Classification of environmental sounds based on image processing of spectrogram

47-166805 GONG ZHIHAO
Supervisor: Prof. Ken Sasaki

The purpose of this research is to combine the visual information of spectrograms with acoustic properties for providing a novel feature extraction method for environmental sounds. Starting with the "spectrogram reading" with the assistance of binarization, it can be found that regions with high acoustic energy distribute in different patterns, which can be considered as a factor for distinguishing sounds. Therefore, along with the acoustic energy level in those regions as processing reference, frequency components with their stability, periodicity and mixture ratios are extracted mostly as the features. Besides, due to the variety of expression forms, continuous sound is divided into two general categories, for which a hierarchical classification model is established to verify the effectiveness of the extracted features.

Key words: environmental sounds, spectrogram, image processing, feature extraction

1 Introduction

Sounds from environment contain abundant information beyond speech or music we normally hear in our daily life. Therefore, automatic recognition of audio events from the environmental sounds will be very useful for caring systems for elderly people, security systems, and mechanical fault diagnosis. However, many researches on feature extraction of environmental sounds are based on conventional methods for speech recognition, such as Mel-frequency Cepstral Coefficient (MFCC) and Linear Predictive Cepstral Coding (LPCC)[1,2]. These parameters are designed for speech sounds that have tonal and harmonic structures. Therefore, they are not necessarily suitable for environmental sounds that are noise-like and less structured compared to speech sounds.

Among several approaches of feature extraction, visualization based methods are gathering more eyes gradually. Time-frequency representation of sound, known as spectrogram, makes the sounds' characteristics more intuitive and inspires researchers to propose new features and their extraction methods. However, most of them only use the texture features of image such as Local Binary Pattern (LBP)[3]. This kind of methods put the distinctive properties of sounds in a black box so that essential benchmarks for classifying environment sounds are still unclear now.

In this research, we firstly summarize the distinctive properties of continuous sound through observation of spectrograms. Then through a series of novel methods that take into account the acoustic meaning, observed properties by human intuition are extracted into features. Finally, by analyzing the result of a hierarchical classification model, we attempt to verify the proposed features and attempt to explore the pivotal features for distinguishing environment.

2 Feature exploration based on spectrogram

Sound is composed of fluctuations with different frequency. These sound components also possess various amplitudes which represent different acoustic energy level in the frequency domain. The perception of different sounds depends largely on those components with high acoustic power. Through short-time Fourier Transform (STFT), we can get a plot with sound components in a time-frequency representation called spectrogram. Those components with high acoustic energy appear as blobs with bright color in the spectrogram, whose distribution is considered as features in the infancy stage of speech recognition. That's also the reason that many previous researches extract texture features of spectrograms. Continuous sounds, which have constant appearance of blobs in their spectrograms similar to speech, are chosen as the object in this research.

To extract those blobs with high acoustic power for further analysis, an adaptive binarization method is conducted on spectrograms. The result of water sound is shown in Fig.1.

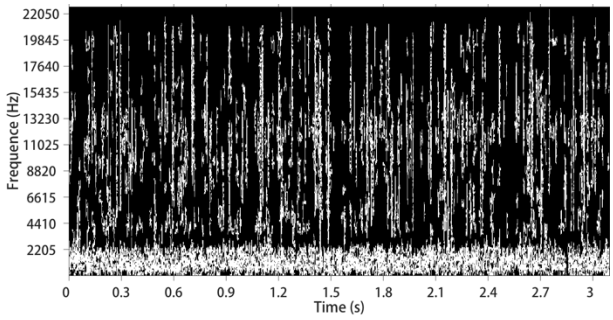

Fig.1 Binary spectrogram with adaptive threshold

After observing the binarized spectrograms of other continuous sounds, the distributions of the white regions that represent basic components of sound can be summarize into two descriptive factors:

1. dense parts in frequency domain.
2. their change mode with time.

However, spectrograms store three-dimensional information: time, frequency and energy (Power Spectral Density, PSD), which means the two factors above neglect the acoustic energy in the blobs. Therefore, in order to analyze the importance of energy

values in extracted non-zero regions and verify the correctness of application of binarization, inverse Short-time Fourier Transform (STFT) is used to generate sound signals only using the coefficients in non-zero positions in binary spectrograms. The process is shown in Fig.2.

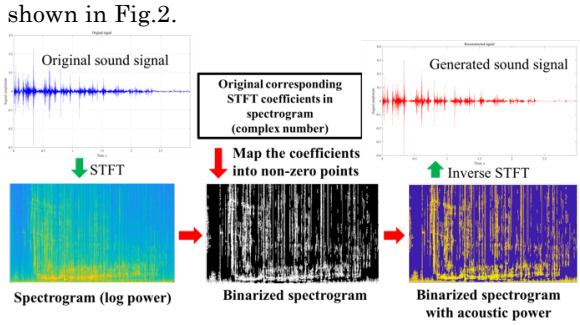

Fig.2 Flow chart of Invers STFT

Through listening to the generated sounds based on several kind of sounds, it can be found that there exists slight difference between original sound and sound generated by original coefficients in non-zero positions. Conversely, sounds generated by coefficients with the same value sound like meaningless robot or electric current. Hence binary spectrogram with grayscale (in Fig.3) which represents the normalized energy with [0,1] will be the original data for subsequent feature extraction.

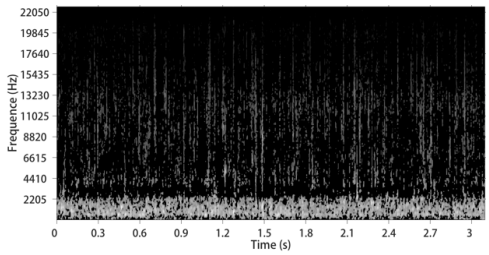

Fig.3 Binarized spectrogram with grayscale energy representation

By combining time-frequency and energy domain for spectrogram analysis, blocks with first level energy can be extracted by separating energy values into several levels. Information of these blocks should represent the most fundamental properties of environment sounds. Comparison of the graphs with these blocks from different sounds like Fig.4 tells us that continuous sound can be divided into two general categories because the significant distinction of the existence for sudden rising up. The general classification structure is show in Fig.5.

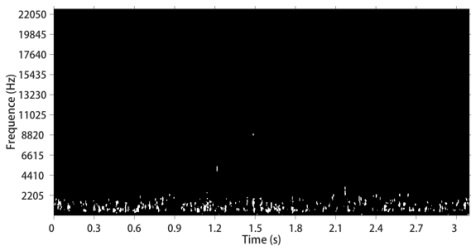

Fig.4 blocks with first level of water sound

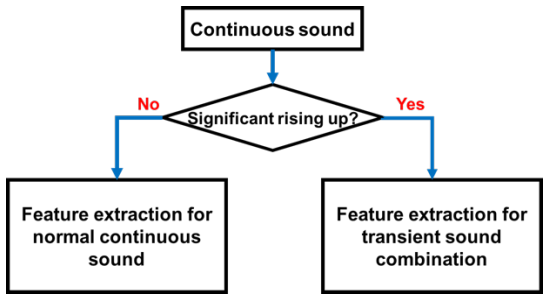

Fig.5 Proposed general classification structure

3 Feature extraction based on property analysis

3.1 General category feature

The simplest approach for detecting the sudden rising up for transient sound combination is summing up graph with first level blocks vertically. Peaks in the summed waveform for transient sound combination will be taller and sparser than those in normal continuous sound.

In order to extracted the two features above more precisely, I introduce a conditional peak detection method to extract peaks both with significant relative and absolute height. The summed waveform with detected peaks are shown in Fig6.

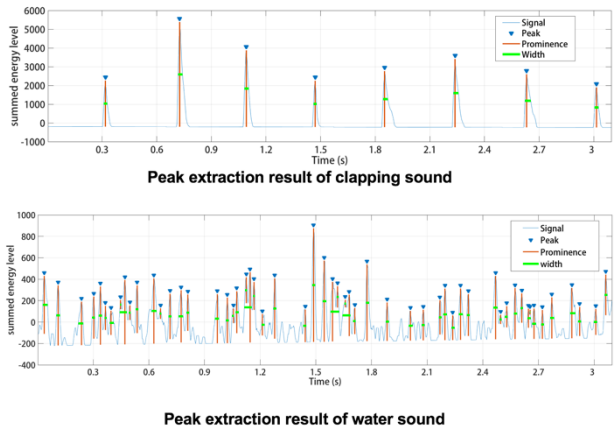

Fig.6 Summed waveform with conditional peak detection

To quantize mentioned features of height and sparse, I proposed sharpness (prominence/width) and time span as the feature for general category classification.

3.2 Features for normal continuous sound

According to the summarized two factors in section 2, by expanding them for normal continuous sound, we can list the features as:

1. 1. Existence grade of vertical stripe like structures for describing regularity in frequency range over 2500Hz.
2. 2. Periodicity for describing temporal transition.
3. 3. Feature group for describing frequency components, their stability, mixture ratios and flatness in frequency domain.

For the first feature, grayscale spectrogram processed by Gabor filter with 0° is segmented into horizontal frames for evaluating the continuity of

regions with high power in frequency domain. As shown in Fig.7, longer the vertical stripe is, the more peaks are overlapped in the summed waveform of the frames.

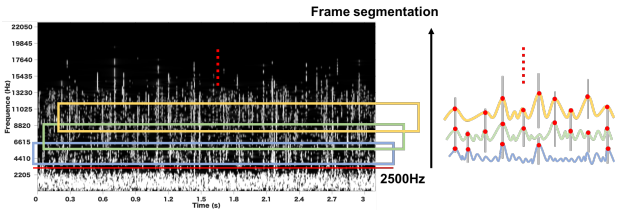

Fig.7 Process for evaluating vertical stripes

For the second and third feature, in order to capture those descriptors almost from human intuitions as precisely as possible, the refining spectrogram is proposed to represent the frequency components in several aspects: frequency belts with dense regions, their relative energy strength and flatness. The generation of this graph is arranging the frequency component detection results into time order. The conditional peak detection is also used for frequency components detection as shown in Fig.8.

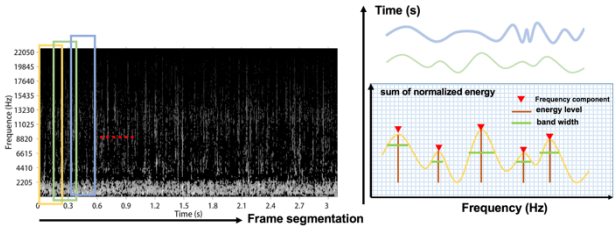

Fig.8 Process of generating refining spectrogram

An example of generated refining spectrogram is shown in Fig.9.

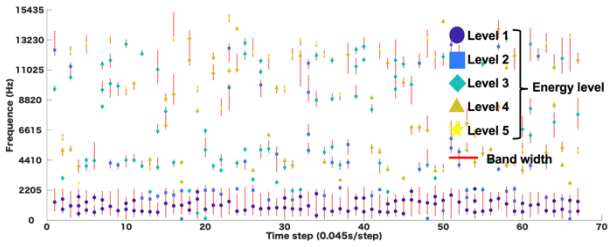

Fig.9 Refining spectrogram of water sound

Turning to the periodicity for normal continuous sound, the essential information is in the components with biggest energy values. As shown in Fig.10, by adapting Fast Fourier Transform (FFT) on the wave formed by those components, reasonable difference is revealed through peak number and coefficient of variation of the FFT result wave.

Then, for the descriptors: 1.average frequency position and their coefficient of variation, 2.average number of peaks that covered by their average frequency band width, are extracted separately by first level components and components with other levels. Especially for the components in first level, bias (Fig.11) is calculated for sound like wind and thunder with components with low average frequency and wide expansion to higher frequency bands. The additional feature for other level components are the relative ratios of energy with the first level.

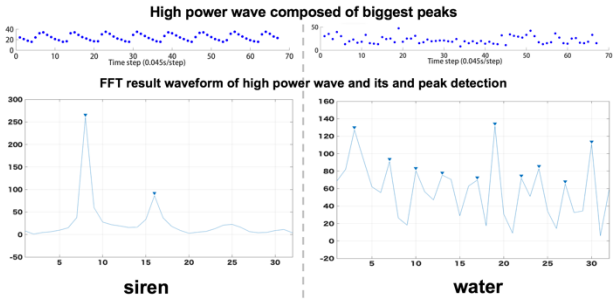

Fig.10 Periodicity feature extraction

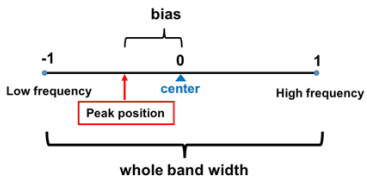

Fig.11 Definition of bias

Finally, features for normal continuous sounds can be arranged into a vector with 24 dimensions.

3.3 Features for transient sound combination

The first task for extracting features of combination of transient sounds is to divide the rising up sections out of spectrogram. Based on the characteristic that the covering range of frequency becomes significantly wider with the occurrence of transient sound, frequency centroid is used to cut the rising up sections out as in Fig.12

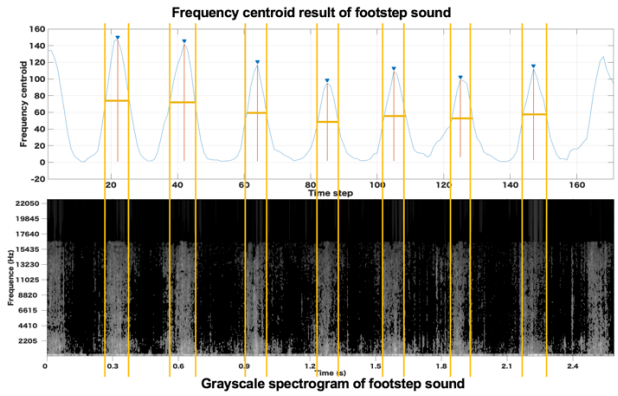

Fig.12 Extraction of the sections with rising up

Then similar to normal continuous sound, 1.average frequency position, 2.their band width and 3.ratios with first level are also extracted as features in a single section. Moreover, for transient sound analysis, the durations for different components could also be pivotal since the sound of knocking on glasses has clear formants with different lengths in time domain. The visualized extraction result is show in Fig.13

Features for transient sound combination can be arranged into a vector with 19 dimensions eventually.

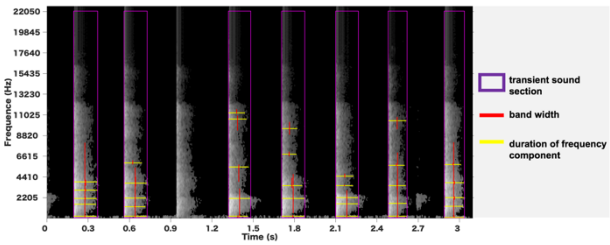

Fig.13 visualized feature extraction result for clapping

4 Recognition Experiments with proposed features

According to the feature extraction structure in Fig.5, a hierarchical classification model is established as shown in Fig.14.

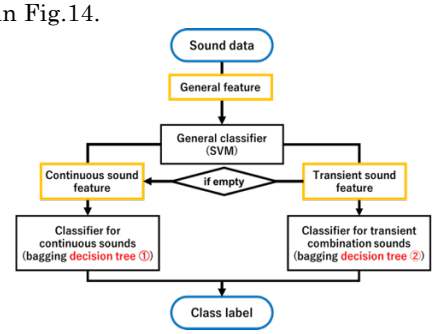

Fig.14 Proposed hierarchical classification model

General classifier is trained with the two-dimensional feature defined in 3.1. The scatter diagram of sharpness and time span with the classification boundary trained by Support Vector Machine (SVM) is shown in Fig.15:

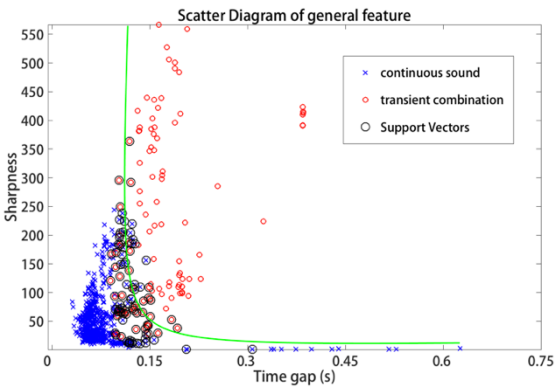

Fig.15 Scatter of general features and trained boundary by SVM

Decision tree ① and ② are trained with proposed features for the two general categories separately with accuracy of 91.1% and 95.4%. Comprehensive result with the proposed hierarchical classification model is 84.81% and summarized as a confusion matrix shown in Fig.16.

From the result it can be found though the recognition rate of rain is 0%, the testing data is mostly set into water sound and similar sound like clapping, which means the proposed features have good ability on describing basic properties of environment sound.

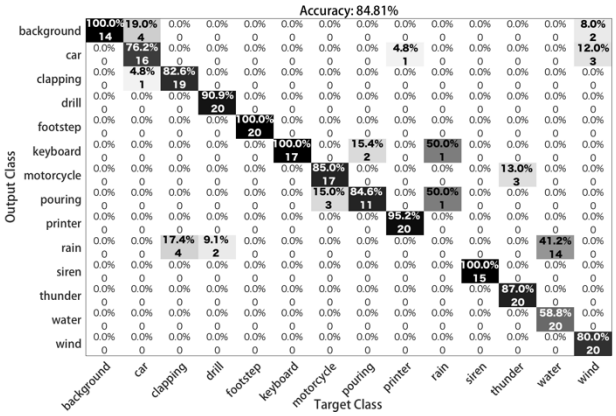

Fig.16 Confusion matrix of comprehensive classification result

Besides, the purpose of using decision tree is to seek the crucial benchmarks of sound distinguishing through exploring internal structure of the tree. By seeing the bifurcation points of the two decision trees, it can be said that the fundamental properties of environmental sound are mostly concentrate on the components with the first level power. But there isn't exist a clear path for a specific kind of sounds, which shows the variety of the expression form for environments sounds and necessity for features to capture characteristics more precisely.

5 Conclusion

Unlike pure texture features in previous researches, this approach provides a new perspective for capturing the graded perception of spectrogram by human intuition. The experiment results suggest that the proposed features possess good ability of capturing basic properties of environment sounds. But the more precise result would need detailed analysis for materials that cause vibration and the parameters used in experiments need optimization with contrast tests.

Reference

1) P. Guyot, "Water flow detection from a wearable device with a new feature, the spectral cover," *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, Annecy, 2012, pp. 1-6.

2) Ç. Okuyucu, "Audio Feature and Classifier Analysis for Efficient Recognition of Environmental Sounds," *2013 IEEE International Symposium on Multimedia*, Anaheim, CA, 2013, pp. 125-132.

3) K. Z. Thwe, "Environmental Sound Classification based on Time-frequency Representation," *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Kanazawa, pp. 251–255, 2017.