Hindawi
Security and Communication Networks
Volume 2021, Article ID 5566347, 12 pages
https://doi.org/10.1155/2021/5566347

Research Article

Feature Extraction Method for Hidden Information in Audio Streams Based on HM-EMD

Jiu Lou, Zhongliang Xu, Decheng Zuo, and Hongwei Liu

Harbin Institute of Technology, School of Computer Science and Technology, Harbin 150001, China

Correspondence should be addressed to Decheng Zuo; zuodc@hit.edu.cn

Received 14 February 2021; Revised 22 March 2021; Accepted 21 July 2021; Published 2 August 2021

Academic Editor: Shehzad Ashraf Chaudhry

Using fake audio to spoof the audio devices in the Internet of Things has become an important problem in modern network security. Aiming at the problem of lack of robust features in fake audio detection, an audio streams' hidden feature extraction method based on a heuristic mask for empirical mode decomposition (HM-EMD) is proposed in this paper. First, using HM-EMD, each signal is decomposed into several monotonic intrinsic mode functions (IMFs). Then, on the basis of IMFs, basic features and hidden information features HCFs of audio streams are constructed, respectively. Finally, a machine learning method is used to classify audio streams based on these features. The experimental results show that hidden information features of audio streams based on HM-EMD can effectively supplement the nonlinear and nonstationary information that traditional features such as mel cepstrum features cannot express and can better realize the representation of hidden acoustic events, which provide a new research idea for fake audio detection.

1. Introduction

With the development of the Internet of Things (IoT) technology, an increasing number of audio and video acquisition devices are now connected to the Internet. Fake audio becomes an increased new threat on voice interfaces due to the recent breakthroughs in speech synthesis and voice conversion technologies. Therefore, the detection of fake audio has become a new hot issue of network security [1, 2]. There are mainly two methods of audio forgery. One is to generate spoofed utterances using text-to-speech (TTS) and voice conversion (VC) algorithms, which is also called logic access (LA) [3] and the other is the use of professional replay devices to get spoof attack, which is also known as physical access (PA) [4]. There are more diversity of audio forgery means and more difficulty in fake audio detection [5]. In this paper, an audio streams hidden feature extraction method based on HM-EMD is proposed and used to extract features of audio streams to detect fake audio.

At present, the fake audio detection is mainly based on acoustic features to build classification model. Linear frequency cepstral coefficients (LFCC) [6], constant-Q cepstral coefficients (CQCC) [7], and mel frequency cepstral coefficients (MFCC) [8] are commonly used in fake audio detection. However, these features are based on fixed filter banks; none of these acoustic features are able to generalize well on unknown spoofing technologies. Subsequently, the end-to-end deep learning method to detect the fake audio is gradually concerned by researchers. Alejandro et al. proposed a cyclic neural network based on optical convolution gate to extract the shallow features at the frame level and the deep features of sequence dependence at the same time [9]. Zeinali et al. used VGg and light CNN to detect fake audio [10]. However, this end-to-end deep learning approach requires large, evenly distributed datasets and relies on certain types of fake audio.

Through the analysis of forged audio, it is found that the AI-based fake audio technology focuses more on speech content and ignores the background sound in an audio stream [11]. The background sound will also change during the replay spoofing. Therefore, the construction of features representing the hidden information in audio scenes can be used to detect the fake audio [12].

In order to focus on local level details of the signal in terms of specific regions (which may be highly discriminative), empirical-mode-decomposition- (EMD-) based approach is explored. EMD has superior time-frequency resolution performance in nonlinear unsteady signal processing and has been applied in counterfeit audio detection [13].

However, the traditional EMD method has a few disadvantages, including mode aliasing and the inconsistency of IMF dimensions after signal decomposition. Hence, accurately estimating the IMF range of a certain frequency distribution is difficult. In 2005, Deering and Kaiser proposed the ensemble empirical mode decomposition (EEMD) decision method [14], which attempts to solve the problem of mode aliasing by introducing Gaussian white noise into the signal to be decomposed. In EEMD, the attributes of Gaussian white noise should be adjusted artificially. However, the Gaussian white noise leaves traces in the IMF decomposed from the signal, thereby resulting in low signal restoration accuracy and extensive calculations. Time-varying filtering-based empirical mode decomposition (TVF-EMD) uses the b-spline time-varying filter for mode selection and thus solves the problem of mode aliasing to a certain extent. However, TVF-EMD must calculate the cutoff frequency first, thus leaving the problem of dimension inconsistency unsolved [15].

To sum up, in order to make full use of the time-frequency analysis advantages of EMD, it is necessary to solve the modal aliasing and frequency inconsistency problems existing in EMD itself. In this paper, a heuristic empirical mode decomposition (HM-EMD) method is proposed to improve the purity of IMFS and solve the problem of inconsistency between mode mixing and IMF dimension. Then, the acoustic hidden component features (AHCF) of the audio stream were constructed and used to locate the acoustic events in audio stream in the acoustic stream classification dataset A of DCASE [16]. Fake audio detection is implemented on ASVSpoof2019 dataset [17]. The experimental results show that the basic features and AHCFs of the audio streams based on HM-EMD can represent the audio background which help to verify the types of the fake audio.

The paper consists of five parts. The first part is an introduction of audio streams. The second part mainly introduces the principle of HM-EMD. The third part describes the mining of hidden information in audio streams based on the proposed HM-EMD. The fourth part presents the results of classification of audio streams on the basis of HM-EMD. The fifth part summarizes the characteristics of the proposed method and presents future research directions.

2. Heuristic Mask for Empirical Mode Decomposition (HM-EMD)

In this section, the classical empirical mode decomposition (EMD) method is first introduced and then follows the analysis of mode aliasing in EMD; finally, the solution of mode aliasing based on heuristic mask signal is proposed in detail.

2.1. Empirical Mode Decomposition Method

2.1.1. Empirical Mode Decomposition. EMD can decompose the original signal $x(t)$ $(t \in N, N = \{0, 1, \ldots n\})$ into a series of IMFs whose upper and lower envelopes have a mean value of 0. This decomposition method does not need to preset basis functions (such as Fourier transform or wavelet analysis), but the IMFs should satisfy the following formulas:

$$\left|\mathrm{Num}_{\mathrm{extream}} - \mathrm{Num}_{\mathrm{cross}}\right| \leq 1, \tag{1}$$

$$\sum_{t \in N} B_{\max}(t) + \sum_{t \in N} B_{\min}(t) = 0, \tag{2}$$

where $\mathrm{Num}_{\mathrm{extream}}$ is the number of extreme points of the data sequence and $\mathrm{Num}_{\mathrm{cross}}$ is the number of zero crossings; $B_{\max}(t)$, $B_{\min}(t)$ are the upper and lower envelopes by cubic spline interpolation with the maximum and minimum points as the control points, respectively. Formula (1) represents the narrow-band constraint condition of the IMF, and formula (2) represents the local symmetry constraint condition. The process of EMD decomposition to obtain an IMF can be expressed as follows (Algorithm 1).

2.1.2. Modal Aliasing of EMD. However, the most significant drawback of EMD is modal aliasing, as shown in Figure 1. Figure 1(b) shows the FFT spectrum corresponding to each IMF in Figure 1(a). It can be seen from the figure that each FFT spectrum contains multiple signals of different frequencies, which means a single IMF contains signals of different frequencies or signals of the same frequency that appear in different IMF components. These are modal aliasing. The main reason for modal aliasing is the absence of extreme value or the inconsistent distribution of extreme value, which makes the variation trend error between the spectral envelope obtained by interpolation and the real signal is too large. At this time, the time-domain signal does not meet the narrow-band requirements of IMF decomposition, resulting in mode aliasing.

In order to solve this problem, mask signal $s(t)$ is usually created to compensate the missing extreme value, and then the values are given, respectively:

$$x_{+}(t) = xt + st, \tag{3}$$

$$x_{-}(t) = xt - st. \tag{4}$$

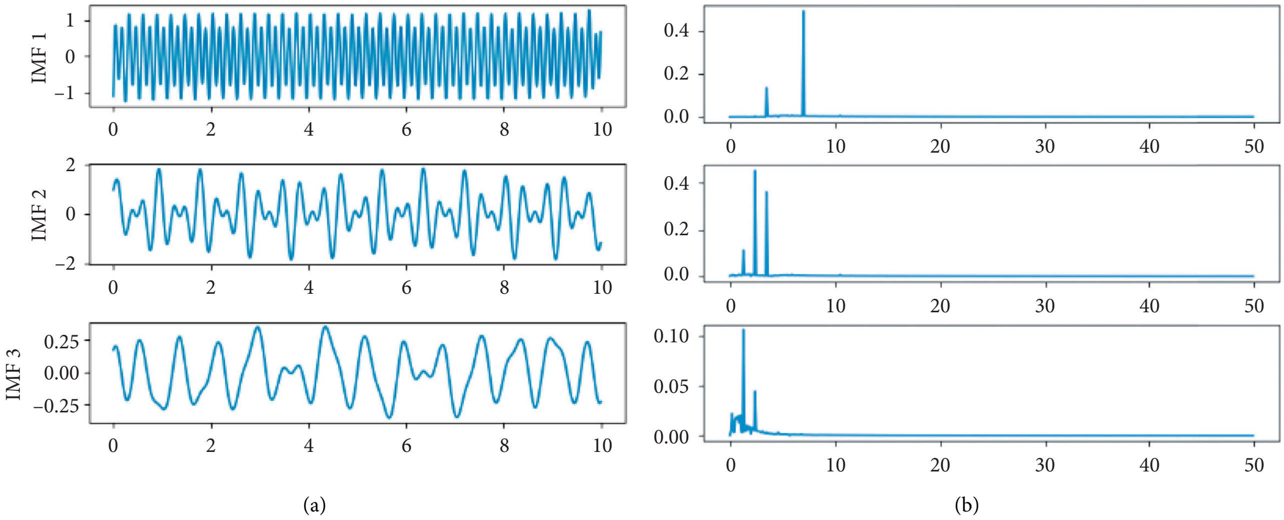
For $x_{-}(t)$ and $x_{+}(t)$, EMD is performed to obtain the natural mode functions $r_{\mathrm{IMF}-}(t)$, respectively. The final IMF is defined as follows:

$$r_{\mathrm{IMF}}(t) = \frac{r_{\mathrm{IMF}+}(t) + r_{\mathrm{IMF}-}(t)}{2}. \tag{5}$$

It can be seen from the above that the extreme value distribution of mask signal $s(t)$ is very important to solve the modal aliasing problem. White noise is usually used as mask signal $s(t)$, but this method does not make full use of the properties of the signal itself and cannot adapt to a variety of signal contents. Therefore, this paper proposes a heuristic mask for empirical mode decomposition method. This

```
Input: original signal $x\,(t)$, supposed IMF number $i$
Output: intrinsic mode functions, IMF
(1) $i = 1$, $x^1(t) = x\,(t)$.
(2) Get the extremum points $\{u_1^{\max}, u_1^{\min}, u_2^{\max}, \ldots\}$ of signal $x^i(t)$, calculate the upper and lower envelope $B_{\max}(t), B_{\min}(t)$ by cubic spline interpolation with the maximum and minimum points as control points, and get the average value of upper and lower envelope $B_{\text{mean}}(t)$ at every points.
(3) $r\,(t) = x^i(t) - B_{\text{mean}}(t)$. If $r\,(t)$ satisfies formulas (1) and (2), then $r\,(t)$ is taken as the $i$th IMF signal $r_{\text{IMF}}^i(t)$, $i = i + 1$; if not, repeat step 2 and 3 for signal $r\,(t)$.
(4) $x^i(t) = x^{i-1}(t) - r_{\text{IMF}}^{i-1}(t)$. Return to step 1 until the termination condition is satisfied.
```

Algorithm 1: Empirical mode decomposition.



(a)

(b)

Figure 1: EMD results for signal $x_1(t) = \sin 2\,\pi * 2.4t + \sin 2\,\pi * 3.5t + \sin 2\,\pi * 7t$: (a) IMFs; (b) FFT spectrum.

method makes full use of the structural attributes of the signal itself to construct variable analysis window and mask signals. The specific principle and implementation process are as follows.

2.2. Heuristic Mask Signals

2.2.1. Basic Principle Analysis. The signal properties need to be established prior to EMD. A time-varying FM/AM model can be used to express any nonstationary signal; that is,

$$x\,(t) = At\,\sin(\omega\,(t)), \tag{6}$$

where $a\,(T)$ is the envelope function and $\omega\,(T)$ is the phase function. The analytical signal is

$$z\,(t) = xt + jH\,[x\,(t)]. \tag{7}$$

Here, $H[\cdot]$ denotes the Hilbert transform. We calculate the instantaneous phase $\omega\,(t) = \arctan(H\,[x\,(t)]/x\,(t))$ and instantaneous frequency $f_{IF}t = (1/2\pi)(\mathrm{d}[\omega\,(t)]/\mathrm{d}t)$. Using Hilbert transform, we can separate the AM and FM components of the IMF to achieve the purpose of modal separation.

For the single component mode, the instantaneous frequency $f_{IF}t$ should be nearly linear, while the variation range of $\omega\,(t)$ should be considerably small. When mode aliasing occurs, $f_{IF}t$ should clearly change without consideration of the end points. Especially, for hidden components, a jump of $f_{IF}t$ occurs at the time point of concealment. We constructed a variable analysis window according to the time-frequency characteristics of instantaneous frequency. Then, we divided the signal into several parts.

If $f_{IF}t$ of the segmented signal is still unstable; then, the modal separation problem can be transformed into the $(\mathrm{d}[\omega\,(t)]/\mathrm{d}t)$ minimisation problem, in which the bandwidth of $\sin(\omega\,(t))$ is minimised. The bandwidth calculation method for nonstationary signals can be obtained by the Carson rule:

$$\mathrm{BW}_{\text{AM–FM}} = 2\Delta f + f_{\text{FM}} + f_{\text{AM}}, \tag{8}$$

where Δf is the deviation of the instantaneous frequency from its mean value and f_{AM} and f_{FM} denote the frequencies of the AM and FM signals, respectively. We can make $\Delta f = 0$ to minimise the bandwidth. In other words, the decomposition frequency of each IMF is expected to be equal to the centre frequency of the instantaneous frequency, that is, equal to the mean value of the instantaneous frequency $\overline{f}_{IF}t$. Then, a mask signal with the same frequency as $\overline{f}_{IF}t$ can be selected and the number of IMFs required can be determined.

2.2.2. Algorithm Description. The HM-EMD algorithm comprises the following steps: variable analysis window construction and mask signal construction.

(1) Variable Analysis Window Construction. The jump point t_i should be picked such that formula (9) is satisfied:

$$\left|f_{IF}t_i - f_{IF}t_{i+1}\right| + \left|f_{IF}t_{i-1} - f_{IF}t_i\right| > \mu_{\Delta f_{IF}t} + \rho\varepsilon_{\Delta f_{IF}t}, \tag{9}$$

where $\Delta f_{IF}t$ is the difference in instantaneous frequencies at t_i, $\mu_{\Delta f_{IF}t}$ is the mean value of $\Delta f_{IF}t$ at all time points, $\varepsilon_{\Delta f_{IF}t}$ is the variance, and ρ is the variable parameter. The original signal is divided into two parts by the time division points t_i and decomposed by EMD independently.

(2) Mask Signal Construction. The sine signal is a common form of a mask signal, and its amplitude and frequency should be determined. As analysed in Section 2.1, the frequency is determined as the average instantaneous frequency $\overline{f}_{IF}$. Hence, the amplitude is also determined as the mean value $\overline{A_{IF}}$ of the instantaneous amplitude. Then, the mask signal s t is defined as

$$st = \overline{A_{IF}} \sin 2\ \pi \overline{f}_{IF}t, \tag{10}$$

where $\overline{A_{IF}} = (1/n)\sum_{t=1}^{n}\sqrt{r_{IF}(t)^2 + H(r_{IF}(t)^2)}$ and $\overline{f}_{IF} = (1/n)\sum_{t=1}^{n}(\mathrm{d}/\mathrm{d}k)(\arctan(H(r_{IF}(t))/r_{IF}(t)))$.

Then, IMFs can be refreshed by formulas (3)–(5), in which the number of IMFs is determined by $\overline{f}_{IF}$ and f_c is the sampling frequency. The algorithm flow is as follows (Algorithm 2):

3. HM-EMD-Based Acoustic Scene Classification

The audio stream contains the hidden acoustic events that can represent the acoustic scene. In this section, HM-EMD is first used to decompose the acoustic scene signals, and the IMF of hidden acoustic events in these acoustic scene signals is analysed. According to the analysis results, a full-band IMF hidden component feature is proposed to represent the hidden acoustic events. Finally, the process of acoustic scene classification using these features is given in detail.

3.1. Acoustic Scene Signal Analysis by HM-EMD. When processing the original signal with HM-EMD, the variable analysis window and mask signal are used to intervene the decomposition of the original signal. The frame length is selected according to the frequency structure of the signal itself, while the frequency domain components corresponding to each IMF are relatively independent, which provides higher interpretability of the features. The instantaneous frequency and amplitude of each IMF also contain all information of IMF components, which means that the instantaneous frequency and amplitude of all IMF components contain most of the information of the signal to be analysed and can be directly used as the basic characteristics of the signal. Figure 2 shows the time-domain waveforms of some typical IMFs with hiding acoustic events in the ambient audio stream, in which only the most significant one of all IMF waveforms is shown. It can be seen that the time-domain waveform characteristics of these events are very obvious, the extreme value and over-average rate are very different, and they are distributed in low, medium, and high frequency bands. Therefore, this paper proposes a full-band IMF hiding component features, which can distinguish them well, to effectively improve the effect of ambient audio stream recognition algorithm. The feature calculation method is shown in Section 3.2.

3.2. Mutagenic Component Features. Figure 2 shows various hidden components in the acoustic scene data. On the one hand, the hidden components cause a significant interference to the signal spectrum, thereby greatly affecting the ambient audio stream recognition effect based on traditional spectrum features (such as MFCC). On the other hand, the types and characteristics of hidden components corresponding to different ambient audio streams also exhibit significant differences. These hidden components are closely related to the types of acoustic events. The features constructed on the basis of hidden components can help to distinguish ambient audio streams. For a hidden component, its frequency, amplitude, and change mode information can effectively reflect its essential attributes. Almost all of such information can be reflected by the envelope shape of the IMF obtained by decomposition. Therefore, we design a set of HACFs. Based on the IMF decomposed by HM-EMD, the features extract the relevant information of hidden components, including the shock intensity feature SH and over-average feature average crossing rate (ACR).

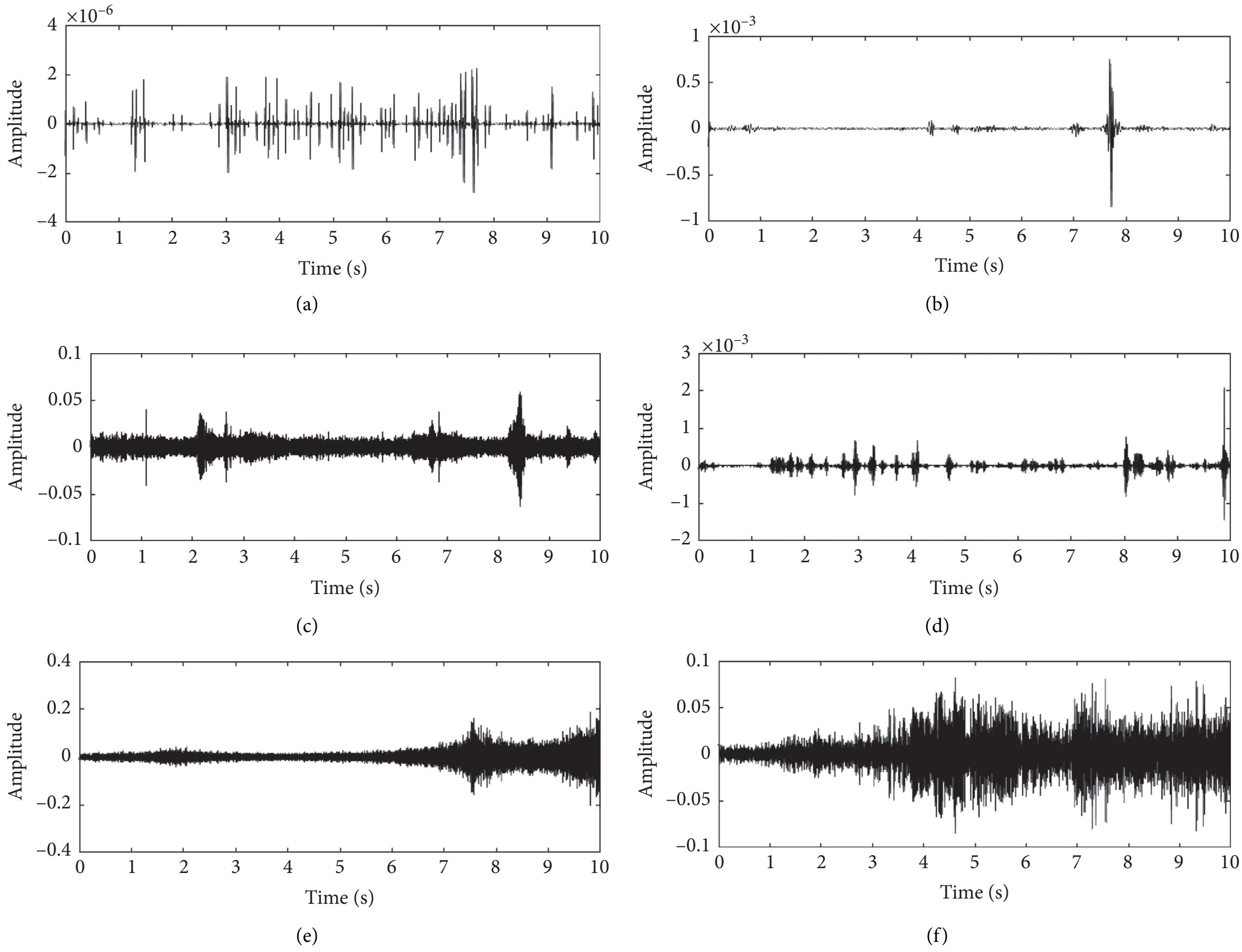
3.2.1. Shock Intensity Feature (SH).

$$\begin{aligned} \mathrm{SH}_{\max j} &= \max\left(r_{\mathrm{IMF}j}^{\mathrm{up}}(t)\right), \\ \mathrm{SH}_{\min j} &= \min\left(r_{\mathrm{IMF}j}^{\mathrm{up}}(t)\right), \end{aligned} \tag{11}$$

where max $(r_{\mathrm{IMF}j}^{\mathrm{up}}(t))$ is the upper limit of the signal amplitude in the jth IMF and min $(r_{\mathrm{IMF}j}^{\mathrm{up}}(t))$ is the lower limit. Both limits represent the change intensities of the hidden components relative to the steady components for measuring the changes in signal amplitude. As the sum of the mean values of the upper and lower envelopes of the IMF is 0, the signal is symmetrical along the time axis, and the information carried by the upper and lower envelopes is almost the same. Therefore, a one-sided envelope is enough to ensure the consistency of the symbols of the two values. The superscript means that the upper envelope is used for calculation.

```
Input: signal $x\ (t)$, supposed IMF number $i$
Output: intrinsic mode function, IMF
(1) $x_1\,(t) = x\ (t)$, $i = 1$.
(2) Get the first IMF of the signal residual $x_i\,(t)$, calculate the mean and variance of $\Delta f_{IF}t$, and use formula (8) to determine whether
    there is a hiding jump point. Variable analysis window is constructed according to the hiding jump point and $x_i\,(t)$ is segmented.
(3) Construct mask signal for each IMF$_i$: $s_i t = \overline{A_{IF_i}} \sin 2\ \pi \overline{f_{IF_i}} t$.
(4) Do EMD on $x_{i+}t = x_i\,(t) + s_i t$ and $x_{i\_}t = x_i\,(t) - s_i t$; get the first IMF $r_{\mathrm{IMF}_i+}(t)$ and $r_{\mathrm{IMF}_i-}(t)$.
(5) Let $r_{\mathrm{IMF}_i}(t) = (r_{\mathrm{IMF}_i+}(t) + r_{\mathrm{IMF}_i-}(t))/2$, and splice all the divided pieces.
(6) $i = i + 1$, $x_i\,(t) = x_{i-1}\,(t) - r_{\mathrm{IMF}_i}(t)$, return to step 2, until $\overline{f_{IF_i}} t < (f_c/2i)$, or no new IMF is required.
```

Algorithm 2: Heuristic empirical mode decomposition with a masking signal.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: IMF waveforms with significant hidden components in different environments of the audio stream. (a) Airport luggage roller: IMF16, low freq. (b) Metro rail joint collision: IMF14, low freq. (c) Chirm: IMF1, high freq. (d) Steps: IMF10, medium freq. (e) Vehicles from far to near: IMF1, high freq. (f) Tram acceleration: IMF4, medium and high freq.

3.2.2. ACR Feature.

$$\mathrm{ACR}_i = \frac{1}{2T}\sum_{i=2}^{T}\left|\mathrm{sgn}\left[r^{\mathrm{up}}_{\mathrm{IMF}j}(t) - \overline{r^{\mathrm{up}}_{\mathrm{IMF}j}(t)}\right] - \mathrm{sgn}\left[r^{\mathrm{up}}_{\mathrm{IMF}j}(t-1) - \overline{r^{\mathrm{up}}_{\mathrm{IMF}j}(t)}\right]\right|. \tag{12}$$

ACR features can express the number of times the upper envelope of an IMF passes through its mean point, that is, the number of times the IMF's upper envelope (time domain amplitude) fluctuates significantly. If the value is large, the IMF amplitude frequently fluctuates near the mean value. For ambient audio stream recognition application scenarios, if the value is greater than a certain threshold (10 Hz or above), the data may not have obvious and meaningful hidden components and the change of the upper envelope near the mean value is only the normal fluctuation of the acoustic signal itself. If the value is less than the threshold, the data may contain significant hidden components, and one-half of the zero-crossing frequency is the frequency of the hidden components.

3.3. Ambient Audio Stream Classification. The process of audio streams classification based on heuristic mask empirical mode decomposition is shown in Figure 3. Firstly, the HM-EMD method is used to decompose the signal into the IMFS set. Then, the basic features are extracted based on the IMFS: the instantaneous frequency, instantaneous amplitude, and the hidden features of AHCFs are all composed of the feature matrix, and the feature matrix is input into the classifier to get the final recognition result. In order to verify the validity of the feature, two kinds of classifiers are selected in this paper. One is a three-layer perceptron model, whose specific structure is shown in Figure 3. The model has a three-layer structure. Sigmoid function is used as activation function for the first two layers; each layer has 500 and 250 neurons, respectively. The output layer uses a SoftMax classifier and has 10 neurons. The second is the TridentRestNet model, which consists of three branches, each of which is ResNet101. The different branches have different convolutional kernels for bottleneck modules, which use $3 * 3$, $5 * 5$, and $7 * 7$, respectively, in order to obtain features at different scales. Finally, all the features are fused together to give the recognition result. The experimental results show that under the two model systems, the system based on HM-EMD features still shows satisfactory results. The specific experimental results and analysis are as follows:

4. Experiments and Results

In this section, we evaluate the performance of the proposed HM-EMD method for the validity of modal separation and the audio stream classification. First, we provide details on the experimental setup which include both evaluation criteria and datasets. Second, the indexes, the results of effectiveness analysis for modal separation and acoustic scene classification methods whose performance are compared with the proposed method are provided. Finally, we compare the performance of HM-EMD with that of the baseline methods based on the experiments which are conducted on the DCASE datasets and ASVSpoof2019 and analyse the experimental results in detail.

4.1. Experimental Setup. We verify the results of this work from two aspects: the validity of modal separation and the validity of the HM-EMD features for environmental audio stream classification.

4.1.1. Validation of Modal Separation. A nonlinearity index is defined in formula (13), and it measures the stability of the decomposition results. The larger the DN is, the greater the nonlinear degree is, indicating the more unstable components; the verification data are the mixed signals of the three modes in Figure 1:

$$\mathrm{DN} = \left[\frac{1}{n}\sum_{t=1}^{n}\left(\frac{f_{IF}t - \overline{f_{IF}t}}{\overline{f_{IF}t}}\right)^{2}\right]^{1/2}. \tag{13}$$

4.1.2. Validation of the Features of HM-EMD for the Classification of Audio Streams. To verify the effectiveness of designing a series of features based on HM-EMD, we use a basic HM-EMD feature matrix and a basic features + HACF matrix as the input parameters of the classifier. Specifically, the number of mask EMD reference IMFs is 20, the number of HM-EMD basic feature's dimension is 20, and HACFs is 3D, whose number of dimension is 20 × 3. The audio frame length is 0.5 s, the interframe overlap is 0.25 s, and the total number of dimensions is 39 × 20 × 3 = 2340. The classical mel frequency cepstral coefficients are selected as the contrast features; they include 13 dimensional MFCCs and delta features. The total number of dimensions is 39, and the audio frame length is 40 ms.

After setting the characteristic parameters, we conducted the test according to the process designed in Figure 3. There are two datasets used in our experiment:

(1) TASK1A dataset of DCASE [16]: the dataset contains data on ten cities and nine devices, that is, three real devices (A, B, C) and six simulated devices (S1–S6). The dataset has good annotation, including three different types of indoor, outdoor, and traffic. It also has ten different ambient audio streams, namely, airport, shopping mall, metro, metro station, pedestrian, street traffic, tram, park and public square, and bus. The acoustic data span a total of 64 h, with 40 h used in dataset training and with 24 h used in verification. Each audio segment is 10 s long, and the sampling rate is 44.1 kHz.

(2) ASVSpoof 2019 dataset [17], which is a dataset aiming to foster the research on countermeasure to detect voice spoofing in automatic speaker verification. The dataset contains synthesized and replayed speech attacks, which are classified as logical access and physical access respectively. There are three subsets under these two tracks, namely, training set, development set, and evaluation set. Actual voice data for both tracks were collected from 107 speakers collected from the VCTK2 database, 46 male speakers, and 61 female speakers. A subset of training and development of physical access was created by simulating room acoustics, including 3 room sizes, 3 reverberation levels, and 3 speaker

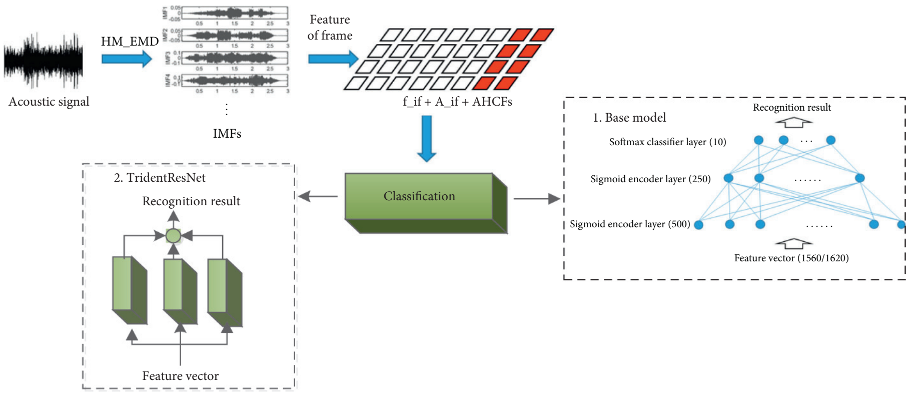


Figure 3: The process of audio streams classification based on heuristic mask empirical mode decomposition.

distances from the ASV microphone. In addition, there are nine recording configurations, with three recording distances to the speakers and three speakers of different qualities. Since this paper focuses on testing the feature of HM-EMD in the recognition of fake audio, the 10 neurons in the output layer of the two classifiers in Figure 3 are replaced with 2 neurons and the training the model again. At this time, the output result of the model is the detection results of fake audio. The results are evaluated by EER (equal error rate). Specific experimental results are shown in Section 4.2.

4.2. Results and Analysis

4.2.1. Effectiveness Analysis for Modal Separation. By comparing the traditional EMD results, we can see $\mathrm{DN}_{\text{HM-EMD}}$ $/\mathrm{DN}_{\text{EMD}} < 1$ for any given case. Hence, the IMF processed by the HM-EMD method has the lowest nonlinearity; that is, the IMF has a high purity and is close to the blind separation result under an ideal state. The separation result is shown in Figure 4. The features based on this high-purity IMF signal can effectively characterize the subtle changes in the signal components in the time and frequency domains. Hence, the method is suitable for all types of acoustic correlation analyses and recognition, especially for the recognition of ambient audio streams with hidden acoustic events.

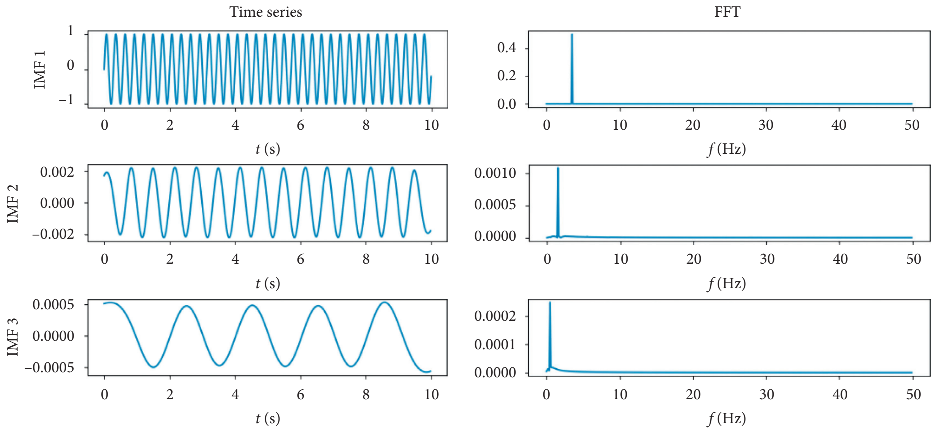
4.2.2. Based on HM Feature Validity of EMD

(1) Ambient Audio Stream Classification. HACFs can be used to identify the hidden components in IMFs and are thus of great significance for ambient audio stream recognition. We verified the discrimination ability of HACFs in different scenarios (Figure 5). The figure shows the scatter projection of some hidden component features in the three-dimensional space. Even the three-dimensional features in a single IMF have a strong scene discrimination ability. HACFs show good discrimination ability among different ambient audio stream categories and thus provide technical support for subsequent ambient audio stream classification.

Figures 6 and 7, respectively, show the acoustic streams classification and recognition results of MFCC features, HM-EMD basic features and HM-EMD basic features + AHCFS features based on simple classifier and complex classifier. The simple classifier is a simple three-layer perceptron, while the complex classifier adopts the optimal classifier in the DCASE competition [18]. As can be seen from the figures, basic environmental features based on HM-EMD in the simple classifier are better than MFCC features in most scenes, and AHCFS features can effectively capture hidden information in the environment, which is improving the accuracy of audio streams classification. In the complex classification model, the improvement of model classification ability can make up the deficiency of feature representation, and the overall recognition rate has increased.

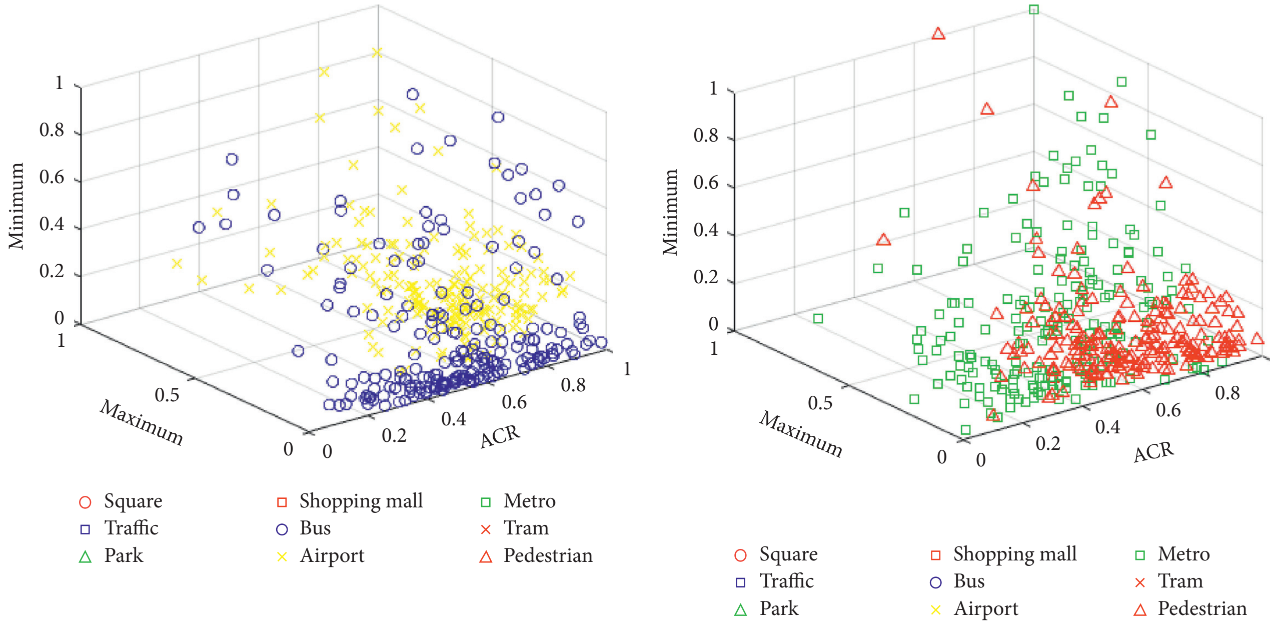
Tables 1 shows the results of acoustic streams classification. It can be seen that the HM-EMD feature is superior to the MFCC feature with different classifiers: given the basic classifier, $f_{IF} + A_{IF}$ is 6.7 percentage points higher than that in the MFCC series. After the addition of HACFs, the recognition rate increases by 17.4 percentage points. This result is close to the classification accuracy of the RESNET network with a 32M model size in the DCASE competition [19], while the simple model used in this paper is only 225K. In a complex classification model, the improvement of model classification can make up for the lack of features to some extent. However, in this case, $f_{IF} + A_{IF}$ + HACFs still improves the accuracy by 1.3%, and the recognition result reaches 75.7%. The $f_{IF} + A_{IF}$ feature can provide instantaneous characteristics in time-frequency domain and HACFs represent the statistical characteristics in time-frequency domain. The combination of the two can help improve the accuracy of the classification of environmental audio streams.

(2) Fake Audio Detection. Table 2 presents the fake audio detection results based on HM-EMD features. It can be seen that the forged audio based on LA is easier to be identified than the forged audio based on PA, and the HM-EMD feature has a more effect on the fake audio of LA attack. After adding the hidden information feature AHCF, the detection error rate of forged audio was reduced by 5.61% and 6.11%, respectively. This is because the LA



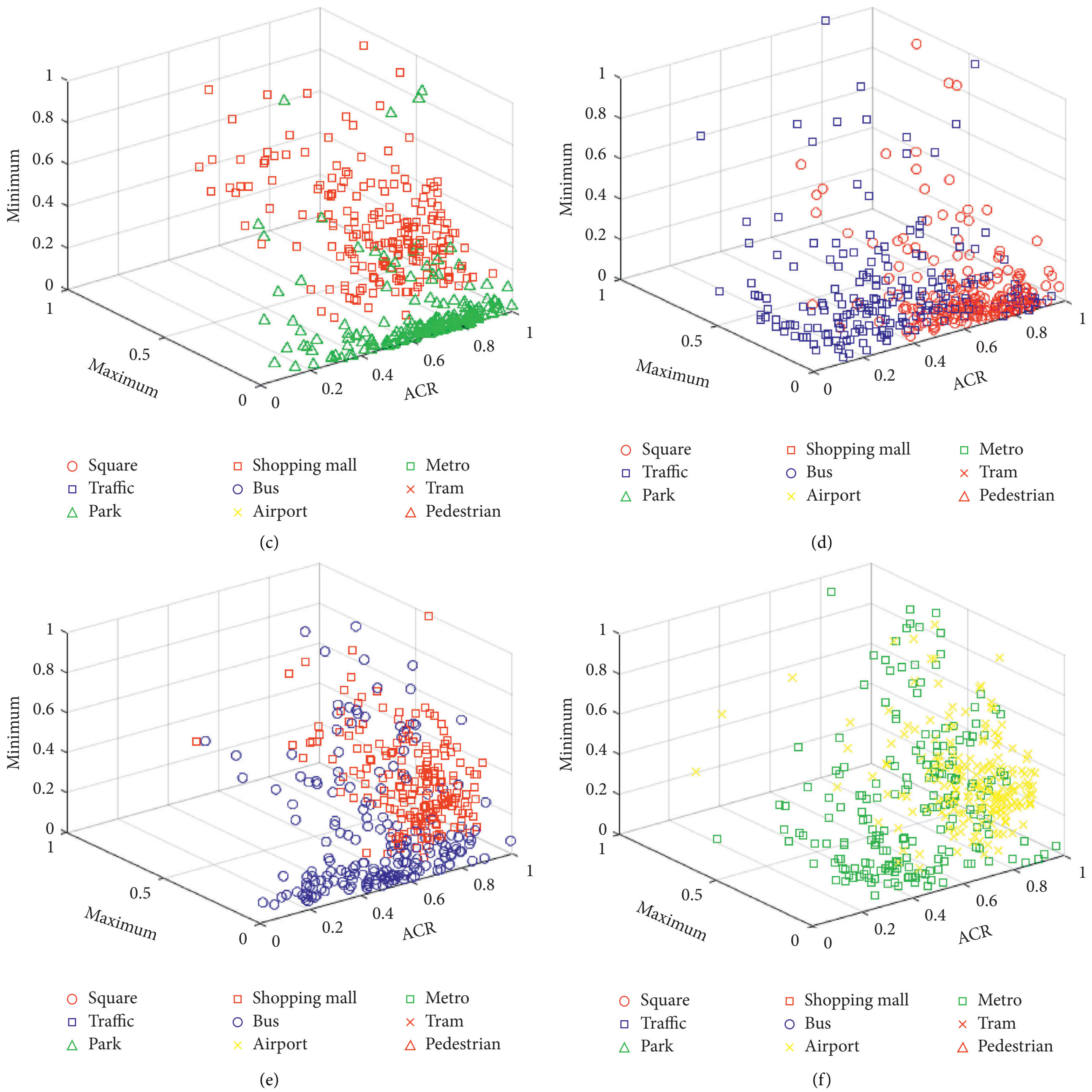
(a) (b)

Figure 4: HM-EMD results of $x_1(t)$.



(a) (b)

Figure 5: Continued.



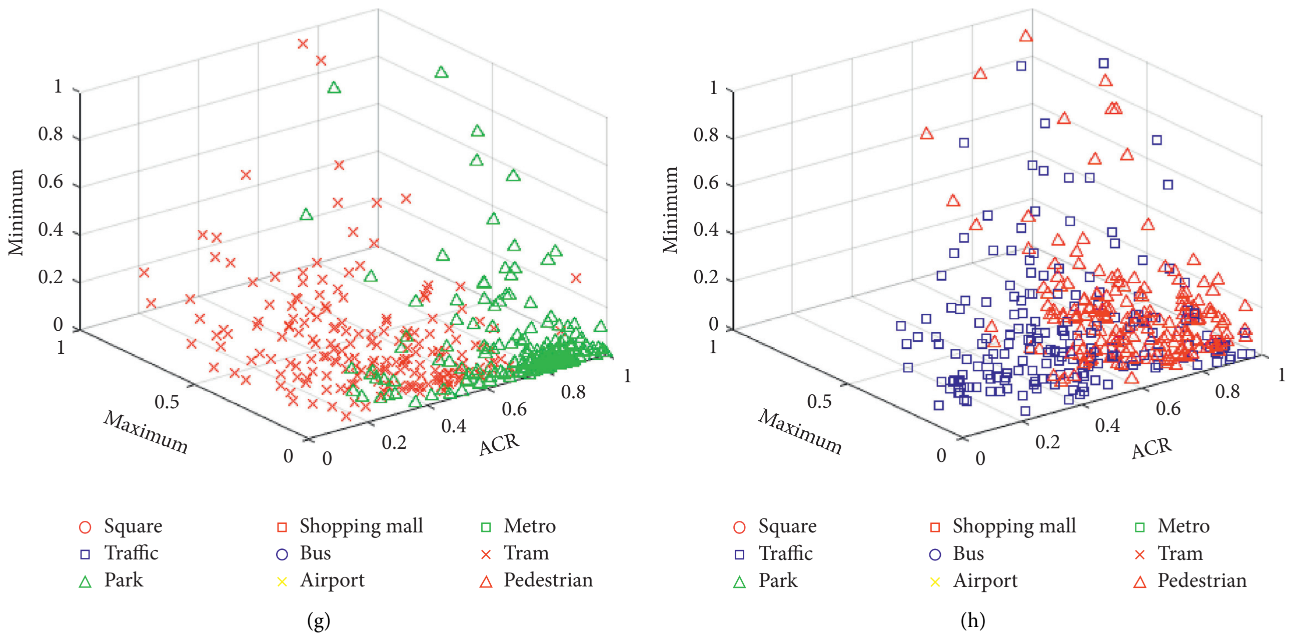
(c)

(d)

(e)

(f)

Figure 5: Continued.



(g)

(h)

Figure 5: HACFS feature distribution in 3D space. (a) IMF1 of bus and airport. (b) IMF1 of metro and pedestrian. (c) IMF1 of park and shopping mall. (d) IMF1 of square and traffic. (e) IMF2 of bus and shopping mall. (f) IMF2 of metro and airport. (g) IMF2 of tram and park. (h) IMF2 of traffic and pedestrian.

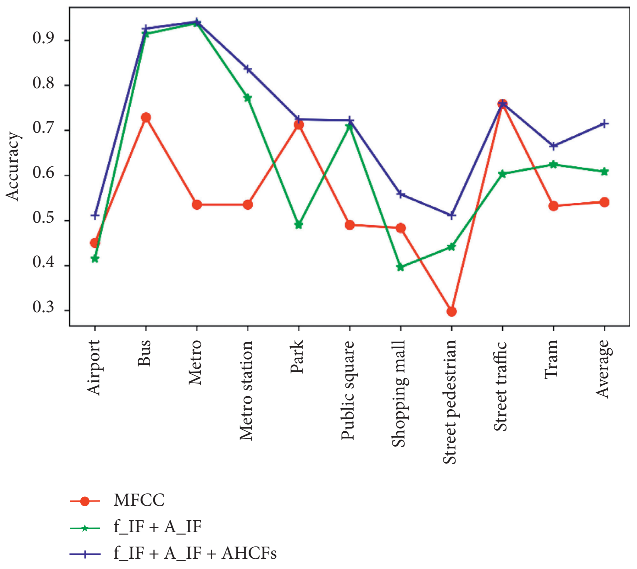


Figure 6: Acoustic streams' classification results of DCASE based on the simple classifier.

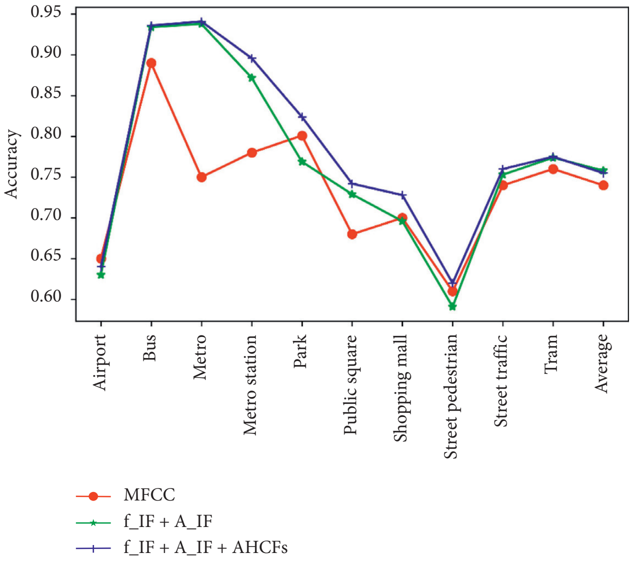


Figure 7: Acoustic streams' classification results of DCASE based on TridentresNet.

ignores background sound in the process of synthesizing the fake audio. In this case, the addition of captured audio background features greatly reduces the detection error rate. It can also be seen from the table that the HM-EMD feature can reduce the error rate of fake audio detection in both simple and complex models, which also proves the effectiveness of the feature extraction method proposed in this paper.

Table 1: Average results of acoustic streams classification based on DCASE dataset.

Classifier	MFCC (accuracy %)	$f_{IF} + A_{IF}$ (accuracy %)	$f_{IF} + A_{IF}$ + AHCFs (accuracy %)
TridentResNet_DevSet	73.7	74.5	75.0
TridentResNet_EvalSet	73.7	74.5	75.0
TridentResNet_Ensemble	74.2	75.0	75.5
TridentResNet_Weighted_Ensemble	74.4	75.2	75.7
3L neural network	54.1	60.8	71.5

Table 2: Fake audio detection results based on ASVSpoof 2019 dataset.

Classifier	Attack type	MFCC (EER %)	$f_{IF} + A_{IF}$ (EER %)	$f_{IF} + A_{IF}$ + AHCFs (EER %)
TridentResNet	LA	7.06	2.82	1.45
	PA	9.55	8.51	7.14
3L neural network	LA	9.95	4.98	3.84
	PA	11.83	11.79	9.44

5. Conclusions

Aiming at the problem of audio fraud existing in the network, this paper proposes a method of feature extraction of hidden information in audio streams based on HM-EMD. Because the audio background information is difficult to be forged, the basic features of audio streams and the characteristic HCFS of hidden information are constructed for fake audio streams based on stable IMFs decomposed by the HM-EMD method. The experimental results show that the HM-EMD-based features have richer characterization ability for hidden acoustic events than MEL cepstrum features and can improve the accuracy of scene classification and fake audio detection. However, since the HM-EMD decomposition process needs to calculate the mask signal according to the structure of the signal itself and use the mask signal to separate the aliasing signal, the algorithm complexity is increased compared with the classical EMD algorithm. Therefore, in the subsequent work, we will consider the idea of coevolutionary framework to optimize the algorithm [20]. At same time, the relationship between the HM-EMD feature system and different hidden acoustic events will be the further exploration point, so as to achieve accurate hidden acoustic event markers in audio streams of different levels and time scales. In general, the feature extraction of audio streams based on HM-EMD is helpful to detect fake audio and provides a new research idea for solving network audio spoofing.

Data Availability

The data used in this study are the public dataset from DCASE challenge (http://dcase.community/challenge2020/task-acoustic-scene-classification and https://datashare.ed.ac.uk/handle/10283/3336).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Jiu Lou and Decheng Zuo conceived and designed the study. Zhongliang Xu performed the simulations. Hongwei Liu reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was in part supported by the National Key Research and Development Program (2018YFC0830602).

References

[1] A. Jati, C.-C. Hsu, M. Pal et al., "Adversarial attack and defense strategies for deep speaker recognition systems," *Computer Speech & Language*, vol. 68, 2021.

[2] X. R. Li, S. L. Ji, C. M. Wu et al., "Survey on deepfakes and detection techniques," *Journal of Software*, vol. 32, no. 2, pp. 496–518, 2021.

[3] T. Chen, A. Kumar, P. Nagarsheth et al., "Generalization of audio deepfake detection," in *Proceedings of the Odyssey 2020 the Speaker and Language Recognition Workshop*, pp. 132–137, Tokyo, Japan, November 2020.

[4] R. K. Das, J. C. Yang, and H. Z. Li, "Long range acoustic features for spoofed speech detection," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1058–1062, Graz, Austria, September 2019.

[5] M. Todisco, X. Wang, V. Vestman et al., "ASVSpoof 2019: future horizons in spoofed and fake audio detection," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1008–1012, Graz, Austria, September 2019.

[6] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients," in *Proceedings of the Odyssey 2016 the Speaker and Language Recognition Workshop*, pp. 283–290, Bilbao, Spain, June 2016.

[7] R. K. Das, J. C. Yang, and H. Z. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *Proceedings of the 2019 IEEE Automatic Speech Recognition and*

Understanding Workshop, pp. 1018–1025, Singapore, December 2019.

[8] Z. Z. Wu, T. Kinnunen, E. S. Chng et al., "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proceedings of the 2012 Conference Handbook—Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, December 2012.

[9] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez et al., "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1068–1072, Graz, Austria, September 2019.

[10] H. Zeinali, T. Stafylakis, G. Athanasopoulou et al., "Detecting spoofing attacks using VGG and SincNet: BUT-omilia submission to ASVspoof 2019 challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1073–1077, Graz, Austria, September 2019.

[11] R. J. Li, M. Zhao, Z. Li et al., "Anti-spoofing speaker verification system with multi-feature integration and multi-task learning," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1048–1052, Graz, Austria, September 2019.

[12] S. H. Mankad, S. Garg, M. Patel, and H. Adalja, "Investigating feature reduction strategies for replay antispoofing in voice biometrics," in *Proceedings of the 8th International Conference on Pattern Recognition and Machine Intelligence*, Tezpur, India, December 2019.

[13] S. H. Mankad and S. Garg, "On the performance of empirical mode decomposition-based replay spoofing detection in speaker verification systems," *Progress in Artificial Intelligence*, vol. 9, no. 4, pp. 325–339, 2020.

[14] R. Deering and J. F. Kaiser, "The use of a masking signal to improve empirical mode decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 485–488, Philadelphia, PA, USA, March 2005.

[15] H. Li, Z. Li, and W. Mo, "A time varying filter approach for empirical mode decomposition," *Signal Processing*, vol. 138, pp. 146–158, 2017.

[16] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," Technical report, Tampere University of Technology, Tampere, Finland, 2018.

[17] X. Wang, Y. S. Junichi, T. Massimiliano et al., "ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech," Technical Report, Cornell University, Ithaca, NY, USA, 2020.

[18] T. Heittola, A. Mesaros, and T. Virtanen, "TAU urban acoustic scenes 2020 mobile, development dataset," *Tampere University, Tampere, Finland*, 2020, Technical Report.

[19] H. J. Shim, J. H. Kim, J. W. Jung et al., "Audio tagging and deep architectures for acoustic scene classification: Uos submission for the DCASE 2020 challenge," 2020, http://dcase.community/documents/challenge2020/technical_reports/DCASE2020_ Shim_ 120.pdf.

[20] W. Deng, S. F. Shang, X. Cai et al., "Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization," *Knowledge-Based Systems*, vol. 224, pp. 1–14, 2021.