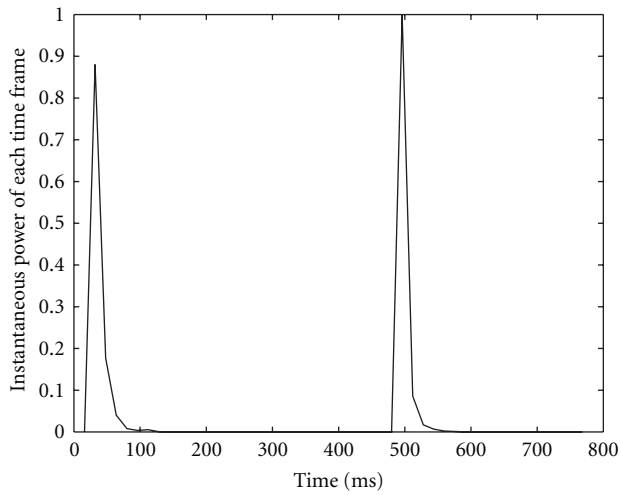

Figure 3: A sample of long-term power-variance pattern (a door-knocking sound).

Table 2: Construction of NNs for classification and recognition parts.

Input layer neuron	48
Intermid layer neuron	32
Output layer neuron	2

long-term power-variance pattern of a door-knocking sound is presented in Figure 3.

This classification stage classifies sounds with short impact sounds as single-impact sounds; sounds of friction, vibration, noises, and electric sounds like phone bells as continuous sounds; some sounds with repetition, for example, hand claps or knocks on a door, as repeated sounds.

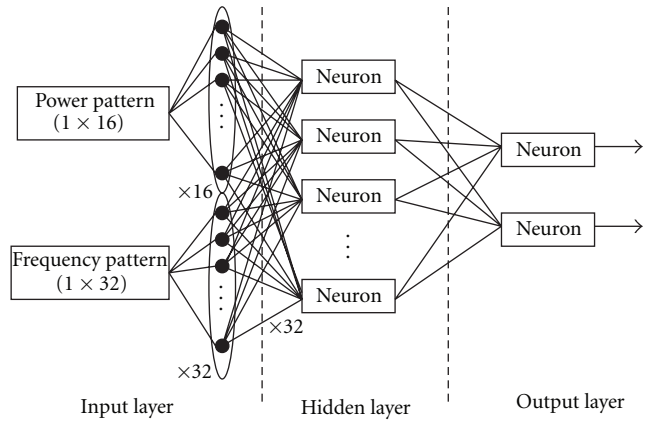

Figure 4: NN in the recognition part.

3.2. Construction of the Recognition Part. For almost all kinds of environmental sounds, the time variances of the frequency characteristics are usually rather stable and there are few marked changes during their period compared with speech sounds. The input data for the recognition part assigns the short-term power-variance pattern to the first 16 inputs and the instantaneous spectrum calculated at the power peak to the remaining 32 inputs, as seen in Figure 4. The output layer of each NN has two neurons that correspond to the results of correct and incorrect matching.

The three NNs in the recognition part correspond to the three target sound categories. Each NN, constructed by a three-layered perceptron, is trained for one target sound category. The final recognition result depends on the difference between the two output neurons of each NN. The NN that obtains the maximum difference of correct and incorrect output is dominant and gives the final recognition result (Figure 2).

Table 3: Results of recognition experiments for environmental sounds in the original database.

Sound kind	First stage rate	Final recognition rate
Boll impact	100%	100%
Metal impact	100%	95%
Door opening/closing	100%	85%
Lock	100%	95%
Switch on/off	100%	100%
Typing	100%	75%
Repeated typing	80%	80%
Knock	90%	90%
Telephone ringing	100%	100%
Japanese vowels	100%	100%
Average		92.0%

Table 4: Results of recognition experiments for environmental sounds in the RWCP database.

Sound kind	First stage rate	Final recognition rate
Wood impact	100%,	96.5%
Metal impact	99.5%,	92.5%
Clap	97.5%,	89.2%
Plastic impact	100%,	100%
Grains falling	100%,	80.0%
Telephone ringing	100%,	88.3%
Metal bell	99.2%,	98.3%
Spray	100%,	95.0%
Whistle	100%,	100%
Drier	100%,	86.0%
Average		92.7%

4. Recognition Experiments

Two experiments using the original prerecorded environmental sound database and the RWCP database were conducted. In all of the experiments, the computer system used was an MS-Windows PC with an Athlone 1600 XP CPU and 512 MB of memory. The NNs were implemented using the MATLAB programming language.

For the original database, 10 samples of each sound kind were used for NN training, and 10 samples of data were used for the recognition tests. The NN training time was about 1 hour in total, and the recognition time for each input data sample was less than 0.1 second. The results of the recognition are listed in Table 3. The average rate of recognition was 92.0%.

From the RWCP database, data for 10 kinds of sounds (total of 45 instances) were selected for the experiments. In the experiments, 10 samples of each sound kind were used for NN training and 20 samples were used for testing. Since there were not enough kinds of repeated sounds in this database, only single-impact and continuous sounds were tested. The required training time was 2 hours, and the recognition time for each data sample was less than 0.1 second. The recognition results are presented in Table 4. The average recognition rate was 92.7%.

5. Conclusion

In this research, we propose a multistage environmental sound recognition method. The method consists of a classification stage and a recognition stage. The classification stage classifies environmental sounds into three categories based on their long-term power-variance patterns, and the recognition stage recognizes the sound kind based on a combination of the short-term power-variance pattern and the instantaneous spectrum at the power peak.

The merit of this method is that it uses a one-dimensional intersectional time-frequency pattern that combines the power-variance pattern and the instantaneous spectrum at the power peak. The recognition rate of the new method was 12% higher than methods using only an instantaneous spectrum at the power peak. The results are also comparable with HMM-based methods, although those methods must accommodate the time variance of the input vector series with more complicated computations.

References

[1] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for rbotos: Sound localization and separation," *Artificial Life and Robotics*, vol. 1, no. 4, pp. 157–163, 1997.

[2] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 225–232, 1999.

[3] K. Hiyane and J. Iio, "Non-speech sound recognition with microphone array," in *Proceedings of the IEEE International Workshop Hands-Free Speech Communication*, 2001.

[4] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated world recognition," *Neural Networks*, vol. 3, pp. 23–43, 1990.

[5] K. Miki, T. Nishiura, S. Nakamura, and G. Kashino, "Environmental sound recognition by HMM," in *Proceedings of the Spring Meet of The Acoustical Society of Japan*, no. 1-8-8, 2000.

[6] A. Sasou and K. Tanaka, "Environmental sound recognition based on AR-HMM," in *Proceedings of the Autumn Meet of The Acoustical Society of Japan*, no. 3-Q-7, 2002.