Ecological Environmental Sounds Classification Based on Genetic Algorithm and Matching Pursuit Sparse Decomposition

Ming Li， Ying Li
College of Mathematics and Computer Science
Fuzhou University
Fuzhou, China

***Abstract*—The Mel-frequency cepstral coefficients (MFCCs) based on human auditory characteristics are widely used for audio recognition. However, the performance of MFCC-based audio recognition degrades due to noise interference. In consideration of this, we propose the matching pursuit (MP) sparse representation algorithm based on genetic algorithm (GA) improved by elite strategy and evolution reversal to accomplish the task of filtering out extraneous noise. In the first step, MP is carried out to represent the ecological environmental signal's inner structure. The second step consists of MFCCs feature extraction. Finally, two different classifiers, Support Vector Machine (SVM) and Gaussian mixture model (GMM) were performed and compared using the proposed features. Experimental results showed that the SVM-based classifier outperforms the GMM classifier and indicated that this method with sparse representation achieved improved performance in noisy environments.**

Keywords-ecological environmental sounds recognition; matching pursuit; genetic algorithm; sparse representation; MFCCs

I. INTRODUCTION

Ecological environmental audio signals, to some extent, can reflect the extent of changes in living environments, the impact of human activities on their living conditions, biological activities and their population changes, and that it can also be employed to provide data support for nature conservation. For example, in natural audio scenes, automatic bird song detection for the existence of the vocalizations of two endangered bird species was applied and used in automatic habitat mapping to reflect trends in bird population sizes[2]. It proposed a method for the development of a soundscape, and verified the potential of the proposed method for classification of environmental sounds within a soundscape development task [4]. The work showed that the acoustic environment was a rich source of context information and could be used as a good indicator of current activity [5]. Therefore, analyzing ecological environmental sound signals is of major importance.

In comparison to other fields in pattern recognition, little work had been carried out regarding environmental sounds recognition. Nevertheless, it has been gained attention in recent years. As compared to speech and music recognition, its recognition rate is lower. In this paper, recordings attained by a recording pen at mountain forest are employed in our experiments.

Audio signals have been traditionally characterized by MFCCs. MFCCs have shown to work well for audio signals, but their performance degrades in presence of noise. Namely, noise influences the quality of audio recordings and thus the reliability of recognition results. In this work, we propose to use the matching pursuit (MP) algorithm based on improved genetic algorithm to obtain sparse representation of ecological environmental sounds, accomplishing the task of filtering out extraneous noise. The searching method used is genetic algorithm improved with elitist strategy and evolution reversal. Noise is random, irrelevant and has no structural characteristics. So after sparse decomposition, the sample signal can be a good elimination of the effect of noise, that is, the reconstructed signal are consistent parts of atoms' structure properties. Because a few atoms can represent the signal, the cost of signal processing can be reduced in this way and the most important characteristics of the signal are maintained. There are no specific requirements for the atom and thus provides great flexibility for applications. It has been widely used in recognition of different types of environmental sounds [8], signal denoising[1], and pattern recognition[6] and so on. Our goal in this paper is to use MP as a way to denoising and to study different ecological environmental sounds in a more general sense.

The classifier SVM is widely used in speech recognition, image recognition, environmental noise classify, and biomedical pattern recognition and so on. Gerosa L. et al utilized two parallel GMM classifiers for discriminating screams from noise and gunshots from noise, and their system achieved precision of 90% [10]. Barkana B.D. and Uzkent B. compared the SVM and k-means clustering classifiers on three environmental noises, and the SVM-based classifier outperformed the k-means clustering classifier [11]. In this paper, two different classifiers SVM and GMM are used for classification.

The remainder of this paper is organized as follows. Decomposing method of matching pursuits is discussed in Section Ⅱ. Then MFCCs feature extraction is discussed in SectionⅢ. SectionⅣ contains simulation experiment and analysis of the results. Finally, concluding remarks and future research directions are given in Section Ⅴ.

Project founded by National Natural Science Foundation of China under Grant No. 61075022

978-1-4673-0964-6/12/$31.00 ©2012 IEEE

II. ANALYSIS OF MP SPARSE DECOMPOSITION METHOD

A. Signal sparse decomposition based on MP

The MP algorithm was originally introduced by Mallat and Zhang [1] for adaptively decomposing signals in an over-complete dictionary of functions, providing a sparse linear expansion of waveforms. As long as the dictionary is over-complete, the expansion is guaranteed to converge to a solution where the residual signal has zero energy. The following description of MP algorithm is based on the descriptions from [1].

Let H be a Hilbert space, the signal $f \in H$, dictionary D be a family $D=(g_{\gamma_i})_{\gamma_i \in \Gamma}$ of waveforms in H such that $\| g_{\gamma_i} \|=1$, where Γ is the parameter set and g_{γ_i} is called an atom. The number of atoms in dictionary is much larger than the signal's length.

- Let $g_{\gamma_0} \in D$, choose an atom g_{γ_0} from D that best matches the f, that is

$$|< f, g_{\gamma_0} >|= \sup_{\gamma \in \Gamma} |< f, g_{\gamma} >| , (1)$$

  Then the signal can be decomposed into covariates on the best atom g_{γ_0} and the residual, that is

$$f =< f, g_{\gamma_0} > g_{\gamma_0} + R^1 f , (2)$$

- Decompose the residual

$$R^k f =< R^k f, g_{\gamma_k} > g_{\gamma_k} + R^{k+1} f , (3)$$

  Where $R^k f$ is the residual vector after approximating f, g_{γ_k} satisfies

$$|< R^k f, g_{\gamma_k} >|= \sup_{\gamma \in \Gamma} |< R^k f, g_{\gamma} >| , (4)$$

- Judge if $\| R^k f \| < \varepsilon$ ($\varepsilon > 0$), ε is the threshold that has been set, if $\| R^k f \| < \varepsilon$, then come to the next step; If not, we come to the beginning.

$$f \approx \sum_{k=0}^{L} < R^k f, g_k > g_k , (5)$$

- The original vector f is decomposed into a sum of dictionary elements that are chosen to best match its residues.

Where L is the total number of iteration ($L \ll N$), in other words, the atoms of the reconstruct signal. Formula (5) and $L \ll N$ represent the idea of sparse representation. It should be noted that with the decomposition, namely, as k is increasing, the residual $R^k f$ is gradually decreasing, until disappear.

From formula (1), we can see that in signal decomposition based on MP, we have to get through with the projection calculation on each atom in the dictionary, whereas the number of atoms in an over-complete dictionary is very large. This is a key factor of the high computational complexity in signal sparse decomposition.

B. Structural characteristics of the over-complete dictionary

The following description of the formation of the dictionary is based on the descriptions from [7].

The dictionary $D=(g_\gamma)_{\gamma \in \Gamma}$ is formed by Gabor; an atom is made up of a modulated Gaussian window function:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) \cos(vt+w) , (6)$$

In which, $g(t)=e^{-\pi t^2}$ is Gauss function; $\gamma=(s,u,v,w)$ is time-frequency parameters. The time-frequency parameters are discretized as follows: $\gamma=(s,u,v,w)=(a^j, pa^j\Delta u, ka^{-j}\Delta v, i\Delta w)$, among which $0<j\le \log_2 N$, $0 \le p \le 2^{-j+1}$, $0 \le k < 2^{j+1}$, $0 \le i \le 12$, $a=2$, $\Delta u = 1/2$, $\Delta v = \pi$, $\Delta w = \pi/6$. In $D=(g_\gamma)_{\gamma \in \Gamma}$, an atom is determined by $\gamma=(s,u,v,w)$. For any scale s >0, translation u, frequency v and phase w, we denote $\gamma=(s,u,v,w)$. The parameters s, v and w define the waveform of an atom. The parameter u defines the center of an atom.

Theoretically, a dictionary with a good structure should contain a number of atoms and types as much as possible to get good effects of sparse decomposition. While for actual calculation, if possible, should not contain similar atoms, to meet the storage and computation requirements. Standing on this point, keep the parameters $u=N/2$ not changing. Thus, the size of the dictionary will be greatly reduced to fulfill requests of general computers' memory [7].

C. Find the best atom by genetic algorithm

In MP decomposition, most of the calculation is spent on looking for the best atom, whereas finding the optimal atom is an optimization problem, and thus the genetic algorithm can be used for optimization, which will greatly reduce the amount of computation. We proposed to apply the genetic algorithm using elitist strategy and evolution reverse in selecting the best atom in the over-complete dictionary. Fig. 1 shows a rough flow chart of the algorithm.

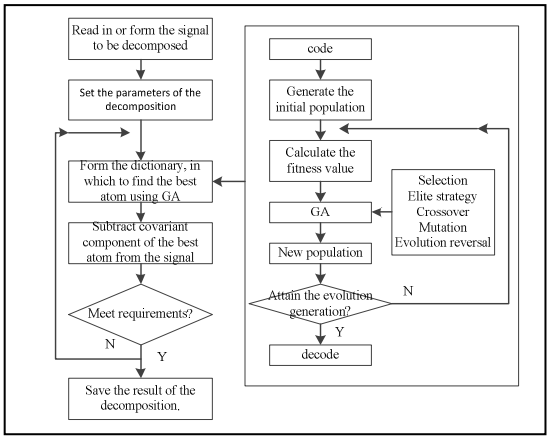

Figure 1. Frame of MP sparse decomposition algorithm based on GA.

If the predetermined evolution generation is attained, the program ends.

The best atoms are strongly expressive to the signal's inner characteristics. So, a few atoms would be able to reconstruct the major structure of the signal contaminated by extraneous noise. As shown in Fig. 2, only 25 atoms can reconstruct the signal's main structure.

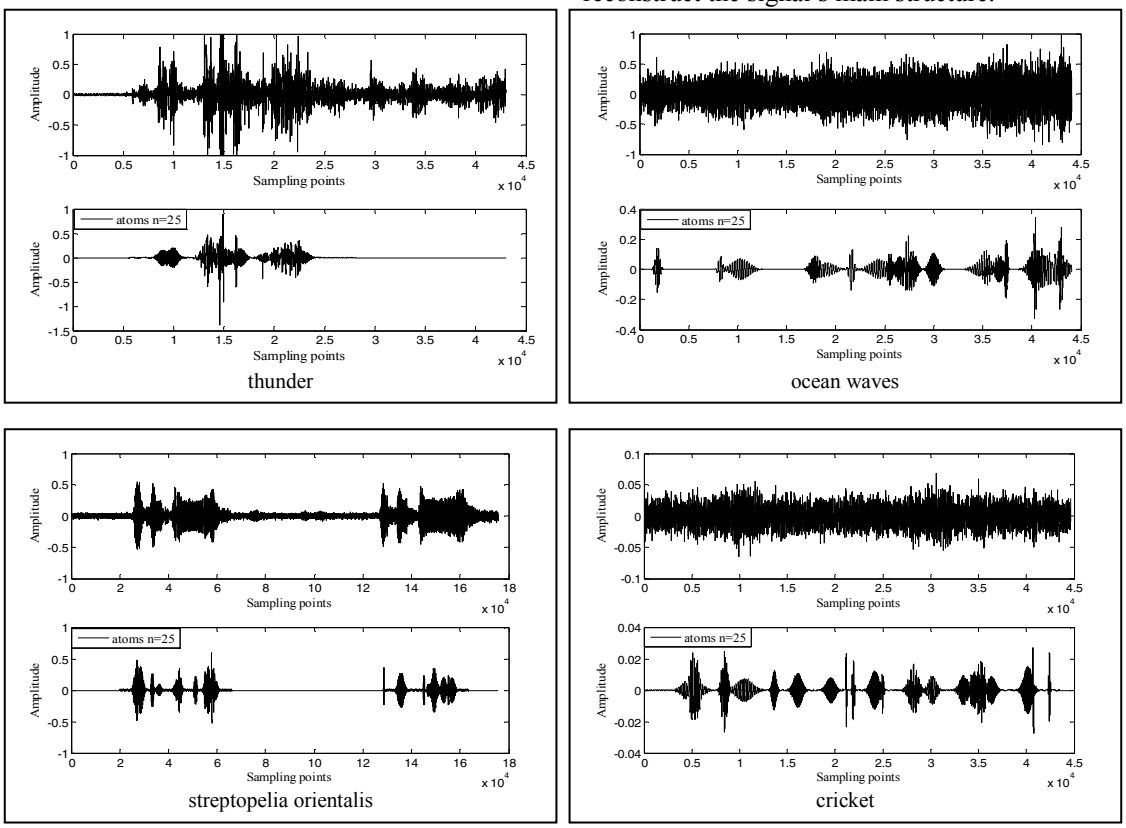

Figure 2. Comparison of the sample signal and the sparse representation signal of n=25

D. Limitations of the algorithm

When the problem size is smaller, we can generally get the optimal solution. While the problem size is larger, only approximate solutions may be obtained. Then, we can increase the population size and maximum genetic generations to make it closer to the optimal solution.

III. MEL-FREQUENCY CEPSTRUM COEFFICIENTS FEATURE EXTRACTION

Choose appropriate features are vital to a recognition system. MFCCs take human perception sensitivity with respect to frequencies into consideration. The frequency bands are equally spaced on the Mel-scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound.

The MFCCs feature extraction is easy and fast to implement. Studies show that MFCCs can be used as audio classification feature [8], and have higher classification accuracy. The computation procedure of MFCCs is as follows:

- The input signal is segmented into frames; calculate the log energy of each frame.
- Each frame is multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame.
- Perform FFT to obtain the magnitude frequency response of each frame.
- Multiple the magnitude frequency response by a set of 24 triangular band-pass filters to get the log energy of each triangular band-pass filter.
- Apply DCT on the 24 log energy obtained from the triangular band-pass filters to have 12 Mel-scale cepstrum coefficients.
- Add the log energy and pitch period as the 13rd, 14rd feature to MFCC respectively. The pitch tracking algorithm is based on sum of magnitude difference square function [12].

Here, the frame size N=128, that is about 11.6ms with overlap of 1/2 of the frame size.

IV. EXPERIMENTS AND RESULTS ANALYSIS

A. Dataset

Our database of ecological environmental sounds is not clean audio. In order to simulate real scenarios, the sounds used in our experiments are from the Freesound Project [4]

and our project team who use the voice recorders to record sounds on the spot. The recording backgrounds are near a mountain stream waterfall，in the street and the mountain path . Table Ⅰ shows the composition of the database.

All the selected sounds are converted to 11025Hz sampling rate, mono-channel and 16 bits. Considering the decomposition rate of MP, we cut the long-term audio sequence into 4s-6s segments. Segments are taken as the basic classification units and segments from the same audio are considered a set.

TABLE I. DATASET OF ECOLOGICAL ENVIRONMENTAL SOUNDS

Index	Type	Sound constitutes	Counts
1	Water	ocean waves, running water	80
2	Weather	raining, snowstorm, thunder and rain	140
3	Mammal	bark, meow, boar	140
4	Bird	coucal, gallinule, silver-eye, common pheasant, Chinese francolin, bamboo partridge, streptopelia orientalis, necklace dove, garrulax canorus, woodpecker call	424
5	Insect	cricket, fly, bee	80
6	Others	train passing, footsteps, frog	120

B. Experimental results and analysis of SVM and GMM

Based on the above MP sparse decomposition, we choose n=25 atoms in our experiments. After extracting MFCCs feature, the SVM and GMM classification model were utilized on the 24 types of sounds.

We choose the RBF kernel to try, and then the penalty parameter c and the kernel parameter variance g is chosen. We performed an eightfold cross validation for the best two parameters c and g. And then use the best parameters to train the model. The overall process is as follows.

- Perform sparse decomposition on sampled signals to obtain reconstructed signals with twenty-five atoms. Performance comparison of a sample reconstructed of different atoms is shown in figure 3.

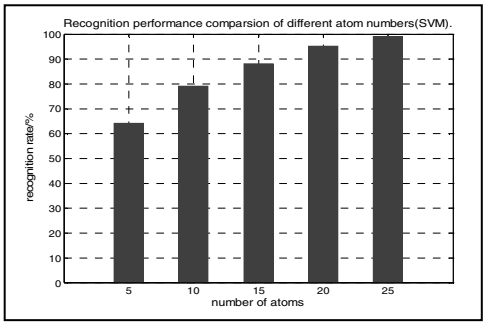

Figure 3. Performance comparison of different atom numbers.

- Extract 14--dimensional MFCCs features of each input sample.
- Conduct scaling on the data to the range of [0, 1]. The main advantage is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another is to avoid numerical difficulties during the calculation.

  Train the SVM classifier using the samples in training set (training set given instance-label pairs and features). And then carry out label prediction on testing dataset using the obtained model (testing set only given instances features). The classification accuracy is shown in Figure 6. Here, the Implementation of the classifier is adopted the LIBSVM toolbox. The process is presented as follows:

  Use eightfold cross validation to find the best parameter (c, g) so that the classifier can accurately predict unknown data. We use a course grid search first. After identifying a better region on the grid, a finer grid search on that region is conducted.

- After the best c&g is found, the whole training set is trained again to generate the final classifier.

Comparison of the classification accuracy on sounds of 24 before and after sparse representation is shown in Fig. 4. The sequencing of sounds is same with those in Table Ⅰ,that is, ocean waves, running water, raining, snowstorm, thunder and rain, bark, meow, boar, coucal, gallinule, silver-eye, common pheasant, Chinese francolin, bamboo partridge, streptopelia orientalis, necklace dove, garrulax canorus, woodpecker call, cricket, fly, bee, train passing, footsteps, frog.

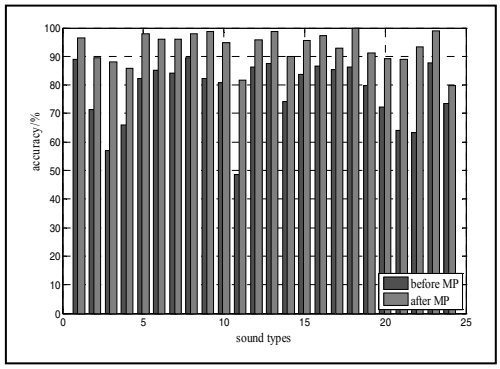

Figure 4. Classification results of SVM

On the whole, the recognition rate is better after sparse reconstruction. As to running water, many samples contain chirpings of insects of high and narrow frequency; as to raining that are typically noise-like with a broad flat spectrum, they may not be effectively modeled, and hence they have a lower recognition rate.

For completeness, we compared the results from the GMM. We used the same database in the SVM classifier to test the performance of the GMM classifier.

In the process of training GMM model, we used k-means algorithm to get the initial cluster center, and the expectation maximization algorithm to estimate parameters of the GMM [13]. With the cluster centers, the MFCCs got from the

previous could be converged into a stable 14-dimensional mean vector. The GMM establish a probability model for each possible class of signal. Tests were carried out on every frame to get the probability score, and the overall score summation for each audio segment decided the sound type. Fig. 5 shows the results of the classification accuracy using SVM and GMM.

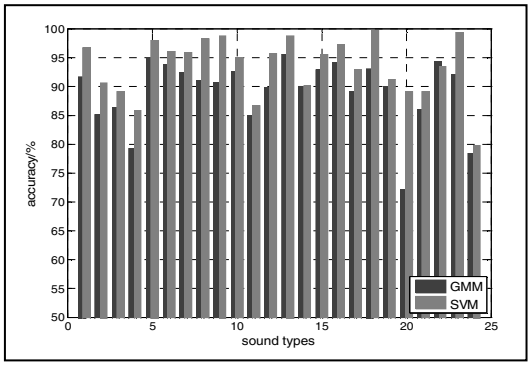

Figure 5. Classification results of SVM and GMM

Table Ⅱ shows recognition rate under the SVM-based classifier for the sound classes of water, weather, mammal, bird, insect and others.

TABLE II. RECOGNITION PERFORMANCE OF SVM

Type	MFCC+SVM	
	Before MP	*After MP*
water	82.41%	93.67%
weather	78.71%	89.59%
mammal	87.06%	97.49%
bird	85.93%	98.71%
insect	77.96%	90.29%
others	80.33%	96.33%

TableⅢ shows recognition rate under the GMM-based classifier.

TABLE III. RECOGNITION PERFORMANCE OF GMM

Type	MFCC+GMM	
	Before MP	*After MP*
Water	74.24%	89.86%
Weather	75.09%	85.38%
mammal	85.66%	92.06%
Bird	81.48%	94.37%
Insect	70.51%	86.96%
others	79.12%	92.02%

Clearly, the SVM-based classifier outperforms the GMM-based classifier. Even at locations of the insect sounds near the pool at night, the SVM with features after MP still yield a high correct rate and was robust against noise. This justifies SVM is suitable for this application.

V. SUMMARY AND PROSPECT

In this paper, we present an ecological environmental sounds classification system using the MP to compute adaptive signal representations and thus eliminating extraneous noise in the signal. The MP was improved by GA using elite strategy and evolution reversal. Then extract MFCC features and be classified by the SVM and GMM. The comparison experiments show that the MFCCs features set after MP sparse decomposition provide significant classification accuracy even under noisy environments, and the SVM classifier outperforms the GMM classifiers. In future, we will add more environmental sounds into our dataset and introduce some new audio features and effective classification algorithms to improve classification accuracy and reduce computation complexity.

ACKNOWLEDGMENT

The authors would like to thank the members in our team for their helpful comments and suggestions.

[1] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” IEEE Trans. Signal Processing, vol. 41, no. 12, pp. 3397 -3415, 1993.

[2] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, “Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring,” Pattern Recognition Letters, vol. 31, no. 12, pp. 1524-1534, 2010.

[3] YIN Zhong-Ke, WANG Jian-Ying, PierreVandergheynst . “Signal sparse decomposition based on GA and atom property,” Journal of the China Rail way Society, vol. 27, no. 3, pp. 58-61, 2005.

[4] I. Paraskevas, S. M. Potirakis and M. Rangoussi, “Natural soundscapes and identification of environmental sounds: a pattern recognition approach, ” Proceedings of the 16th Int. Conference on Digital Signal Processing, Santorini, Greece, pp. 5-7, July, 2009.

[5] Smith D, Ma L and Ryan N . “Acoustic environment as an indicator of social and physical context,” Personal and Ubiquitous Computing, vol. 10, no. 4, pp. 241-254, 2006.

[6] Arthur P. Lobo and Philipos C. Loizou, “Voiced/unvoiced speech discrimination in noise using gabor atomic decomposition,” Proceedings of 2003 IEEE. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 03), Hong Kong, China, Vol. 1, No. 4, pp. 820-823, Apr, 2003.

[7] YIN Zhongke, WANG Jianying, SHAO Jun, "Sparse decomposition based on structural properties of atom dictionary, "Journal of Southwest University, vol. 40, no. 2, pp. 173-178, April. 2005 .

[8] S. Chu, S. Narayanan, C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features,” Proc. Audio, Speech, & Language Processing, vol. 17, no. 6, pp. 1142-1158, Aug 2009.

[9] Ma L, Milner B, Smith D, "Acoustic environment classification, " ACM Trans. Speech Lang. Process, vol. 3, no. 2, pp. 1-22, 2006.

[10] L. L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi and A. Sarti, "Scream and gunshot detection in noisy environments," in 15th European Signal Processing Conference (EUSIPCO-07), Sep., Poznan, Poland, 2007.

[11] Buket D. Barkana, Burak Uzkent, “Environmental noise classifier using a new set of feature parameters based on pitch range,” Applied Acoustics, vol. 72, no. 11, pp. 841-848, 2011.

[12] LIU Jian, ZHENG Fang, WU Wen-Hu , “Real-time pitch tracking based on sum of magnitude difference square function, ”Journal of Tsinghua University(Science and Technology), vol. 46, no. 1, pp. 47-51, 2006.

[13] Jwu-Sheng Hu, Chieh-Cheng Cheng and Wei-Han Liu, “Robust speaker’s location detection in a vehicle environment using GMM models,” IEEE Transactions on Systems, Man and Cybernetics, Part B, vol. 36, no. 2, pp. 403–412, April 2006.