

# Environmental Sound Recognition Using Double-Level Energy Detection

Xiaoxia Zhang, Ying Li

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China.  
Email: 641868500@qq.com

Received April, 2013.

## ABSTRACT

The performance of classic Mel-frequency cepstral coefficients (MFCC) is unsatisfactory in noisy environment with different sound sources from nature. In this paper, a classification approach of the ecological environmental sounds using the double-level energy detection (DED) was presented. The DED was used to detect the existence of the sound signals under noise conditions. In addition, MFCC features from the frames which were detected the presence of the sound signals by DED were extracted. Experimental results show that the proposed technology has better noise immunity than classic MFCC, and also outperforms time-domain energy detection (TED) and frequency-domain energy detection (FED) respectively.

**Keywords:** Ecological Environmental Sounds; Double-Level Energy Detection; Time-Domain Energy Detection; Frequency-Domain Energy Detection; Mel-Frequency Cepstral Coefficients

## 1. Introduction

The sound recognition is a fundamental problem of the sound signals processing. It has important applications in many fields like media search [1], military [2], security supervision [3] etc. Sound data contains a wealth of useful information. Through the recognition and analysis of sound, we can get lots of environmental characteristics, such as climate, geography, time, species, etc. Sound can get messages that the vision cannot capture. Moreover, the sound can be obtained anytime and anywhere, it is not limited to the light, and it is not necessary within the field of vision. The required storage space is smaller than that of the video signals. The sound data has many advantages.

In the practical application of sound, the sound sources are not clearly known, which lead to designing an appropriate sound signals detection method becomes more difficult. Energy detection does not need to know a priori knowledge of the sound signals and it is easy to implement. Therefore, the energy detection has a greater advantage in this case. The time-domain energy detection (TED) runs faster, but the detection accuracy is not well, while the frequency-domain energy detection (FED) has higher detection accuracy, but runs more slowly. So we construct the double-level energy detection (DED) by combining the respective advantage of TED and FED. Under the condition of guaranteeing certain detection accuracy, this method is simple, effective and has lower

complexity than the separate use the time-domain or frequency-domain energy detection.

The Mel-Frequency Cepstral Coefficients (MFCC) is the most common feature used in many sound recognition systems [4]. The MFCC feature fully considers the characteristics of human hearing, which has a good performance in recognition. When MFCC is used to analyze the sound signals with flat-spectrum noise, the effect is not good, so the classification results of the MFCC decrease significantly in the background noise. To solve this abuse, we combine DED with MFCC. This method has a certain degree of noise immunity, and the feature vector is denoted as DED\_MFCC. As the model created by SVM shows more robust, we extract the DED\_MFCC to train the SVM classifier.

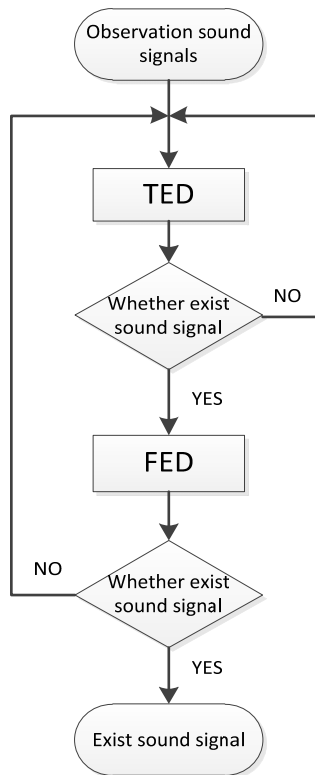
This paper is organized as follows. Section 2 presents the principles of DED. Section 3 introduces the feature extraction process and section 4 introduces the classification approach. In section 5, the experimental setup and the achieved results are presented. Finally, the conclusion of our work is given in section 6.

## 2. Double-Level Energy Detection

The energy detector is a kind of ideal signals detection tools, which can detect the existence of the sound signals in the noise environment. At present, energy detection can be divided into time-domain energy detection (TED) and frequency-domain energy detection (FED). These

two methods have their own respective strengths and weaknesses [5]. TED has the advantages including relatively simple, short running time, but disadvantage of low accuracy of the sound signal detection. Because of the discrete fast Fourier transform (FFT), FED can improve the application flexibility and accuracy of detection, but the running speed slows down.

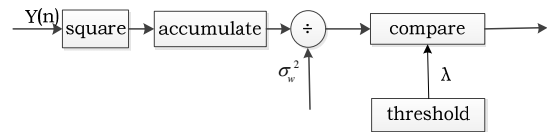
Given the advantages and disadvantages of the two methods described above, we use TED to detect the observation sound signals firstly, if there are no sound signals being detected, indicating that the sound is the noise; otherwise, we use FED to detect the observation sound signals, if it cannot detect the sound signals, it can be concluded that the observation sound signals do not contain sound signals, on the contrary, we can determine the observation sound signals contain the sound signals. The flow of DED is illustrated in **Figure 1**. Under the condition of guaranteeing certain detection accuracy, this method is simple, effective and has lower complexity than the separate use the time-domain or frequency domain energy detection.



**Figure 1. Flow of DED.**

## 2.1. Time-Domain Energy Detection

The principle of time-domain energy detection [6-8] is shown in **Figure 2**. Where  $Y(n)$  is the observation vector,  $\sigma_w^2$  is the noise variance,  $\lambda$  is the threshold that is set for a specific probability of false alarm (PFA).  $Y(n)$  goes



**Figure 2. Time-domain energy detection.**

through the operations of modulus square and accumulation as:

$$T = \sum_n (Y[n])^2 \quad (1)$$

The judgment formula in **Figure 1** is

$$\begin{matrix} H_1 \\ \frac{T}{\sigma_w^2} > \lambda \\ H_0 \end{matrix} \quad (2)$$

if  $\frac{T}{\sigma_w^2} > \lambda$ , the sound signals exist; if  $\frac{T}{\sigma_w^2} < \lambda$ , the sound signals do not exist.

In Equation (2),  $H_0$  means the sound signals do not exist, while  $H_1$  means the observation vector contains the sound signals. So the entire detection process is the hypothesis of a binary test:

$$Y(n) = \begin{cases} W(n) & H_0 \\ S(n) + W(n) & H_1 \end{cases} \quad (3)$$

where  $n = 1, \dots, N$ ,  $N$  is the number of samples.

This work uses a pre-given probability of false alarm (PFA), and the test statistic can be approximated by a Gaussian distribution:

$$\begin{aligned} H_0: T &\sim \text{Normal}(N\sigma_w^2, 2N\sigma_w^4), \\ H_1: T &\sim \text{Normal}(N(\sigma_w^2 + \sigma_s^2), 2N(\sigma_w^2 + \sigma_s^2)^2) \end{aligned} \quad (4)$$

where  $N$  is the number of samples (detection time, its value equals to the length of the frame),  $\sigma_w^2$  is the noise variance,  $\sigma_s^2$  is the sound signals variance.

When the signals do not exist, through the known probability of false alarm (PFA), we can obtain the threshold of the judgment. In the case of  $H_0$ ,  $T$  is in line with the Gaussian distribution, so the PFA is:

$$PFA = P\left(\frac{T}{\sigma_w^2} > \lambda \mid H_0\right) \quad (5)$$

then

$$Q\left(\frac{\lambda - N}{\sqrt{2N}}\right) = PFA \quad (6)$$

where  $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{y^2}{2}\right) dy$ .

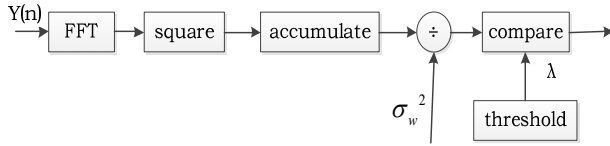
Therefore, through the given  $N$  and  $PFA$ , the threshold can be obtained by Equation (6):

$$\lambda = N + \sqrt{2NQ^{-1}(PFA)} \quad (7)$$

It can be seen from the above analysis, for the given PFA and the noise variance, we can calculate the judgment threshold, and then we can conclude that which frame contains the sound signals by Equation (2).

## 2.2. Frequency-Domain Energy Detection

The principle of frequency-domain energy detection is shown in **Figure 3**.



**Figure 3. Frequency-domain energy detection.**

Compared to the time-domain energy detection, frequency-domain energy detection firstly puts the observation vector through the FFT module to transform the time-domain signals into the frequency-domain signals. Then get the frequency-domain energy by putting the frequency domain signals through the modules of squaring and accumulating. Finally, compare the value of the frequency domain energy which is divided by noise variance with the threshold, which determines whether there are sound signals. If the probability of TED is  $P$ , when the observation signals contain the sound signals actually, the omission probability is  $1-P$  after the comparison with the threshold. If the TED determines that there are the sound signals in the observation signals, then start FED to detect the unknown sounds.

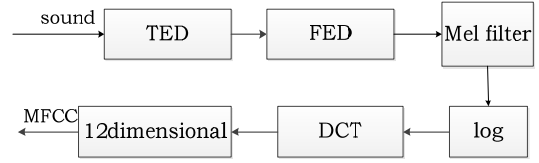
## 3. Extraction of DED\_MFCC

MFCC analyzes sound signals from the perspective of the human ear frequency level of the nonlinear psychological sense. It uses a nonlinear Mel-frequency scale to simulate the human auditory system. Values of the Mel-frequency scale are roughly logarithmic with the linear frequency, and more in line with the human auditory characteristics.

The calculation of MFCC parameters is based on the frequency reference of “bark” [9]. The relationship with the frequency conversion is:

$$f_{mel} = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \quad (8)$$

The energy detection method does not have the pre-processing of pre-emphasis. We simply divide the sound signals into frames and inter-frame without overlap, which thereby greatly enhances the efficiency of running. In this paper, the process of DED\_MFCC extraction is shown in **Figure 4**.



**Figure 4. Extraction of DED\_MFCC.**

The steps of extracting DED\_MFCC are as follows:

- 1) The input sound is divided into successive frames, 256 samples per frame, inter-frame without overlap.
- 2) Each frame is coupled with the Hanning window, then discard the frames without the sound signals though the TED.
- 3) Fast Fourier Transform is applied to the frame, then use the FED to discard the frames which is the false positives by the TED.
- 4) The energy spectrum generated from the FED passes through a set of the triangular Mel-scale filter bank, and the output is  $m(l), l = 1, 2, \dots, L$ .  $L$  is set to 24. The span of each triangular filter in the Mel-scale is equal, and it is set to 112Mel.
- 5) Take the logarithm of all of the filter output, then apply the Discrete Cosine Transform (DCT) to get a group of DED\_MFCC:

$$ded\_mfcc(j) = \sum_{l=1}^L \log m(l) \cos \left\{ \left( l - \frac{1}{2} \right) \frac{j\pi}{L} \right\}, 0 \leq j \leq L \quad (9)$$

In this paper, we use the first 12 coefficients as DED\_MFCC.

## 4. SVM Classification Algorithm

The support vector machine (SVM) is first proposed by Cortes and Vapnik etc. It shows many unique advantages in solving the problems of the small samples, nonlinear and high dimensional pattern recognition. SVM is built on the basis of VC dimension theory and structural risk minimization of the statistical learning theory. The basic principle is to correctly separate the two sample points in the separating hyperplane, and to maximize the minimum distance of the plus or minus class samples to the separating hyperplane.

Assume that  $\{x_1, x_2, \dots, x_1\}$  is the training sample,  $\{y_1, y_2, \dots, y_1\}$  is the corresponding class label. SVM finds the separating hyperplane with the largest interval by solving the following quadratic programming problem:

$$\begin{aligned} \min_{(w,b,\xi)} & \frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \\ \text{s.t.} & y_i [w^T x_i + b] - 1 + \xi_i \geq 0 \\ & \xi_i \geq 0, i = 1, 2, 3, \dots, n \end{aligned} \quad (10)$$

where  $C$  is a coefficient that regulates the value between the misclassification and the robustness of the classifica-

tion (width of margin).

Many linearly inseparable problems in the real world can be converted into linearly separable by mapping to high-dimensional space with the SVM kernel function [10]. At present, the most commonly used kernel functions include linear kernel, polynomial kernel, RBF kernel and sigmoid kernel. In the experiment of this paper, we use the LIBSVM package which is designed by DR. Lin Zhiren of Taiwan University.

## 5. Experiments

### 5.1. Experimental Setup

The sounds of the ecological environment which are used in these experiments are the variety of birds singing. There is a total of 12 kinds of birds singing here including flour chicken, Zhu turtledove, Dong chicken, male thrush, blackwater chicken, hair chicken, mother partridge, mountain turtledove, water rails, white-eye, the mother pheasant, mother bamboo chicken. There are 20 samples of each kind of birds singing, of which 10 samples are used for training and 10 samples for testing, so a total of 240 sound samples. These bird singings were recorded by voice recorder in the outdoors, and the length of each sound sample is more than two seconds, the sampling rate is 44100 Hz. In this work, the signal to noise ratio (SNR) of the sounds which is used to train the SVM models is 60 dB in the training step, which is done by adding the noise at 60 dB SNR to the clear sound data. In the testing step, we use the sounds with different SNRs. The extracted features include classic MFCC, TED\_MFCC (use the time-domain energy detection only), FED\_MFCC (use the frequency-domain energy detection only) and DED\_MFCC, and they are all 12-dimensional feature vectors. The kernel function used in SVM is RBF kernel. The PFA was set to  $10^{-8}$ .

### 5.2. Evaluation Results

In order to observe the noise immunity of the DED\_MFCC feature on the ecological environment sounds classification, we use SVM to construct the classification models based on MFCC, TED\_MFCC, FED\_MFCC and DED\_MFCC respectively. We use 30 frames of each sample for training, and 256 frames of each sample for testing. Here, the noise we add to the clear bird singing is the white Gaussian noise. The classification accuracy of the test samples corresponding to the different SNRs is shown in **Table 1**. The results of the experiments show that, in the case of SNR 50 dB and above, the four kinds of features do not have big difference. But with the enhancement of the noise, the MFCC has a sharp decline in the recognition rate. Compared to MFCC, TED\_MFCC has an improvement up to about 25%, while FED\_MFCC

**Table 1. classification results under the white Gaussian noise.**

feature	Signal Noise Ratio (SNR)					
	60 dB	50 dB	40 dB	30 dB	20 dB	10 dB
MFCC	100.00	90.83	68.33	45.83	37.50	24.17
TED_MFCC	100.00	100.00	91.67	78.33	67.50	43.33
FED_MFCC	100.00	100.00	91.67	85.83	71.67	61.67
DED_MFCC	100.00	100.00	91.67	85.83	71.67	61.67

and DED\_MFCC up to about 35%. Therefore, it can be seen from the experiments that FED\_MFCC and DED\_MFCC show better robustness in a noisy environment.

In order to demonstrate DED\_MFCC not only has a higher recognition rate but also lower time complexity, here we study the time complexity of the feature extraction of the test samples and classification. The time complexity means the running time of the program, the unit is seconds(s). Due to using the same amount of training data, the time complexity of training the model is the same. In these experiments, the training step is the same, and we use all of the 240 samples for testing and classification. The result is the average of 10 experiments, and it is shown in **Figure 5** and **Figure 6**. From **Figure 5**, we can see that the time complexity of MFCC feature extraction is higher than that of FED\_MFCC, which is because FED\_MFCC does not have the pre-emphasis, while MFCC needs it and MFCC extracts feature of all the frames; TED\_MFCC and DED\_MFCC reduce the number of FFT, so their time complexity of feature extraction is lower than that of FED\_MFCC. From **Figure 6**, we can see that the time complexity of MFCC classification is higher than that of the others, which is because MFCC classifies all the frames, while the others only classify the frames through the time-domain or frequency-domain energy detection; the detected accuracy of the FED is higher than that of the TED, that means the number of frames detected by the FED is less than the number of frames detected by the TED, so the classification time complexity of FED\_MFCC and DED\_MFCC is lower than that of TED\_MFCC. From **Figure 5** and **Figure 6**, as SNR decreases, the number of the frames which are detected is also reduced. So the time complexity of feature extraction and classification is reduced too. Therefore, DED\_MFCC has a lower time complexity of feature extraction and classification with the advantages of the TED\_MFCC and the FED\_MFCC.

In order to observe the performance of the method in the natural ambient noise, we use the sound of brook which was recorded in the outdoors instead of the white Gaussian noise. That means the clear bird singings were added with the sound of water with different SNRs. The

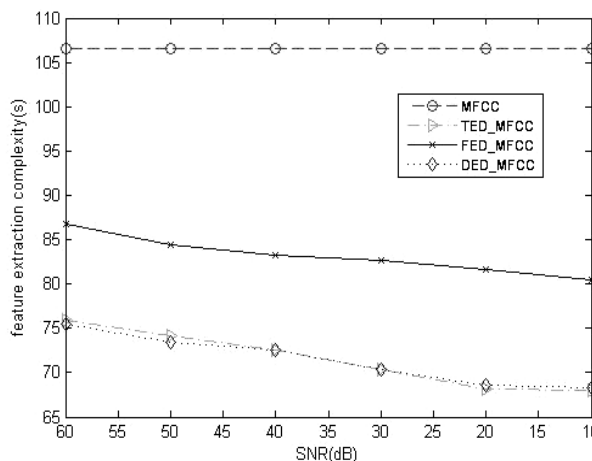


Figure 5. Feature extraction time complexity under the Gaussian white noise.

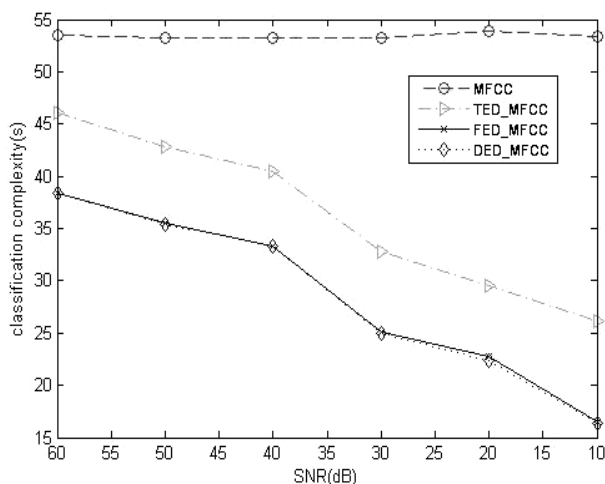


Figure 6. Classification time complexity under the Gaussian white noise.

results are shown in **Table 2**. From these experiments, we can see that this method is applicable to not only the white Gaussian noise in the simulative laboratory conditions, but also the noise of the natural environment. And when the SNR is below 30 dB, the classification results are even better than those of the white Gaussian noise.

## 6. Conclusions and Future Work

In this paper, we present a feature extraction method of MFCC based on double level energy detector. The experimental results show that the classification result is slightly better than the classic MFCC in the case of the high SNRs; but in the case of low SNRs, this method has good robustness, and the classification accuracy has greatly improved compared with the classic MFCC. In terms of the time complexity, the proposed method combines the advantages of the time-domain energy detection and the frequency-domain energy detection, so it has

Table 2. classification results under the noise of the natural environment.

feature	Signal Noise Ratio (SNR)					
	60dB	50dB	40dB	30dB	20dB	10dB
MFCC	100.00	91.67	80.83	49.17	39.17	24.17
TED_MFCC	100.00	100.00	91.67	78.33	70.83	61.67
FED_MFCC	100.00	100.00	91.67	85.33	73.33	65.83
DED_MFCC	100.00	100.00	91.67	85.83	73.33	65.83

lower time complexity in both the process of feature extraction and classification. In addition, it performs better in the natural ambient noise than in the Gaussian white noise. There are two disadvantages in the proposed method: first of all, the recognition rate is still less than ideal in the case of low SNRs; second, the sound of the ecological environment is limited to birds singing. In the future works, we will look for new features which have better noise immunity combined with this proposed method to improve the classification effect, and cover more kinds of ecological environmental sounds.

## 7. Acknowledgements

This work is supported by the National Natural Science Fund Project (No. 61075022).

## REFERENCES

- [1] R. Typke, F. Wiering and R. Veltkamp, "A Survey of Music Information Retrieval Systems," *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, 11-15 September 2005, pp. 153-160.
- [2] L. Gerosa, G. Valenzise and M. Tagliasacchi, *et al.*, "Scream and Gunshot Detection in Noisy Environments," *Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007)*, Poznan, 3-7 September 2007, pp. 1216-1220.
- [3] C. Zieger, A. Bruti and P. Svaizer, "Acoustic Based Surveillance System for Intrusion Detection," *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)*, Genoa, 2-4 September 2009, pp. 314-319. [doi:10.1109/AVSS.2009.49](https://doi.org/10.1109/AVSS.2009.49)
- [4] A. Dufaux, "Detection and Recognition of Impulsive Sounds Signals," Institute de Microtechnique Neuchatel, Switzerland, 2001.
- [5] M. Q. Wu, W. Ma and C. X. Xu, "A Low-Power Algorithm of Uniting the Time-Domain and Frequency-Domain Thresholds," China Patent, No. 101848044, 2010.
- [6] H. Urknowitz, "Energy Detection of Unknown Deterministic Signals," *Proceedings of the IEEE*, Vol. 55, No. 4, 1967, pp. 523-531. [doi:10.1109/PROC.1967.5573](https://doi.org/10.1109/PROC.1967.5573)

- [7] L. Vergara, J. Moragues, J. Gosalbez, *et al.*, "Detection of Signals of Unknown Duration by Multiple Energy Detectors," *Signal Processing*, Vol. 90, No. 2, 2010, pp. 719-726. [doi:10.1016/j.sigpro.2009.08.007](https://doi.org/10.1016/j.sigpro.2009.08.007)
- [8] S. M. Kay, "Fundamentals of Statistical Signal Processing: Detection Theory," 1st Edition, New Jersey: Prentice-Hall, 1998.
- [9] Y. Li, "A Quick Classification for Area Environmental Audio Data Based on Local Search Tree," *Proceedings of the 2009 International Conference on Environmental Science and Information Application Technology (ESIAT 2009)*, Wuhan, China, 4-5 July 2009, pp. 569 -574. [doi:10.1109/ESIAT.2009.15](https://doi.org/10.1109/ESIAT.2009.15)
- [10] V. David and A. Sánchez, "Advanced Support Vector Machines and Kernel Methods," *Neurocomputing*, Vol. 55, No. 1-2, 2003, pp. 5-20. [doi:10.1016/S0925-2312\(03\)00373-4](https://doi.org/10.1016/S0925-2312(03)00373-4).