Diversity-Robust Acoustic Feature Signatures Based on Multiscale Fractal Dimension for Similarity Search of Environmental Sounds

journal or publication title	IEICE Transactions on Information and Systems
volume	104
number	10
page range	1734-1748
year	2021-10
URL	http://id.nii.ac.jp/1261/00000209/

doi: 10.1587/transinf.2021EDP7016

IEICE TRANS. INF. & SYST., VOL.E104–D, NO.10 OCTOBER 2021

PAPER

Diversity-Robust Acoustic Feature Signatures Based on Multiscale Fractal Dimension for Similarity Search of Environmental Sounds

Motohiro SUNOUCHI[†a)] ***and*** **Masaharu YOSHIOKA**[††b)]**,** ***Members***

SUMMARY This paper proposes new acoustic feature signatures based on the multiscale fractal dimension (MFD), which are robust against the diversity of environmental sounds, for the content-based similarity search. The diversity of sound sources and acoustic compositions is a typical feature of environmental sounds. Several acoustic features have been proposed for environmental sounds. Among them is the widely-used Mel-Frequency Cepstral Coefficients (MFCCs), which describes frequency-domain features. However, in addition to these features in the frequency domain, environmental sounds have other important features in the time domain with various time scales. In our previous paper, we proposed enhanced multiscale fractal dimension signature (EMFD) for environmental sounds. This paper extends EMFD by using the kernel density estimation method, which results in better performance of the similarity search tasks. Furthermore, it newly proposes another acoustic feature signature based on MFD, namely very-long-range multiscale fractal dimension signature (MFD-VL). The MFD-VL signature describes several features of the time-varying envelope for long periods of time. The MFD-VL signature has stability and robustness against background noise and small fluctuations in the parameters of sound sources, which are produced in field recordings. We discuss the effectiveness of these signatures in the similarity sound search by comparing with acoustic features proposed in the DCASE 2018 challenges. Due to the unique descriptiveness of our proposed signatures, we confirmed the signatures are effective when they are used with other acoustic features.

key words: *environmental sound analysis, fractals, content-based retrieval, feature extraction*

1. Introduction

Acoustic feature extraction is a basic audio signal processing issue. Acoustic features are important and necessary for various contexts and applications related to environmental sound recognition (ESR), such as large-scale content-based retrieval, auditory scene analysis, visualization, and event detection for surveillance. During the last decade, handy digital sound recorders have gained popularity, and at present, not only professional creators, but also amateurs have started recording environmental sounds and sharing them on web services such as Freesound [1], [2] and SoundCloud [3]. These sound recordings are not only appreciated as music works, but also sampled for creating sound effects, new music works, and live performances in music genres such as ambient, drone, and electronic [4], [5]. These sound recordings are also utilized for research to analyze and understand the variety of sound environments that we live in [6].

1.1 Applications Using Acoustic Features for Environmental Sounds

In recent years, the research on ESR for understanding a scene and its context has received considerable attention [7]. The workshop challenges on *Detection and Classification of Acoustic Scenes and Events* (DCASE) have demonstrated performance evaluations of systems for the detection and classification of sound events [8]. Based on the best result from Task 1B of DCASE2020, Koutini *et al.* evaluated their Receptive Field (RF) regularized CNN model with some parameter reduction methods [9].

Classification is a basic application that uses acoustic features. In 2003, Cowling and Sitte [10] presented a comprehensive comparative study of classification techniques that use various acoustic features for environmental sounds. They reported that the test patterns using each of the Mel-Frequency Cepstral Coefficients (MFCCs) and the continuous wavelet transform achieved the best recognition performance. In 2009 and 2012, Chu *et al.* [11] and Mogi *et al.* [12] reported that recognition systems that use the Matching-Pursuit-based acoustic feature as a time-domain feature shows better classification performance than systems that use the popular MFCCs only as a frequency-domain feature. In 2013, Bauge *et al.* [13] proposed a new acoustic feature for environmental sounds based on the scattering transform. This feature is robust against frequency transposition.

Content-based retrieval is another basic application that uses acoustic features. Web-based sound archives such as Freesound and SoundCloud are becoming popular and the amount of their sound content is increasing. The online users who utilize these sound archives can share and browse sound content by means of content-based retrieval. In 2008, Xue *et al.* [14] proposed a similarity search system, which employs a cluster-based indexing approach for environmental sounds. In 2010, Roma *et al.* [15] proposed a method for the retrieval of environmental sounds using the general sound-events taxonomy defined based on the principles of ecological acoustics. Chechik *et al.* [16] compared the scalability of several classification methods using MFCCs for a large-scale content-based sound retrieval. In 2013,

Manuscript received January 18, 2021.
Manuscript revised May 14, 2021.
Manuscript publicized July 2, 2021.

[†]The author is with the Design Department, Sapporo City University, Sapporo-shi, 005–0864 Japan.
[††]The author is with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: sunouchi@media.scu.ac.jp
b) E-mail: yoshioka@ist.hokudai.ac.jp

DOI: 10.1587/transinf.2021EDP7016

Sunouchi and Tanaka [17] proposed a new acoustic feature signature, namely, the enhanced multiscale fractal dimension signature (EMFD) and demonstrated the effectiveness of EMFD for content-based similarity search of environmental sounds.

In recent years, the workshop challenges on DCASE have focused on improving machine learning methods for ESR and produced high-performance results for their tasks. Acoustic features are still essential as input data for the machine learning methods for ESR. Hence finding new acoustic features that can properly describe the features of environmental sounds is fundamental in improving the performance of these ESR applications. In addition, by studying how acoustic features can describe the features of environmental sounds and affect the performance of ESR tasks, we can develop an understanding of how we are listening to environmental sounds.

1.2 Acoustic Feature Extraction for ESR

The environmental sounds outside a recording studio are produced by action and movement. We can identify things by listening to their acoustic properties, which are the results of the sound production process. However, environmental sound signals of the same type cannot be physically identical to each other due to the difference in their production processes. Furthermore, the different sound signals generated by simultaneous events are mixed with each other, which makes the properties of each sound source obscured [18].

Various acoustic features have been proposed for content-based audio retrieval. The feature selection is an important process for ESR [19], [20]. Cepstral features that include MFCCs and their first and second derivatives (MFCCsΔ and MFCCs$\Delta\Delta$) are widely used as frequency-domain acoustic features. MP-based acoustic feature has been proposed as one of the useful time-domain features for ESR [11], [12], [21].

Recent researches have focused on the evaluation of time-domain features of environmental sounds. For ESR, we need acoustic features that describe the non-stationary characteristics of target sounds as a time-domain feature and are robust against the diversity of environmental sounds [7]. We have recognized there may be three main causes of the diversity of environmental sounds.

D1) Small fluctuations of sound source parameters, such as carrier signal frequency, due to the individuality of the sound source.
D2) Background noises that the person who recorded the target sound did not expect to record.
D3) Mixed composition of different types of sound sources.

For the third cause D3, it is necessary to apply, for example, independent component analysis or non-negative matrix factorization to the sound signal before the feature extraction process [12], [22]. In this study, we focus on the extraction of new acoustic feature signatures that are robust against the diversities caused by D1 and D2.

1.3 Problems of EMFD Signature and Their Solutions

In our previous work [17], we proposed an EMFD signature that can describe both the frequency-domain features and time-domain features of target sounds. The EMFD signature is a feature vector, which consists of the time-varying multiscale fractal dimension (MFD) values. We demonstrated that EMFD improves the performance of similarity search by supplementing MFCCs. Unfortunately, it is found that EMFD includes error values that depend on the number of analysis windows and the histogram's bin size used for computing its histogram. Furthermore, EMFD seems to be oversensitive while discriminating the features of environmental sounds, and may lack robustness against the diversity of environmental sounds.

In this study, we extend the EMFD signature by improving the process of computing its histogram using the kernel density estimation method. By optimizing the bandwidth parameter used for kernel density estimation, the histogram of the enhanced multiscale fractal dimension using kernel density estimation signature (EMFD-KDE) becomes sufficiently smooth and robust against the diversity of environmental sounds as an acoustic feature signature. In Sect. 2, we present the basic theory and characteristics of EMFD. In Sect. 3, we propose a method to compute the EMFD-KDE signature. In Sects. 5 and 6, we demonstrate that EMFD-KDE improves the performance of the similarity search system.

Furthermore, we enhance the idea of EMFD and propose a new acoustic feature signature, namely very-long-range multiscale fractal dimension signature (MFD-VL). The environmental sounds have important acoustic features over a long time period. However, EMFD cannot describe the time-domain feature for time periods longer than 10 ms. In Sect. 4, we propose a method to compute the MFD-VL signature. In addition, we demonstrate that MFD-VL can describe the features of the time varying envelope for long periods of time, and that it has the robustness against the diversity causes D1 and D2.

In Sect. 7, we conclude that the proposed feature signatures of EMFD-KDE and MFD-VL solve the problems of EMFD and are effective when they are used with other acoustic features, including MFCCs and acoustic features proposed in the DCASE 2018 challenges.

2. Basic Theory of Enhanced Multiscale Fractal Dimension Signature

Mandelbrot, who advocated a concept of fractal in 1975 for the first time, demonstrated that some structures in nature could be modeled well by the theory of fractals [23]. One of the most important characteristics of fractals is that they have self-similarity properties at multiple scales. In the field of acoustics, Voss and Clarke analyzed the power spectrum of fluctuating physical variables including frequency, loud-

ness and pitch in music and speech [24]. They obtained the $1/f^{\gamma}$ ($0.5 \lesssim \gamma \lesssim 1.5$) aspects in the power spectrum of each variable against the frequency of a signal passed through a low-pass filter having a range 0 Hz – 1 Hz. Hsu [25] compared the fractal geometry of classical music works, and found that there is a relation, defined by the theory of fractals, between the interval of successive notes and their frequency of occurrence.

2.1 Multiscale Fractal Dimension

A fractal dimension is an index value that can describe the characteristics of a fractal by quantifying their complexity as a ratio of the change in detail to the change in scale. Acoustic features based on the fractal dimension have been proposed and utilized for various practical applications in the fields such as acoustics, music analysis, image analysis, physics, physiology, and neuroscience. Maragos *et al.* [26], [27] proposed the short-time fractal dimension of speech signals as an acoustic feature and used it for speech segmentation and sound classification. Zlatintsi and Maragos [28], [29] proposed a multiscale fractal dimension (MFD) profile as a short-time descriptor and found that this descriptor can discriminate several aspects among different musical instruments.

2.2 Steps to Compute the EMFD Signature

In our previous work [17], we developed EMFD as a feature signature of environmental sounds for a similarity search system. The EMFD is computed as follows.

2.2.1 Preprocessing Target Sounds

The maximum amplitude of each target sound that is to be analyzed must be first normalized to −0.1 db. They are converted to the standard format with the following specifications: sampling rate of 44.1 kHz and bit depth of 16 bits.

2.2.2 Computing the Area of Minkowski Sausage

The fractal dimension of a sound signal can be computed based on the Minkowski-Bouligand dimension. A covering area can be drawn by moving a unit disk of radius r along the curve of the waveform. This covering area is called a Minkowski Sausage. The center of the unit disk should be at any position on the curve of waveform and the width of Minkowski Sausage becomes $2r$. Figure 1 shows the Minkowski Sausage obtained by moving the unit disk along the waveform. To compute the area of the Minkowski Sausage of a discrete sound signal, the unit disk vector $C(r)$ is defined as Eq. (1), where r denotes the radius of the unit disk and i denotes the discrete position on the horizon. Figure 2 shows how the model of the unit disk is built. The vertical distance from the center to the top of unit disk at each horizontal position is denoted by the unit disk vector $C(r)$. Let n be the sampling position, r the radius of the unit

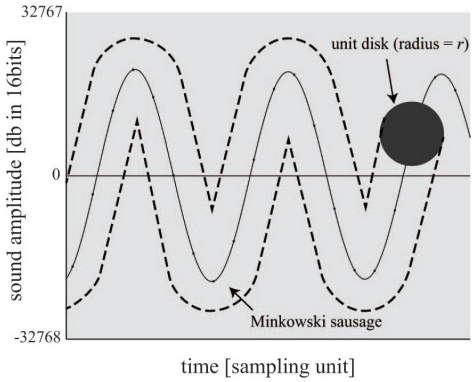

Fig. 1 A sound waveform and a Minkowski sausage

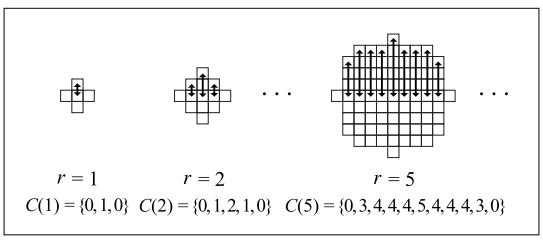

Fig. 2 Mesh-Approximation of a unit disk

disk, p the discrete position of the unit disk, and $\mathrm{sig}(x)$ the amplitude of sound signal at each sampling position x. The area of the Minkowski Sausage $area(n, r)$ at each sampling position n is computed as Eq. (2).

$$C(r) = \left\{ \mathrm{floor}\left(\sqrt{2ri - i^2} \right) \,\middle|\, 0 \le i \le 2r, i \in \mathbb{Z} \right\} \tag{1}$$

$$\begin{aligned} \mathrm{area}(n, r) = &\max_{\substack{0 \le p \le 2r \\ p \in \mathbb{Z}}} \left(\mathrm{sig}\,(n - r + p) + \mathrm{floor}\left(\sqrt{2rp - p^2} \right) \right) \\ &- \min_{\substack{0 \le p \le 2r \\ p \in \mathbb{Z}}} \left(\mathrm{sig}\,(n - r + p) - \mathrm{floor}\left(\sqrt{2rp - p^2} \right) \right) \end{aligned} \tag{2}$$

2.2.3 Definition of Multiscale Fractal Dimension

The MFD values are computed for each analysis window whose period is 50ms. Let $A(r)$ be the area of the Minkowski Sausage drawn by the unit disk of radius r in each analysis window. The MFD of each analysis window is defined by Eq. (3). The minimum radius (r = 1) corresponds to the sampling period of the signal (1/44.1 ms) and the range of r from 1 to 132 corresponds to the range of the time scales from 1/44.1 to 3 ms.

$$MFD = \left\{ 2 - \frac{\log\left(A(r+1)/A(r)\right)}{\log\left((r+1)/r\right)} \,\middle|\, 1 \le r \le 132,\ r \in \mathbb{Z} \right\} \tag{3}$$

2.2.4 Definition of the EMFD Signature

In our previous work [17], we found that MFD has informa-

$$MFD_{enhanced}(x) = 2 - \frac{\log\left(A(r(x+1))/A(r(x))\right)}{\log\left(r(x+1)/r(x)\right)}, \text{where } r(x) = \text{round}(1.4^x) \tag{4}$$

$$AW(sound,\ period) = \left\{0, period, \ldots, \text{floor}\left(\frac{\text{the length of } sound}{period} - 1\right) \times period\right\} \tag{5}$$

$$FAW(dbin,\ rbin) = \{\ t \ \mid\ t \in AW(sound,\ 50),\ 1 + (dbin - 1)/32 \le MFD_{enhanced}(rbin) \text{ of analysis window } t < 1 + dbin/32\} \tag{6}$$

$$EMFD(sound) = \left\{\frac{\text{card}\left(FAW(dbin, rbin)\right)}{\text{card}\left(AW(sound,\ 50)\right)} \;\middle|\; 1 \le dbin \le 32,\ dbin \in \mathbb{Z},\ 1 \le rbin \le 16,\ rbin \in \mathbb{Z}\right\} \tag{7}$$

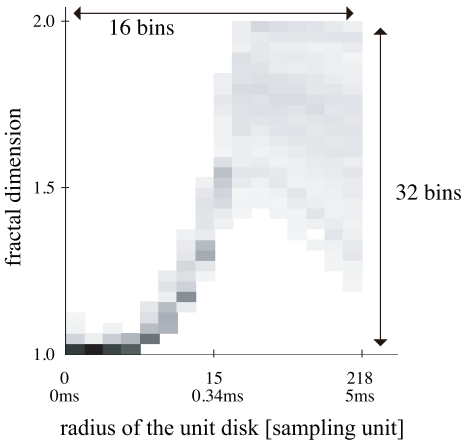

Fig. 3 Visualization of the EMFD signature of a cuckoo sound S_{cuckoo}

tive values for unit disks larger than the disk with a radius of 3 ms ($r = 132$). The maximum radius of the unit disk was extended to 218, which corresponds to 5 ms (1/10 of the period of analysis window), and the discrete values of the unit disk were modified to have exponential values. The enhanced MFD value at the x-th discrete value of the unit disk is defined as Eq. (4). The enhanced MFD values are computed for each analysis window. The EMFD signature is then defined as the two-dimensional histogram (16×32) of the time-varying enhanced MFD. Let *period* be the period (ms) of the analysis window and *sound* be the target sound. The set of analysis windows of the target sound is defined as Eq. (5). Let *rbin* be the bins that correspond to a series of 16 numbers used to define the different radius of the unit disk for computing the enhanced MFD, and *dbin* be the bins that correspond to a series of the 32 small intervals into which the range of fractal dimension is divided. The set of analysis windows whose enhanced MFD values fall into the bin (*dbin*, *rbin*) is defined by Eq. (6). The set of values in each bin of the EMFD histogram is defined by Eq. (7). Figure 3 shows the histogram that visualizes the EMFD signature of a cuckoo sound S_{cuckoo}. The length of S_{cuckoo} is 21.08s.

2.3 Known Characteristics of the EMFD Signature

Zlatintsi and Maragos [29] concluded that the MFD profiles are useful for quantifying the multiscale complexity and fragmentation of the different states of the instrument sound waveforms. In our previous work [17], we confirmed that EMFD describes the frequency-domain features and several other effective features of environmental sounds that MFCCs cannot describe. Furthermore, we confirmed that the EMFD signature has robustness against changes in volume levels and phase shifting of sound signals in the analysis window.

2.4 Problems of the EMFD Signature

The EMFD includes error values that depend on the number of analysis windows and the histogram's bin size, which is defined for computing its histogram, as shown in Eq. (7). In Sect. 3, we extend the existing EMFD by employing a kernel density estimation method to solve this problem.

Another problem of EMFD is that it cannot describe time-domain features for a time period longer than 10 ms, although environmental sounds have important acoustic features over a long time period. To solve this problem, we introduce MFD-VL signature in Sect. 4 as a newly developed time-domain acoustic feature.

3. Extending EMFD Employing a Kernel Density Estimation Method

As mentioned in Sect. 2.2, EMFD is computed as the two-dimensional histogram of time-varying enhanced MFD values. Let $NFAW(bin)$ be the number of analysis windows whose enhanced MFD values fall into the bin, and $NAW(sound)$ be the total number of analysis windows of the target sound *sound*, as defined in Eq. (8). The EMFD value of each bin $EMFD(sound,\ bin)$ is computed as Eq. (9). This method has the following two problems.

$$NAW(sound) = \text{card}(AW(sound,\ 50)) \tag{8}$$

$$EMFD(sound,\ bin) = \frac{NFAW(bin)}{NAW(sound)} \tag{9}$$

The first problem is that the value of each bin necessarily includes an error, because the value can be one of the discrete values given by the density of analysis windows. In particular, a lower number of analysis windows for the target sound increases the errors. Ideally, the EMFD histogram should be a continuous probability distribution of the time-varying enhanced MFD values, regardless of the number of analysis windows.

The second problem is that EMFD computed by the existing method is oversensitive to discriminate the features of environmental sounds. The tones and frequencies of each environmental sound may often vary, depending on the recording conditions and individual characteristics of the sources that generate this sound, even if person try to record the same type of environmental sounds in the same way. Therefore, a feature signature of environmental sounds should have robustness against the diversity of environmental sounds caused by D1 defined in Sect. 1.2.

To solve these problems, we introduce the kernel density estimation method to compute the EMFD histogram.

3.1 Definition of the EMFD-KDE Signature

The kernel density estimation method is employed to compute the probability distribution of the enhanced MFD values at each radius of the unit disk. The values of each bin of EMFD-KDE are defined as shown in Eq. (4), Eq. (8), Eq. (10), Eq. (11), and Eq. (12), where $K(\cdot)$ is the kernel function which is a Gaussian function, and h in Eq. (11) is the smoothing parameter called bandwidth.

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{10}$$

$$f_{emfd\text{-}kde}(dbin_{val}, rbin) = \frac{1}{NAW\,h} \times \sum^{NAW} K\left(\frac{dbin_{val} - MFD_{enhanced}(rbin)}{h}\right) \tag{11}$$

$$EMFD\text{-}KDE = \left\{ f_{emfd\text{-}kde}\left(1 + \frac{dbin - 0.5}{32}, rbin\right) \;\middle|\; 1 \le dbin \le 32,\ dbin \in \mathbb{Z}, 1 \le rbin \le 16,\ rbin \in \mathbb{Z} \right\} \tag{12}$$

3.2 Optimization of the Bandwidth for Kernel Density Estimation

The bandwidth h is a smoothing parameter, which is usually determined by the trade-off between the number of data samples and their standard deviation. Let n be the number of data samples and σ be the standard deviation of the data samples. The bandwidth h of a Gaussian kernel density estimator is given by the normal reference rule defined by Eq. (13). The normal reference rule is most commonly used to determine the bandwidth [30].

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-\frac{1}{5}} \tag{13}$$

We define the bandwidth $h_{rbin}(\alpha)$, which is optimized for each radius of the unit disk, as Eq. (14), Eq. (15), and Eq. (16), where *avg* is the arithmetic mean of the enhanced MFD values of each analysis window at *rbin*, and σ_{rbin} is the standard deviation of the enhanced MFD values at *rbin*. The smoothing parameter α in Eq. (16) is a constant. Through the experiments with different values of α, the constant value is determined so that the best result for the target task is obtained.

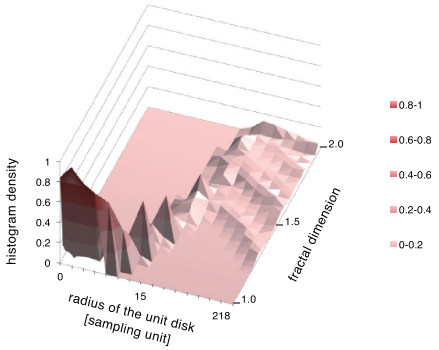

Fig. 4 The 3D histogram image that visualizes the existing EMFD signature of the cuckoo sound S_{cuckoo}

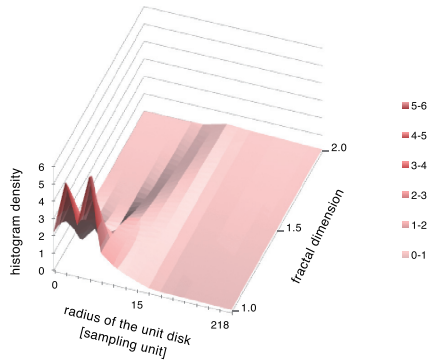

Fig. 5 The 3D histogram image that visualizes the EMFD-KDE signature of the cuckoo sound S_{cuckoo}. The bandwidth is $h_{rbin}(32)$.

$$avg = \frac{1}{NAW} \sum^{NAW} MFD_{enhanced}(rbin) \tag{14}$$

$$\sigma_{rbin} = \sqrt{\frac{1}{NAW} \sum^{NAW} (MFD_{enhanced}(rbin) - avg)^2} \tag{15}$$

$$h_{rbin}(\alpha) = 1.06\sigma_{rbin} NAW^{-\frac{1}{5}} \alpha \tag{16}$$

Figure 4 shows the 3D histogram visualizing the EMFD signature of the cuckoo sound S_{cuckoo}. Figure 5 shows the 3D histogram visualizing the EMFD-KDE signature of the same cuckoo sound S_{cuckoo}. The 3D histogram of the EMFD-KDE signature is much smoother than that of the EMFD signature. At each radius of the unit disk, the larger standard deviation of the enhanced MFD values σ_{rbin} results in the smoother histogram.

4. Very Long Range Multiscale Fractal Dimension Signature

The environmental sounds have important acoustic features with varying time periods. However, EMFD cannot describe the time-domain features for time periods longer than 10 ms.

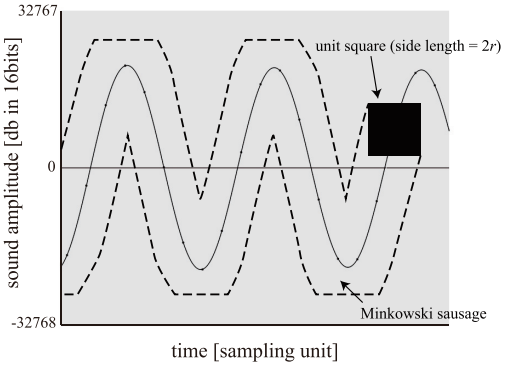

Fig. 6 A sound waveform and a Minkowski Sausage obtained by moving the unit square

To solve this problem, we propose a new acoustic feature signature for environmental sounds, based on the multiscale fractal dimension. This feature signature is called very-long-range multiscale fractal dimension signature (MFD-VL). The basic idea of MFD-VL is to extend the size range of the unit figure to consider the larger ones, which are used to compute the area of the Minkowski Sausage. In this section, we define the method to compute the MFD-VL signature and demonstrate its characteristics.

4.1 Definition of the MFD-VL Signature

The multiscale fractal dimension values of MFD-VL are computed for an entire target sound, and not for each fixed-length analysis window of the target sound. A unit square, instead of a unit disk, is used to compute the area of the Minkowski Sausage for MFD-VL. Figure 6 shows the Minkowski Sausage obtained by moving the unit square along the waveform. The method using the unit square is much faster than the one using the unit disk. Let n denote the sampling position, r the half side-length of the unit square, p the discrete position of the unit square, and $sig(x)$ the amplitude value of the sound signal at each sampling position x. The area of the Minkowski Sausage $area_{sq}(n, r)$ at each sampling position n is computed as Eq. (17).

$$area_{sq}(n,r) = \max_{\substack{0\le p\le 2r \\ p\in\mathbb{Z}}} (\mathrm{sig}(n-r+p)+r) - \min_{\substack{0\le p\le 2r \\ p\in\mathbb{Z}}} (\mathrm{sig}(n-r+p)-r) \quad (17)$$

The half side-length r of the unit square for each scale was defined as Eq. (18), where sf is the sampling frequency of the target sound. In this study, sf is 44100 Hz. Let $A_{sq}(r)$ denote the area of the Minkowski Sausage obtained for an entire target sound by moving the unit square whose side length is $2r$. The MFD-VL signature is defined as Eq. (19). The MFD-VL signature is a feature vector that contains 10 elements.

$$r(x) = \mathrm{round}\left(sf \times 2^{-\frac{x+2}{2}}\right) \quad (18)$$

$$MFD\text{-}VL = \left\{ 2 - \frac{\log\left(A_{sq}(r(x))/A_{sq}(r(x+1))\right)}{\log\left(r(x)/r(x+1)\right)} \;\middle|\; 0\le x\le 9,\ x\in\mathbb{Z} \right\} \quad (19)$$

4.2 Basic Characteristics of the MFD-VL Signature

We found several basic characteristics of MFD-VL through the experiments using test sounds.

4.2.1 MFD-VL's Descriptiveness of the Beats of Single Sine Waves

The MFD-VL signature is expected to describe the acoustic features over very long time-periods. We found that MFD-VL can discriminate frequencies of amplitude envelopes between 22.6 Hz and 1 Hz. The range of the wavelength of the amplitude envelopes corresponds to the range of the side length of the unit square between 0.044 s and 1 s. Let f_{beat} denote the frequency of the beats and $f_{content}$ denote the frequency of single sine waves inside the amplitude envelopes. The set of test sounds SS_{t1} is defined as Eq. (20), Eq. (21), and Eq. (22). Each test sound is filtered by the pink noise filter function f_{pn} Eq. (20). A sound that is artificially synthesized using pure tones usually has distinct or sparse spectra. This kind of sounds may cause numerical instabilities while calculating their acoustic features. To solve this problem, the pink noise filter function f_{pn} is used to add a background pink noise, which is defined as Eq. (20), where sig is an input signal and $Noise_{pink}$ is a background pink noise whose maximum amplitude is normalized to -0.1 db. The signal-to-noise ratio is 24 db. The pink noise, known as $1/f$ noise, is a signal whose power spectral density is inversely proportional to the signal frequency. The pink noise signal is known to widely exist in the natural world. The frequency components below 40 Hz contained in the pink noise are cut off by using a low cut filter before the amplitude normalization because the components with lower frequencies cannot be recorded nor played using common microphones and speakers.

$$f_{pn}(sig) = \frac{15}{16} sig + \frac{1}{16} Noise_{pink} \quad (20)$$

$$s_{t1}(f_{beat}, f_{content}) = f_{pn}(\cos(\pi f_{beat}\, t)\sin(2\pi f_{content}\, t)) \quad (21)$$

$$SS_{t1} = \{ s_{t1}(f_{beat}, f_{content}) \mid f_{content} = 440,\ f_{beat} \in \{0.5, 1, 2, 4, 8, 16, 32\} \} \quad (22)$$

Figure 7 shows the line charts of the MFD-VL values of single sine waves of frequency 440 Hz, which is filtered by Eq. (20), and those of the test sounds SS_{t1}. In Fig. 7, the frequencies of the beats are indicated by the troughs of the line chart, in which the side length of the unit square is less than the wavelengths of the beats. This characteristic can be understood morphologically as shown in Fig. 8. When the

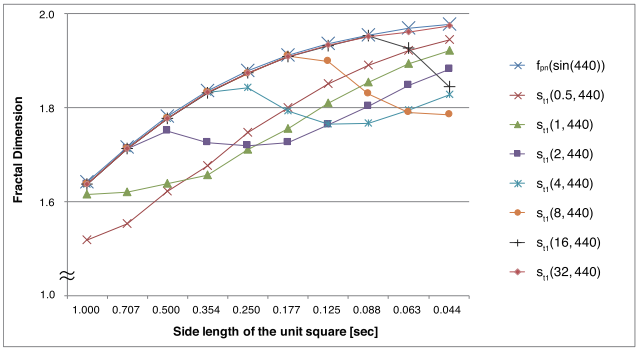

Fig. 7 Line charts of the MFD-VL signatures of single sine waves of frequency 440Hz, with and without a beat. The beat frequencies are 0.5Hz, 1Hz, 2Hz, 4Hz, 8Hz, 16Hz, and 32Hz.

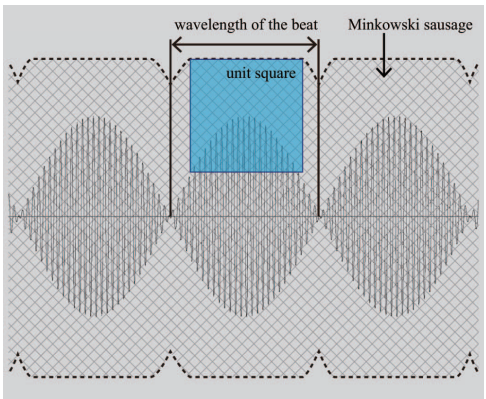

Fig. 8 Minkowski Sausage obtained by moving the unit square whose side length is less than the wavelength of the beat.

side length of the unit square is more than the wavelength of the beat, the area of the Minkowski Sausage becomes almost the same as that of a single sine wave without beats. When the side length of the unit square is less than the wavelengths of the beats, the shorter side length of the unit square results in a smaller area of Minkowski Sausage.

4.2.2 MFD-VL's Descriptiveness of Amplitude Envelope Shapes

Here we analyze some other characteristics of the MFD-VL that are related to the descriptiveness of the amplitude envelope shapes. Let f_{pulse} be the frequency of the rectangular pulse waves and w_{pulse} be the ratio of the rectangular pulse width to the wavelength of the rectangular pulse waves. The rectangular pulse function $rect(f_{pulse},\ w_{pulse},\ t)$ for generating the amplitude envelopes is defined as Eq. (23). Let $f_{content}$ be the frequency of a single sine wave inside the amplitude envelopes generated by the rectangular pulse function. The test sound s_{t2} is defined as Eq. (23) and Eq. (24). The test sound is filtered by the pink noise filter function f_{pn} Eq. (20).

$$rect(f_{pulse},\ w_{pulse},\ t) = \begin{cases} 1 & \left(t \bmod \frac{1}{f_{pulse}} <= \frac{w_{pulse}}{f_{pulse}}\right) \\ 0 & (otherwise) \end{cases} \quad (23)$$

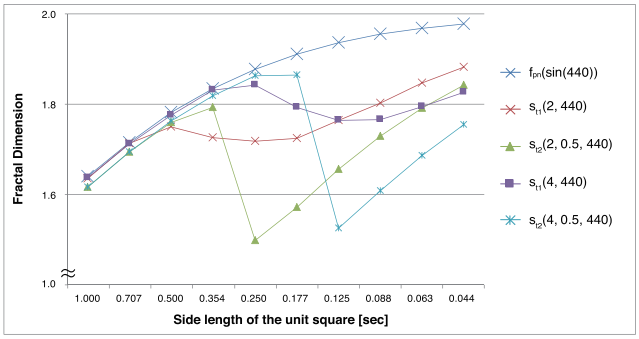

Fig. 9 Line charts of the MFD-VL signatures for single sine waves having a frequency of 440 Hz; those masked by the cosine function, and those masked by the rectangular pulse function.

$$s_{t2}(f_{pulse}, w_{pulse}, f_{content}) = f_{pn}(rect(f_{pulse},\ w_{pulse},\ t)\sin(2\pi f_{content}\ t)) \quad (24)$$

We define the set of test sounds SS_{t2} as Eq. (25). The line charts of MFD-VL of the single sine wave of frequency 440 Hz, which is filtered by Eq. (20), and SS_{t2} are showed in Fig. 9. Here, we compare the line charts of MFD-VL for s_{t1} and those for s_{t2}. This comparison shows that the bottom of the line chart trough of s_{t2} for $f_{pulse} = 2$ is deeper than that of s_{t1} for $f_{beat} = 2$, and that the bottom of the line chart trough of s_{t2} for $f_{pulse} = 4$ is also deeper than that of s_{t1} for $f_{beat} = 4$.

$$\begin{aligned} SS_{t2} = &\{s_{t1}(f_{beat}, f_{content}) \mid f_{content} = 440,\ f_{beat} \in \{2, 4\}\} \\ &\cup \{s_{t2}(f_{pulse}, w_{pulse}, f_{content}) \\ &\quad \mid f_{content} = 440,\ f_{pulse} \in \{2, 4\},\ w_{pulse} = 0.5\} \end{aligned} \quad (25)$$

For another comparison, we define a set of test sounds SS_{t3} as Eq. (26). The set of test sounds SS_{t3} contains single sine waves having a frequency of 440 Hz masked by the rectangular pulse functions with various widths w_{pulse}. In Fig. 10, we compare the line charts of MFD-VL of single sine wave having a frequency of 440 Hz, $s_{t1}(f_{content} = 440, f_{beat} = 4)$, and SS_{t3}. This comparison shows that the narrower width of the amplitude envelopes made by the rectangular pulse function results in deeper troughs in the line chart.

$$\begin{aligned} SS_{t3} = &\{s_{t2}(f_{pulse}, w_{pulse}, f_{content}) \\ &\quad \mid f_{content} = 440,\ f_{pulse} = 4,\ w_{pulse} \in \{0.2, 0.5, 0.8\}\} \end{aligned} \quad (26)$$

4.3 MFD-VL's Descriptiveness of Amplitude Envelopes of Simulated Environmental Sounds

Environmental sounds such as chirping of insects and birds, and water streams have amplitude-modulated waveforms. It is well-known that some type of environmental sounds can be simulated by the granular synthesis technique [31]. The

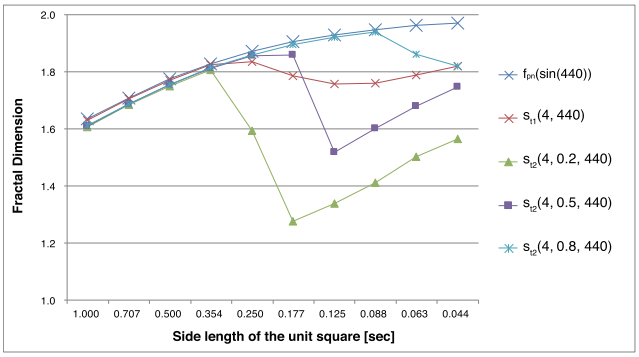

Fig. 10 Line charts of the MFD-VL signatures of single sine waves having a frequency of 440 Hz; those masked by the cosine function, and those masked by the rectangular pulse function having various widths.

granular synthesis technique splits a sound signal into small pieces called grains. Each grain has an envelope that contains an actual sonic content. If the content signal is a simple sinusoid, the synthesized sound signal is considered to be the same as a sound signal produced using the amplitude-modulation technique.

For example, the sound signal of the chirping of a cricket has a carrier wave with a frequency of around 5 kHz, and the carrier wave is modulated into trains of syllable chirps whose rate is near 30 Hz [32]. The sound of a cricket can be roughly simulated by the granular synthesis with an envelope created by the Hann window, which is defined as Eq. (27). Let f_c be the frequency of the carrier wave and f_e be the frequency of the envelope rate. The model of the chirping of a cricket $sim_{cricket}$ can be defined, as shown in Eq. (28) and Eq. (29). The sound signal of the chirping of a *Gryllus bimaculatus*, which is a species of cricket living in Japan, can be simulated as $sim_{cricket}(t, 5800, 30)$. It is known that the chirping of a *Gryllus bimaculatus* sometimes has continuous 3-syllable chirps that are repeated three times per second. Let f_c denote the frequency of the carrier wave, sn be the number of continuous syllable chirps, f_{rep} be the number of repetitions per second of the set of continuous syllable chirps, and f_e be the frequency of continuous chirps. Then, the model of chirping of a cricket $sim_{cricket2}$ is defined using Eq. (30) and Eq. (31). Figure 11 shows the sound waveform of signal $sim_{cricket2}(t, 5800, 3, 2.73, 30)$.

$$\omega(x) = \frac{1}{2}(1 - \cos(2\pi x)) \tag{27}$$

$$T_s(t, f) = \min\left(t \bmod \frac{1}{f}, \frac{1}{1.1f}\right) \tag{28}$$

$$sim_{cricket}(t, f_c, f_e) = \sin(2\pi f_c t)\,\omega(1.1 f_e T_s(t, f_e)) \tag{29}$$

$$rect_{cricket}(f_e,\ sn,\ f_{rep},\ t) = \begin{cases} 1 & \left(t \bmod \frac{1}{f_{rep}} \le \frac{sn}{f_e}\right) \\ 0.05 & (otherwise) \end{cases} \tag{30}$$

$$sim_{cricket2}\left(t, f_c, sn, f_{rep}, f_e\right) = rect_{cricket}(f_e,\ sn,\ f_{rep},\ t)\, sim_{cricket}(t, f_c, f_e) \tag{31}$$

We define the set of test sounds SS_{t4} as Eq. (32). For preventing sparse spectra of the synthesized sound, each test sound is filtered by the pink noise filter function f_{pn} Eq. (20), which is defined in Sect. 4.2.1. Figure 12 shows the line charts of MFD-VL of SS_{t4}. Figure 13 shows the 13-coefficients MFCCs values (MFCC13) of SS_{t4}. The SPTK toolkit [33] was used to compute MFCC13. The MFD-VL line charts of the sound set between $sim_{cricket}$ (group1) and $sim_{cricket2}$ (group2) are clearly different. These results indicate that the MFD-VL signature can clearly discriminate the two amplitude envelope signals.

$$SS_{t4} = \{f_{pn}(sim_{cricket}(t, f, 30)) \mid f \in \{5300, 5800, 6300\}\} \cup \{f_{pn}(sim_{cricket2}(t, f, 3, 2.73, 30)) \mid f \in \{5300, 5800, 6300\}\} \tag{32}$$

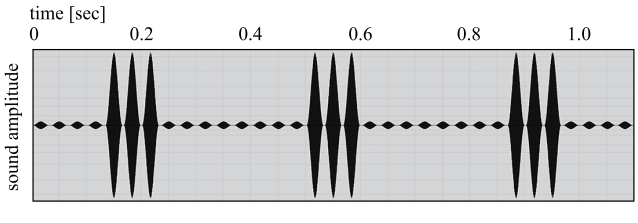

Fig. 11 Sound waveform of the signal $sim_{cricket2}(t, 5800, 3, 2.73, 30)$

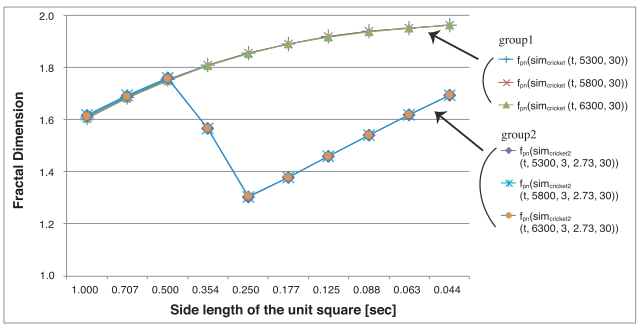

Fig. 12 Line charts of the MFD-VL signatures of simulated cricket sounds. These show the MFD-VL's robustness against fluctuations of a carrier signal's frequency.

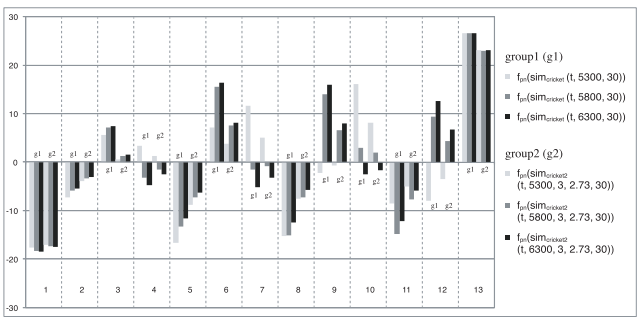

Fig. 13 MFCC13 values of simulated cricket sounds. These show the descriptiveness of varying carrier signals' frequency.

4.4 MFD-VL's Robustness against Fluctuations of a Carrier Signal's Frequency

We have assumed that one of the main causes of the diversity of environmental sounds is small fluctuations of sound source parameters, such as carrier signal frequency, due to the individuality of the sound source (D1). For example, the frequency of the human voice varies with the size and form of the individual's vocal tract which are the sound source parameters with fluctuations. However, we can recognize the sound recordings of talking by children, adults, and aged persons as the human voice in the same manner. The set of test sounds SS_{t4} contains the synthesized sound signals of the chirping of a cricket with different frequencies of a carrier wave. We can evaluate the descriptiveness of acoustic features for the small fluctuations of sound source parameters by comparing the feature vectors of test sounds contained in SS_{t4}.

Figure 12 shows the robustness of MFD-VL against the fluctuation of the carrier signal's frequency contained in the amplitude envelopes. We defined the discrimination rate of both MFD-VL and MFCC13 for the quantitative evaluation. Let $se(ts, n)$ denote the n-th element value of MFD-VL signature of the test sound clip ts. The MFD-VL discrimination rate DR_{mfdvl} for comparing two test sounds A and B is calculated using Eq. (33). The $RANGE_{mfdvl}$ in Eq. (33) denotes the width of the range of $se(ts, x)$, which may take a value between 1 and 2. Therefore, $RANGE_{mfdvl}$ is calculated as $1\ (= 2 - 1)$. The MFCC13 discrimination rate DR_{mfcc13} for comparing two test sounds A and B was defined as Eq. (34) where $se2(ts, n)$ was defined as the n-th dimension value of the MFCC13 vector of the test sound clip ts. The constant $RANGE_{mfcc13}$ was set to 56.3. This number was determined from the difference between the minimum and the maximum values of MFCC13 of 3000 environmental sounds in a dataset for the experimental evaluation which is described in Sect. 5. The sounds in the dataset were collected from the Freesound project [2].

$$DR_{mfdvl} = \sqrt{\frac{1}{10}\sum_{n=1}^{10}\left(\frac{se(A,n) - se(B,n)}{RANGE_{mfdvl}}\right)^2} \tag{33}$$

$$DR_{mfcc13} = \sqrt{\frac{1}{13}\sum_{n=1}^{13}\left(\frac{se2(A,n) - se2(B,n)}{RANGE_{mfcc13}}\right)^2} \tag{34}$$

Table 1 lists the comparisons of the two discrimination rates DR_{mfdvl} and DR_{mfcc} to distinguish two test sounds with various carrier signal frequencies which are contained in SS_{t4}. This comparison shows that to distinguish any two test sounds with various carrier signal frequencies, the values of DR_{mfdvl} are not more than 0.22%, and those of DR_{mfcc} are not less than 2.90% and no more than 16.15%. These results indicate that the MFD-VL signature shows robustness against the fluctuation of the carrier signal frequency, which is the diversity cause D1 defined in Sect. 1.2. The MFD-VL signature can clearly discriminate the characteristics of the amplitude envelopes with its robustness against the fluctuation of the carrier signal frequency.

Table 1 Comparison of discrimination rates of varying carrier signals' frequencies between the MFD-VL signature and MFCC13

	DR_{mfdvl}	DR_{mfcc}
$f_{pn}(sim_{cricket}(t, 5300, 30))$ and $f_{pn}(sim_{cricket}(t, 5800, 30))$	0.02%	16.15%
$f_{pn}(sim_{cricket}(t, 5800, 30))$ and $f_{pn}(sim_{cricket}(t, 6300, 30))$	0.20%	4.31%
$f_{pn}(sim_{cricket2}(t, 5300, 3, 2.73, 30))$ and $f_{pn}(sim_{cricket2}(t, 5800, 3, 2.73, 30))$	0.11%	7.31%
$f_{pn}(sim_{cricket2}(t, 5800, 3, 2.73, 30))$ and $f_{pn}(sim_{cricket2}(t, 6300, 3, 2.73, 30))$	0.22%	2.90%

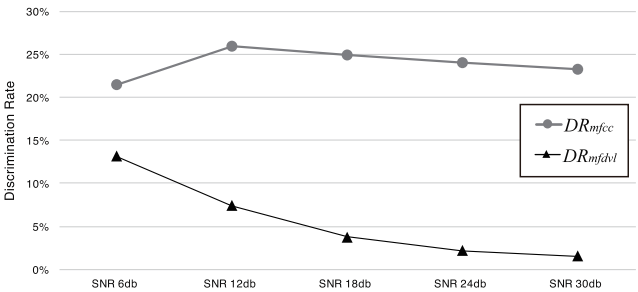

Fig. 14 Line charts of the discrimination rates DR_{mfdvl} and DR_{mfcc} to distinguish the simulated cricket sound $S_{cricket2}$ and each sound in the set of test sounds SS_{t5}.

4.5 MFD-VL's Robustness against Noises

To evaluate the robustness of MFD-VL against noises, we use test sounds generated from the simulated cricket sound and pink noise signals. Let $S_{cricket2}$ denote $sim_{cricket2}(t,\ 5800,\ 3,\ 2.73,\ 30)$, $Noise_{pink}$ denote a pink noise signal whose maximum amplitude is normalized to -0.1 db, and β be a constant number to control the signal-to-noise ratio (SNR). Then, a set of test sounds SS_{t5} is calculated using Eq. (35). The set of constant numbers $\beta \in \{0.67, 0.8, 0.89, 0.94, 0.97\}$ corresponds to the set of SNRs $\{6db, 12db, 18db, 24db, 30db\}$ of SS_{t5}. The frequency components below 40 Hz contained in the pink noise are cut off by using a low cut filter before the amplitude normalization because the components with lower frequencies cannot be recorded nor played using common microphones and speakers.

$$SS_{t5} = \{\beta\, S_{cricket2} + (1-\beta)\, Noise_{pink} \mid \beta \in \{0.67, 0.8, 0.89, 0.94, 0.97\}\} \tag{35}$$

Figure 14 shows the line charts of the discrimination rates DR_{mfdvl} and DR_{mfcc} for distinguishing the simulated cricket sound $S_{cricket2}$ and each sound in the set of test sounds SS_{t5}. The results indicate that the MFD-VL signature has the higher robustness against the pink noise than MFCC13, and that the MFD-VL signature is not relatively affected by the pink noise. It is confirmed that the MFD-VL signature is certainly robust against the diversity cause D2 defined in Sect. 1.2.

5. Experimental Evaluation (EXP1)

5.1 Experimental Setups

To evaluate a descriptiveness of the acoustic feature signatures based on the multiscale fractal dimension, we developed a similarity search system using the k-nearest neighbors (k-NN) method. We examined the performances of the similarity search tasks by using the proposed acoustic feature signatures.

To produce a sound dataset of environmental sounds, sound samples were collected from the Freesound project, which has stored many types of environmental sounds uploaded by users worldwide. The Freesound allows users to share their sounds and describe metadata regarding their shared sounds on the web. Each sound is labeled with a group of tags that are relatively well maintained as user generated contents [1]. The tags represent the objects the users were listening to in their listening experiences. Based on the following rules defined by the authors, the sounds were chosen and imported to our dataset to be used in the similarity search system. The sounds that were tagged with "field-recording" and with lengths between 1 and 600 s were chosen. This means that these sound clips have been recorded outside a recording studio. Each sound source outside the studio is typically unique and never the same. Therefore, sound clips that are tagged with "field-recording" have the diversity cause D1. We imported the top 3000 sounds in the descending order of downloaded number by unspecified users counted by the Freesound's system for each sound. Each sound was converted to a uniform format (1 channel, 44100 Hz sampling frequency, 16 bits bit depth, and maximal amplitude normalized to −0.1 db) for normalization before extracting acoustic features including EMFD, EMFD-KDE, MFD-VL, and MFCCs. The average length of the imported sounds is 70.4 s.

5.2 Acoustic Feature Extraction

One of the most well-known acoustic features used for ESR is MFCCs. Here, we used MFCCs for comparing the descriptiveness with our newly-proposed feature signatures. The SPTK toolkit was used to compute 13-coefficients MFCCs (MFCC13) and MFCC39, which represents the first and the second-order derivatives of MFCC13. MFCC13 and MFCC39 were computed using a fixed width analysis window of length 50 ms. The feature sets of MFCC13 and MFCC39 consist of mean values of their coefficients of the analysis window. EMFD and EMFD-KDE consist of the 512 elements defined in Sects. 2 and 3, and MFD-VL consists of the 10 elements defined in Sect. 4. In this experiment (EXP1), we determined that the smoothing parameter α used for computing the EMFD-KDE signature in Eq. (16) is 32.

Table 2 lists different feature sets to be compared through experimental evaluation. L1 represents the total number of features in the concatenated feature sets and L2 represents the number of features in the feature sets after dimensionality reduction through principal components analysis (PCA). To achieve the best possible performance of the similarity search using k-NN method, PCA was applied to the feature vectors of the most frequently downloaded 600 sounds in the dataset to extract its eigenvectors for dimensionality reduction. The "prcomp" function of R language was used for PCA processing. The corresponding L2s of feature sets 1, 3, 6, and 8 were determined so that each of their cumulative contribution ratios was 99%. The L2s of feature sets 2, 4, 5, 7, 9, and 10 were fixed to 114.

Table 2 List of acoustic feature sets for the comparison of their descriptiveness.

	Acoustic Feature Sets	L1	L2
FS1	1 MFCC13	13	11
	2 MFCC13 + EMFD (× 5.2)	525	114
	3 MFCC13 + MFD-VL (× 6)	23	16
	4 MFCC13 + EMFD-KDE (× 1)	525	114
	5 MFCC13 + EMFD-KDE (× 1) + MFD-VL (× 1)	535	114
FS2	6 MFCC39 [MFCC13 + 13Δ + 13ΔΔ]	39	24
	7 MFCC39 + EMFD (× 5)	551	114
	8 MFCC39 + MFD-VL (× 5.5)	49	27
	9 MFCC39 + EMFD-KDE (× 1)	551	114
	10 MFCC39 + EMFD-KDE (× 1) + MFD-VL (× 0.8)	561	114

Feature sets 1 and 6 are the standard sets for comparing with other feature sets. Each feature set from 2 to 5 consists of MFCC13, and each feature signature is based on the multiscale fractal dimension. We defined that each feature set from 1 to 5 belongs to the group FS1. In addition, each feature set from 7 to 10 consists of MFCC39 and each feature signature is based on the multiscale fractal dimension. Moreover, we defined each feature set from 6 to 10 to belong to group FS2.

The suffix "(× γ)" of each feature vector denotes a weighting coefficient γ. Each value of the feature vectors is multiplied by γ when its feature vector is combined with other feature(s). Through the experimental evaluation, the weighting coefficient γ for each feature vector was appropriately chosen to perform the best result.

5.3 Evaluation Method

The similarity search system using k-NN method returns a search result list of environmental sounds based on the distance in the space of the selected feature set through a search-key sound. To evaluate the performance using each feature vector in Table 2, we defined the similarity index SI between the tag group of the search-key sound $tags_{key}$ and that of the retrieved sound $tags_s$ as Eq. (36). This index is known as the Jaccard similarity coefficient that measures similarity between finite sample sets.

$$SI = \frac{\mathrm{card}\left(tags_{key} \cap tags_s\right)}{\mathrm{card}\left(tags_{key} \cup tags_s\right)} \tag{36}$$

To improve an accuracy of similarity index SI, we removed the commoner morphological and inflexional endings from all tags by using Porter Stemmer [34] in advance. Furthermore, the predefined stop words include sound formats, such as "mp3" and "stereo," and tool makers, such as "sony" and "tascam," were removed from the tag groups for computing the SIs. The tags that contain the text "fieldrecord" were removed from the tag groups because all sounds in the dataset have them. For each of the 3000 sounds in the dataset, SIs between a search-key sound and each retrieved sound in the search-result list were computed. We compared the average values of the SIs of 3000 sounds for each acoustic feature set.

5.4 Evaluation Results

Table 3 shows the SIs of "top n" for each feature set. The SI of "top n" is the average value of the SIs between a search-key sound and each retrieved sound in the top n rank of the search-result list. Table 3 shows the average values of the SIs of "top n" computed for each of the 3000 sounds. For reference, the average value of the SI between two randomly chosen sounds in the dataset is 0.014.

By comparing feature sets 4 and 9 with 2 and 7, respectively, it was confirmed that the descriptiveness of EMFD-KDE is superior to that of EMFD. SI of "top 1" of feature set 4 was 8.1% higher than that of feature set 2. SI of "top 1" of the feature set 9 was 4.9% higher than that of feature set 7.

By comparing feature sets 3 and 8 with 1 and 6, respectively, it was confirmed that the newly-developed MFD-VL signature improves the performance of similarity search result. The MFD-VL signature can describe some acoustic features that MFCCs cannot. SI of "top 1" of the feature set 3 was 8.2% higher than that of the feature set 1. SI of "top 1" of feature set 8 was 1.6% higher than that of the feature set 6.

The results obtained using the feature sets 5 and 10 achieved the best similarity search performance in each group of feature vectors FS1 and FS2. SI of "top 1" of feature set 5 was 17.2% higher than that of the feature set 1. SI of "top 1" of the feature set 10 was 8.7% higher than that of the feature set 6. It was confirmed that EMFD-KDE and MFD-VL are effective as an acoustic feature signature for the environmental sounds.

Table 3 Evaluation results of each feature set quantified using SIs

	Acoustic Feature Sets - FS1				
	1	2	3	4	5
top 1	0.212	0.226	0.229	0.244	**0.248**
top 2	0.170	0.178	0.184	0.193	**0.200**
top 3	0.145	0.150	0.156	0.164	**0.168**

	Acoustic Feature Sets - FS2				
	6	7	8	9	10
top 1	0.233	0.238	0.237	0.249	**0.253**
top 2	0.187	0.186	0.190	0.199	**0.204**
top 3	0.159	0.157	0.162	0.170	**0.172**

6. Experimental Evaluation Using Public Dataset and Other Acoustic Features (EXP2)

6.1 Experimental Setups

We developed another similarity search system to evaluate the proposed acoustic feature signatures by comparing them with the top-ranked acoustic features used in Task2 of the DCASE 2018 challenge [35]. As a sound dataset of this similarity search system, we used the train set of "Freesound Dataset Kaggle 2018" (FSDKaggle2018) [36] which contains 9473 sound clips provided as uncompressed PCM 16 bit, 44.1 kHz, mono audio files. The significant characteristics of the dataset are as follows:

- The sound clips are unequally distributed in the 41 categories of Google's AudioSet Ontology [37]. The minimum number of sound clips per category is 94, and the maximum is 300.
- Each sound clip is annotated with a single ground-truth label of the categories.
- The duration of the sound clips ranges from 300ms to 30s.

The maximal amplitude of sound clips was normalized to −0.1db before the acoustic feature extraction.

6.2 Acoustic Feature Extraction

We chose the top-ranked acoustic features proposed in the DCASE 2018 challenge task2, including log-mel energies, Perceptual weighted power spectrogram, and Logarithmic-filtered log-spectrogram, to compare their descriptiveness with that of the EMFD-KDE and MFD-VL signatures. Table 4 shows the acoustic feature sets to be used in EXP2. The significant processes of the feature extraction are as follows:

Table 4 List of acoustic feature sets for the comparison of their descriptiveness.

Labels	Acoustic Feature Sets	L1	L2
1'	MFCC39	39	25
2'	log-mel energies	1600	25
3'	Perceptual weighted power spectrogram	128	25
4'	Logarithmic-filtered log-spectrogram	128	56
5'	EMFD-KDE	512	249
6'	MFD-VL	10	5
7'	MFCC39 + EMFD-KDE	551	269
8'	MFCC39 + MFD-VL	49	28
9'	MFCC39 + EMFD-KDE + MFD-VL	561	272
10'	log-mel energies + EMFD-KDE	2112	195
11'	log-mel energies + MFD-VL	1610	27
12'	log-mel energies + EMFD-KDE + MFD-VL	2122	198
13'	Perceptual weighted power spectrogram + EMFD-KDE	640	269
14'	Perceptual weighted power spectrogram + MFD-VL	138	29
15'	Perceptual weighted power spectrogram + EMFD-KDE + MFD-VL	650	273
16'	Logarithmic-filtered log-spectrogram + EMFD-KDE	640	296
17'	Logarithmic-filtered log-spectrogram + MFD-VL	138	60
18'	Logarithmic-filtered log-spectrogram + EMFD-KDE + MFD-VL	650	300

- All acoustic features except MFD-VL were computed using fixed-width analysis windows of length 50ms every 25ms.
- For computing the MFD-VL signature, the length of the sound signal must be longer than 1s. In case the target sound clip is shorter than or equal to 1s, we repeated the sound signal so that the total length of the repeated signals becomes longer than 1s.
- Let the smoothing parameter α be 1, which is used for computing the EMFD-KDE signature in Eq. (16).
- Dimensionality reduction was performed through PCA using *scikit-learn* library. The numbers of features L2s after dimensionality reduction were determined so that each of their cumulative contribution ratios was 98%.

6.3 Evaluation Method

For each of the 9473 sounds in the dataset, the relevances between the search-key sound and each retrieved sound in the search-result list were computed using their category labels. Then, we computed the mean value of Precision@k for each query on the entire dataset to measure the top-k ranking quality of the search result lists.

6.4 Evaluation Results

Table 5 shows the Precision@k scores at ranking positions 1, 3, and 10 on the entire dataset. Bold numbers are the best Precision@k scores for each ranking position k. The results obtained using the feature set 8' (MFCC39 + MFD-VL) achieved the best similarity search performance. By comparing feature sets (2' with 10', 11', and 12'), (3' with 13', 14', and 15'), and (4' with 16', 17', and 18') respectively, it was confirmed that both the EMFD-KDE and MFD-VL

Table 5 Precision@k scores at ranking positions 1,3, and 10 on the entire dataset.

	Feature Sets								
	1'	2'	3'	4'	5'	6'	7'	8'	9'
Precision@1	0.561	0.464	0.473	0.500	0.487	0.177	0.543	**0.572**	0.555
Precision@3	0.492	0.391	0.400	0.425	0.418	0.161	0.471	**0.502**	0.480
Precision@10	0.395	0.299	0.304	0.322	0.334	0.141	0.378	**0.402**	0.387

	Feature Sets								
	10'	11'	12'	13'	14'	15'	16'	17'	18'
Precision@1	0.548	0.477	0.552	0.534	0.522	0.542	0.534	0.540	0.544
Precision@3	0.471	0.403	0.474	0.463	0.453	0.472	0.457	0.464	0.466
Precision@10	0.366	0.307	0.369	0.367	0.354	0.375	0.361	0.359	0.369

Table 6 Precision@3 scores for each category.

Categories	Acoustic Feature Sets																	
	1'	2'	3'	4'	5'	6'	7'	8'	9'	10'	11'	12'	13'	14'	15'	16'	17'	18'
Acoustic_guitar	0.523	0.409	0.419	0.409	0.459	0.159	0.517	**0.544**	0.518	0.489	0.417	0.488	0.518	0.472	0.523	0.468	0.450	0.480
Applause	**0.809**	0.763	0.742	0.779	0.731	0.244	0.761	0.797	0.776	0.784	0.770	0.782	0.769	0.779	0.771	0.779	0.799	0.780
Bark	0.386	0.332	0.326	0.427	0.305	0.109	0.370	0.416	0.378	0.382	0.354	0.381	0.375	0.381	0.378	0.367	**0.502**	0.372
Bass_drum	0.472	0.467	0.458	0.553	0.644	0.172	0.647	0.541	0.648	0.628	0.492	0.628	0.659	0.539	**0.661**	0.646	0.617	0.652
Burping_or_eructation	0.340	0.305	0.322	0.405	0.344	0.065	0.375	0.405	0.410	0.390	0.319	0.392	0.398	0.408	0.405	0.416	0.411	**0.430**
Bus	**0.514**	0.330	0.291	0.281	0.187	0.177	0.235	0.502	0.235	0.284	0.336	0.297	0.214	0.346	0.226	0.223	0.330	0.239
Cello	0.623	0.493	0.497	0.439	0.319	0.169	0.388	**0.634**	0.402	0.477	0.507	0.482	0.388	0.570	0.396	0.369	0.477	0.371
Chime	0.380	0.339	0.301	0.336	0.446	0.087	0.461	0.383	**0.470**	0.426	0.357	0.429	0.458	0.354	0.458	0.455	0.368	0.461
Clarinet	**0.747**	0.616	0.558	0.589	0.501	0.326	0.559	0.744	0.557	0.600	0.613	0.601	0.557	0.606	0.558	0.539	0.601	0.550
Computer_keyboard	0.294	0.148	0.123	0.148	0.325	0.227	0.353	0.345	**0.359**	0.291	0.168	0.300	0.322	0.232	0.317	0.319	0.283	0.328
Cough	0.281	0.163	0.174	0.214	0.303	0.060	0.340	0.300	**0.359**	0.281	0.171	0.285	0.335	0.230	0.346	0.331	0.274	0.350
Cowbell	0.754	0.749	0.740	0.771	0.634	0.391	0.670	0.775	0.675	0.735	0.747	0.735	0.675	0.770	0.682	0.656	**0.780**	0.665
Double_bass	0.761	0.719	0.682	0.606	0.621	0.191	0.670	**0.766**	0.681	0.763	0.737	0.762	0.696	0.738	0.703	0.678	0.663	0.691
Drawer_open_or_close	**0.272**	0.217	0.255	0.203	0.203	0.074	0.266	0.241	0.255	0.251	0.247	0.264	0.238	0.243	0.234	0.238	0.268	0.228
Electric_piano	0.518	0.369	0.400	0.273	0.289	0.240	0.338	**0.598**	0.342	0.398	0.396	0.407	0.371	0.473	0.382	0.331	0.373	0.331
Fart	0.446	0.256	0.401	0.320	0.383	0.091	0.424	0.432	0.444	0.474	0.263	0.477	0.460	0.409	0.472	0.481	0.348	**0.486**
Finger_snapping	0.519	0.507	0.544	0.632	0.547	0.490	0.558	**0.647**	0.575	0.561	0.544	0.561	0.561	0.570	0.564	0.581	0.638	0.564
Fireworks	**0.544**	0.338	0.364	0.417	0.381	0.087	0.462	0.491	0.480	0.450	0.359	0.456	0.429	0.422	0.447	0.436	0.436	0.444
Flute	0.464	0.348	0.371	0.384	0.462	0.140	0.530	0.488	0.534	0.471	0.357	0.473	0.530	0.427	**0.536**	0.483	0.411	0.496
Glockenspiel	0.809	0.635	0.606	0.621	0.613	0.270	0.674	**0.830**	0.688	0.766	0.638	0.773	0.745	0.649	0.748	0.660	0.638	0.674
Gong	**0.567**	0.434	0.421	0.461	0.329	0.158	0.404	0.559	0.412	0.449	0.435	0.454	0.390	0.476	0.416	0.373	0.494	0.392
Gunshot_or_gunfire	0.141	0.147	0.136	**0.243**	0.177	0.050	0.206	0.138	0.220	0.206	0.145	0.213	0.186	0.138	0.181	0.224	**0.243**	0.229
Harmonica	0.505	0.352	0.392	0.432	0.479	0.105	0.564	0.491	0.574	0.481	0.360	0.481	0.545	0.499	0.554	0.580	0.461	**0.596**
Hi-hat	0.530	0.447	0.413	0.492	0.466	0.143	0.548	0.553	**0.557**	0.520	0.453	0.522	0.499	0.457	0.513	0.489	0.504	0.491
Keys_jangling	0.441	0.257	0.285	0.384	0.475	0.110	0.530	0.439	0.537	0.535	0.247	**0.547**	0.496	0.353	0.516	0.489	0.463	0.492
Knock	0.493	0.381	0.397	0.449	0.394	0.232	0.425	0.515	0.446	0.410	0.411	0.409	0.434	0.510	0.447	0.419	**0.579**	0.434
Laughter	0.334	0.179	0.211	0.253	0.289	0.113	0.391	0.362	**0.401**	0.336	0.196	0.340	0.322	0.272	0.338	0.369	0.302	0.390
Meow	0.211	0.095	0.071	0.140	0.148	0.067	0.200	**0.219**	0.194	0.125	0.110	0.127	0.161	0.092	0.159	0.172	0.185	0.185
Microwave_oven	**0.235**	0.135	0.123	0.132	0.091	0.053	0.135	0.210	0.148	0.158	0.155	0.160	0.123	0.153	0.123	0.139	0.201	0.139
Oboe	0.835	0.713	0.826	0.717	0.734	0.401	0.800	**0.837**	0.798	0.780	0.710	0.784	0.807	**0.837**	0.816	0.771	0.715	0.780
Saxophone	0.627	0.434	0.531	0.451	0.420	0.251	0.472	**0.652**	0.480	0.516	0.442	0.519	0.506	0.628	0.509	0.471	0.480	0.484
Scissors	0.189	0.077	0.049	0.151	0.116	0.168	0.154	0.249	0.165	0.119	0.095	0.123	0.123	0.168	0.133	0.133	**0.270**	0.151
Shatter	0.387	0.263	0.281	0.370	0.404	0.084	0.538	0.413	**0.563**	0.502	0.279	0.510	0.452	0.349	0.486	0.446	0.441	0.497
Snare_drum	0.342	0.362	0.311	0.377	0.412	0.123	0.460	0.346	0.469	0.486	0.374	**0.488**	0.441	0.363	0.447	0.458	0.440	0.447
Squeak	**0.212**	0.119	0.140	0.129	0.121	0.061	0.153	0.207	0.159	0.153	0.127	0.154	0.150	0.177	0.154	0.149	0.141	0.151
Tambourine	0.769	0.697	0.732	**0.772**	0.712	0.163	0.738	0.701	0.742	0.742	0.701	0.744	0.738	0.739	0.738	0.738	0.756	0.733
Tearing	0.358	0.304	0.287	0.363	0.349	0.081	0.396	0.372	0.408	0.379	0.309	0.383	0.364	0.301	0.371	0.407	0.389	**0.410**
Telephone	0.169	0.211	0.158	0.206	0.125	0.069	0.147	0.175	0.150	0.189	0.225	0.203	0.156	**0.250**	0.158	0.161	0.242	0.167
Trumpet	0.584	0.419	0.462	0.452	0.507	0.180	0.559	**0.604**	0.564	0.509	0.427	0.511	0.557	0.516	0.570	0.523	0.500	0.532
Violin_or_fiddle	**0.554**	0.470	0.402	0.506	0.489	0.158	0.532	0.533	0.543	0.508	0.476	0.509	0.531	0.451	0.537	0.494	0.517	0.498
Writing	0.390	0.274	0.299	0.336	0.402	0.109	0.447	0.384	**0.454**	0.420	0.296	0.422	0.428	0.368	0.438	0.443	0.358	0.453
Number of the best scores	9			2				**10**	7			2		2	2		5	4

signatures improve the performance of the similarity search task. By comparing feature sets 1' to 6', we confirmed that the Precision@10 score obtained using the feature set 5' (EMFD-KDE) achieved better performance than those obtained using the feature sets 2' (log-mel energies), 3' (Perceptual weighted power spectrogram), and 4' (Logarithmic-filtered log-spectrogram), respectively.

Table 6 shows the Precision@3 scores of the search-result list grouped by the 41 categories for each feature set. Bold numbers are the best Precision@3 scores for each category. The top five numbers of the categories in which each feature set obtained the best Precision@3 score are 10 categories for feature set 8' (MFCC39 + MFD-VL), 9 categories for feature set 1' (MFCC39), 7 categories for feature set 9' (MFCC39 + EMFD-KDE + MFD-VL), 5 categories for feature set 17' (Logarithmic-filtered log-spectrogram + MFD-VL), and 4 categories for feature set 18' (Logarithmic-filtered log-spectrogram + EMFD-KDE + MFD-VL).

7. Conclusion

Recent research on ESR focused on the evaluation of time-domain features of environmental sounds. For ESR, an acoustic feature must describe the nonstationary characteristics of target sounds as time-domain features and must be robust against the following three main causes of the diversity of environmental sounds.

D1) Small fluctuations of sound source parameters, such as carrier signal frequency, due to the individuality of the sound source.
D2) Background noises that the person who recorded the target sound did not expect to record.
D3) Mixed composition of different types of sound sources.

In this study, we have focused on the extraction of acoustic feature signatures that are robust against the diversities caused by D1 and D2.

In a previous study, we proposed the EMFD feature signature to describe both frequency- and time-domain features of target sounds.

However, we also recognized the following problems in EMFD.

p1) EMFD includes error values.
p2) It is oversensitive to discriminate the features of environmental sounds.
p3) It lacks the robustness against the diversity of environmental sounds.
p4) It cannot describe the time-domain features for periods longer than 10 ms.

To solve these problems, we proposed the EMFD-KDE and MFD-VL feature signatures.

The newly-proposed EMFD-KDE feature signature is the probability distribution of the enhanced MFD values at each radius of the unit disk computed using the kernel density estimation method. In Sect. 3.2, we studied the method to optimize bandwidth h used for kernel density estimation. Based on the normal reference rule, we defined bandwidth $h_{rbin}(\alpha)$ optimized for each radius of the unit disk by using Eq. (16). Through the experiments with different values of the smoothing parameter α, we determined that the best result of the similarity search task EXP1 is obtained for $\alpha = 32$ and that of EXP2 is obtained for $\alpha = 1$. We assume that this difference of the optimized value of α for each task is caused by the diversity difference of datasets for each task. In the FSDKaggle2018 dataset for EXP2, a number of the ground truth labels have been manually verified, the number of categories is only 41, and the length of the sound clip is shorter than 30sec. The dataset for EXP1 is more diverse and complex than the FSDKaggle2018 dataset.

In Sect. 4, we proposed the MFD-VL signature and demonstrated its characteristics through experiments using the simulated cricket's sounds as follows.

- MFD-VL can discriminate the frequencies of the amplitude envelopes between 22.6 and 1 Hz.
- It can discriminate the shapes of the amplitude envelopes.
- It is robust against the fluctuation of the carrier signal frequency (, i.e., the diversity cause D1 defined in Sect. 1.2).
- It is robust against background noises (, i.e., the diversity cause D2 defined in Sect. 1.2).

The MFD-VL signature is expected to describe the time-domain features for periods longer than 10 ms. The MFD-VL signature shows stability and robustness against background noises and small fluctuations of the carrier signal's frequency.

From the experimental evaluation results (EXP1), we confirmed that the descriptiveness of EMFD-KDE supplementing MFCC13 and MFCC39 is evidently higher than that of EMFD supplementing MFCC13 and MFCC39. We conclude that the smoothness of the EMFD-KDE signature can solve problems p1, p2, and p3. Furthermore, the experimental evaluation results showed that the MFD-VL signature supplementing EMFD-KDE and MFCCs improves the performance of the similarity search. The MFD-VL signature functions as an effective time-domain feature and can solve problems p3 and p4.

In Sect. 6, we conducted another experimental evaluation (EXP2) using the FSDKaggle2018 dataset and other acoustic features. We confirmed clear evidence that both the EMFD-KDE and MFD-VL signatures have the unique descriptiveness of environmental sound. These signatures are effective when they are used with other acoustic features, including MFCC39 and the top-ranked acoustic features in the DCASE 2018 challenge task2.

The EMFD-KDE signature has 512 feature elements, which is relatively more than those of other feature signatures. This implies that the EMFD-KDE signature requires more computational time than the conventional methods for feature extraction. For the similarity search task, feature signatures of library sounds can be computed *a priori* which implies that we do not care about the computation time of

the EMFD-KDE signature. While the EMFD-KDE also requires more searching time than conventional method, this time is negligible compared to the time required for the retrieval of matched sounds from the library. The computational time of EMFD-KDE and the length of searching time using EMFD-KDE do not matter.

Environmental sounds have the acoustic features in their frequency domain, as well as other important features in the time domain with various time scales. We conclude that both the EMFD-KDE and MFD-VL signatures can describe the essential acoustic features of environmental sounds with robustness against the diversity of environmental sounds. Further studies are needed to evaluate the performance of other applications using these acoustic feature signatures, such as classification tasks using machine-learning systems.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 18K11378.

References

[1] V. Akkermans, F. Font, J. Funollet, B. de Jong, G. Roma, S. Togias, and X. Serra, "FREESOUND 2.0: An Improved Platform for Sharing Audio Clips," 12th Int. Soc. Music Inf. Retr. Conf., 2011.
[2] Music Technology Group of Universitat Pompeu Fabra, "The Freesound Project.," https://www.freesound.org/ (Retrieved 2020-12-20).
[3] SoundCloud Limited, "SoundCloud – Hear the world's sounds," https://soundcloud.com/ (Retrieved 2020-12-20).
[4] F. López, "Environmental sound matter," 1998. [Online]. Available: http://www.franciscolopez.net/pdf/env.pdf [Accessed: 23-Apr-2021].
[5] D. Michael, "Toward a Dark Nature Recording," Organised Sound, vol.16, no.3, pp.206–210, 2011.
[6] T.H. Park, J. Lee, J. You, M.-J. Yoo, and J. Turner, "Towards Soundscape Information Retrieval (SIR)," Proc. ICMC—SMC —2014, Athens, Greece, 2014, pp.1218–1225, 2014.
[7] S. Chachada and C.-C.J. Kuo, "Environmental sound recognition: A survey," APSIPA Trans. Signal Inf. Process., vol.3, 2014.
[8] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," IEEE Trans. Multimed., vol.17, no.10, pp.1733–1746, 2015.
[9] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "CP-JKU Submissions to DCASE'20: Low-Complexity Cross-Device Acoustic Scene Classification with RF-Regularized CNNs," DCASE2020 Challenge, Tech. Rep., 2020.
[10] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," Pattern Recognit. Lett., vol.24, no.15, pp.2895–2907, 2003.
[11] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental Sound Recognition With Time-Frequency Audio Features," IEEE Trans. Audio. Speech. Lang. Processing, vol.17, no.6, pp.1142–1158, 2009.
[12] R. Mogi and H. Kasai, "Noise-Robust environmental sound classification method based on combination of ICA and MP features," Artif. Intell. Res., vol.2, no.1, p.107, 2012.
[13] C. Bauge, M. Lagrange, J. Anden, and S. Mallat, "Representing environmental sounds using the separable scattering transform," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.8667–8671, 2013.
[14] J. Xue, G. Wichern, H. Thornbug, and A. Spanias, "Fast query by example of environmental sounds via robust and efficient cluster-based indexing," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.5–8, 2008.
[15] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra, "Ecological Acoustics Perspective for Content-Based Retrieval of Environmental Sounds," EURASIP J. Audio, Speech, Music Process., vol.2010, pp.1–11, 2010.
[16] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," Vancouver, British Columbia, Canada, pp.105–112, 2008.
[17] M. Sunouchi and Y. Tanaka, "Similarity Search of Freesound Environmental Sound Based on Their Enhanced Multiscale Fractal Dimension," Sound Music Comput. Conf. 2013, SMC 2013, pp.715–721, 2013.
[18] S. Handel, "Timbre perception and auditory object identification," Hearing, Academic Press, p.468, 1995.
[19] D. Mitrović, M. Zeppelzauer, and H. Eidenberger, "On Feature Selection in Environmental Sound Recognition," 51st Int. Symp. ELMAR, no.September, pp.28–30, 2009.
[20] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Features for Content-Based Audio Retrieval," Adv. Comput., 2010, vol.78, ch.3, pp.71–150, 2010.
[21] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," IEEE Trans. Signal Process., vol.41, no.12, pp.3397–3415, 1993.
[22] S. Innami and H. Kasai, "NMF-based environmental sound source separation using time-variant gain features," Comput. Math. with Appl., vol.64, no.5, pp.1333–1342, 2012.
[23] B. Mandelbrot, "The Fractal Geometry of Nature," W.H. Freeman and Company, 1982.
[24] R.F. Voss and J. Clarke, " '1/f noise' in music and speech," Nature, vol.258, no.5533, pp.317–318, 1975.
[25] K.J. Hsu and A.J. Hsu, "Fractal Geometry of Music," Proc. Natl. Acad. Sci., vol.87, no.3, pp.938–941, 1990.
[26] P. Maragos, "Fractal aspects of speech signals: dimension and interpolation," [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, pp.417–420, 1991.
[27] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: computation and application to automatic speech recognition," J. Acoust. Soc. Am., vol.105, no.3, pp.1925–1932, 1999.
[28] A. Zlatintsi and P. Maragos, "Musical Instruments Signal Analysis and Recognition Using Fractal Features," 19th Eur. Signal Process. Conf. (EUSIPCO 2011), Barcelona, Spain, no.Eusipco, pp.684–688, 2011.
[29] A. Zlatintsi and P. Maragos, "Multiscale Fractal Analysis of Musical Instrument Signals With Application to Recognition," IEEE Trans. Audio. Speech. Lang. Processing, vol.21, no.4, pp.737–748, 2013.
[30] D.W. Scott, "Multivariate Density Estimation: Theory, Practice, and Visualization," 1st ed. Wiley, 1992.
[31] D. Keller and B. Truax, "Ecologically-based granular synthesis," Proc. Int. Comput. Music Conf., Ann Arbor, USA, 1998, pp.117–120, 1998.
[32] J. Thorson, T. Weber, and F. Huber, "Auditory behavior of the cricket," J. Comp. Physiol. A. Neuroethol. Sens. Neural. Behav. Physiol., vol.146, no.3, pp.361–378, 1982.
[33] SPTK working group, "Speech Signal Processing Toolkit (SPTK)," http://sp-tk.sourceforge.net/, (Retrieved 2020-12-20).
[34] M.F. Porter, "An algorithm for suffix stripping," Progr. Electron. Libr. Inf. Syst., vol.14, no.3, pp.130–137, 1980.
[35] E. Fonseca, M. Plakal, F. Font, D.P.W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of Freesound audio with AudioSet labels: task description, dataset, and baseline," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), pp.69–73, 2018.
[36] E. Fonseca, X. Favory, J. Pons, F. Font, M. Plakal, D.P.W. Ellis,

and X. Serra, "FSDKaggle2018," 29-Jan-2019. [Online]. Available: https://zenodo.org/record/2552860. [Accessed: 23-Apr-2021].

[37] J.F. Gemmeke, D.P.W. 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," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.776–780, 2017, doi: 10.1109/ICASSP.2017.7952261.

Motohiro Sunouchi received masters of environmental studies in human and engineered environmental studies from the University of Tokyo in 2004. He is currently pursuing the Ph.D. in information science at the Hokkaido University. He has been a research assistant since 2007 and a senior lecturer since 2016 with the Department of Design, Sapporo City University, Japan. His research interests lie in the areas of audio signal processing and auditory culture.

Masaharu Yoshioka is a Professor of Faculty of Information Science and Technology, Global Station of Big Data and Cybersecurity, and Institute for Chemical Reaction Design and Discovery of Hokkaido University. He received the B.E. and M.E. degrees of precision engineering and the Ph.D. degree of precision machinery engineering from University of Tokyo, Japan, in 1991, 1993, and 1996, respectively. From April 1996 to March 2000, he was a Research Associate of National Center for Science and Information Systems, Japan. From April 2000 to May 2001, he was a Research Associate of National Institute of Informatics, Japan. From June 2001, he joined the Graduate School of Engineering as a Associate Professor and this school is reorganized as Graduate School of Information Science and Technology in 2004. From January 2019, he became a Professor of Faculty of Information Science and Technology. He also joined Institute for Chemical Reaction Design and Discovery from January 2020. From October 2017, he also serve as a visiting researcher at RIKEN Center for Advanced Intelligence Project. His research interests includes application of knowledge engineering technology for information access and knowledge management, Linked Open Data, and application of knowledge engineering technology for a particular research domain (e.g., cheminformatics and nanoinformatics).