International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011
1793-8201

Automatic Recognition and Retrieval of Wild Animal Vocalizations

S.Gunasekaran and K.Revathy

Abstract—**Automatic animal sound classification and retrieval is very helpful for bioacoustic and audio retrieval applications. In this paper we propose a system to define and extract a set of acoustic features from all archived wild animal sound recordings that is used in subsequent feature selection, classification and retrieval tasks. The database consisted of sounds of six wild animals. The Fractal Dimension analysis based segmentation was selected due to its ability to select the right portion of signal for extracting the features. The feature vectors of the proposed algorithm consist of spectral, temporal and perceptual features of the animal vocalizations. The minimal Redundancy, Maximal Relevance (mRMR) feature selection analysis was exploited to increase the classification accuracy at a compact set of features. These features were used as the inputs of two neural networks, the k-Nearest Neighbor (kNN), the Multi-Layer Perceptron (MLP) and its fusion. The proposed system provides quite robust approach for classification and retrieval purposes, especially for the wild animal sounds.**

Index Terms— **FD - Fractal Dimension, KNN - k-Nearest Neighbor classifier, MLP - Multilayer Perceptron, mRMR - minimal Redundancy - Maximal Relevance.**

I. Introduction

Recognizing sources in the environment from the sounds they produce is one of the primary functions of the auditory system. Sound producing objects have acoustic properties, which are the result of the production process. These properties enable us to recognize sound sources by listening. The property includes the type of excitation, the physical construction, and the shape and size of the resonance structures. The type of excitation varies from one to another, and has significant influence on the sound.

The goal of automatic auditory scene analysis is to create computer systems that can learn to recognize the sound sources in a complex auditory environment. In this paper, the vocalization of six different wild animals will be considered. This includes sounds produced by bear, lion, gorilla, elephant, wolf and eagle. The data of the selected six wild animals is summarized in Table 1. The table contains English names and family, class, order and vocalizations (sound types).

The system described here is still in the early stages of implementation. One of the goals of this paper is to solicit feedback from members of the bioacoustic research community about the proposed approach. This paper is organized as four parts, first part explains about the sound segmentation procedure, second part briefs about the feature extraction, third part briefs about the feature selection and the fourth part describes the training and testing of the neural network classifiers and classifier fusions.

Manuscript received December 18, 2009.

S.Gunasekaran is doing research with Department of Computer Science, University of Kerela, Trivandrum, India. phone: +91-90361-96856; e-mail: yesgunaa@ gmail.com

Dr. K.Revathy is with Department of Computer Science, University of Kerela, Trivandrum, India as Professor (Retd). e-mail: revathy_srp@yahoo.com

II. Sound Segmentation

The different elements composing the system will be described in the following order: the sound segmentation, the feature extraction functions module, feature selection module, its link to Neural Networks block. A Sound Segmentation process first ensures the suppression of silent portions of the animal sound signals. When the signal drops under a certain level, the portion is discarded, because features extracted from such part present completely incorrect information to the classifier.

Fractal dimension is widely used for image segmentation. Also few attempts have been made for speech and music recognition using fractal based segmentation [6][8]. In this research we have used fractal dimension - D to segment the animal vocalizations. Fractal dimension of the wild animal sound signal has been computed using one of the most popular methods, 1-D Box-counting (Minkowski dimension) method.

A. Box-Counting method

Box counting in general involves covering a fractal with a grid of n-dimensional boxes with side length δ and counting the number of non-empty boxes N(δ). Boxes of recursively different sizes are used to cover the fractal. An input signal with N elements is used as input where N is power of 2. The slope β obtained in a bi-logarithmic plot of the number of boxes used against their size then gives the fractal dimension where $D_B = -\beta$. For a smooth 1-D curve

$$N(\delta) \approx \frac{L}{\delta} \tag{1}$$

Where L is the length of the curve. So box-counting could be generalized as

$$N(\delta) \propto \frac{1}{\delta^{D_B}} \tag{2}$$

$$D_B = \lim_{\delta \to 0} \frac{\ln N(\delta)}{\ln \delta} \tag{3}$$

TABLE I. DETAILS ABOUT THE SELECTED SIX WILD ANIMALS

English name	Family	Class	Order	Order
Bear	Ursidae	Mammalia	Carnivora	Moan, Bark, Huff, Growl, Roar
Eagle	Accipitridae	Aves	Falconiformes	Squeak, chirp
Asian Elephant	Elephantidae	Mammalia	Proboscidea	Rumbling, trumpet, snort, bark, roar, cry, chirp
Gorilla	Hominidae	Mammalia	Primates	Hoot, scream, rumble, pig grunt, chest beat, laugh
Lion	Felidae	Mammalia	Carnivora	snarl, purr, hiss, cough, meow, woof, roar
Wolf	Canidae	Mammalia	Carnivora	Howl, Growl, bark-howl, whuff, whimper

B. Estimation of Box-Counting Dimension

The sampling rate of the sound data, Fs, was 44.1 kHz and 16-bit accuracy was used. The system was developed in the Matlab environment, and the Signal Processing Toolbox was utilized. The input animal vocalization signal was divided into 1024 samples size frames and then normalized. The fractal dimension of the frame has been calculated using 1-D box counting algorithm. The mean of the results acquired with box sizes of 2 to 64 was used as a fractal dimension D for that block. Values computed with box sizes above 64 have no significant information about the signal. Portion of the signal with fractal dimension FD below 1.95 was chosen for feature extraction and the rest was discarded.

III. FEATURE EXTRACTION

Feature extraction is a process where a segment of an audio is characterized into a compact numerical representation. There are many features that can be used to characterize audio signals. Generally they can be grouped into five categories: temporal, spectral, perceptual, harmonic and statistical.

- **Temporal features** – Features that are calculated from the input waveform. Example: Auto-correlation coefficients, Zero Crossing Rate, Log-attack time, Temporal centriod, Signal power.
- **Spectral features** – Features that are computed from STFT of the input signal. Example: Spectral centroid, Spread, Skewness, Rolloff, Kurtosis, Variation, Flatness, Crest
- **Perceptual features** – Features that are computed from the human perceptual model. Example: Perceptual spectral centroid, Spread, Skewness, Rolloff, Kurtosis, Variation, Slope, Decrease, Tristimulus, Odd-to-Even ratio.
- **Harmonic features** – Features that are computed from the sinusoidal harmonic model of the signal. Example: Harmonic/noise ratio, Odd to even ratio, Tristimulus.

In this section we describe several audio features that are used in this system.

Fundamental frequency: For an harmonic signal, the fundamental frequency is the frequency so that its integer multiple best explain the content of the signal spectrum.

Spectral Flatness: The spectral flatness is a measure of the noisiness (flat, de-correlation)/ sinusoidality of a spectrum or a part of it. It is computed by the ration of the geometric mean to the arithmetic mean of the energy spectrum value. For a tonal signal SFM is close to 0, for noisy signal it is close to 1.

Log-Attack Time: The log-attack time is the algorithm (decimal base) of the time duration between the time the signal starts to the time it reaches its stable part. It has been proved to be one of the most perceptually important descriptors. It can be estimated taking the logarithm of the time from the start to the end of the attack.

$$LAT = \log_{10} (stop\_attack - start\_attack) \quad (4)$$

Temporal Centroid: The temporal centroid is the time averaged over the energy envelop. It allows distinguishing percussive from sustained sounds. It has been proved to be one of perceptually important descriptors.

Auto-correlation: The cross-correlation represents the signal spectral distribution but in the time domain (the cross-correlation of the signal is the inverse Fourier transform of the spectrum energy distribution of the signal). In order to obtain cross-correlation coefficients independent from the sampling rate of the signal, the signal is first down-sampled at 11025Hz. From the cross-correlation of the signal we keep the first 12 coefficients.

Zero-crossing Rate: The zero-crossing rate is a measure of the number of time the signal value cross the zero axe. Periodic sounds tend to have a small value of it, while noisy sounds tend to have a high value of it. It is computed at each time frame on the signal.

Mel Frequency Cepstral Coefficients (MFCC) : The MFCC represents the shape of the spectrum with very few coefficients. The cepstrum is the Fourier transform of the logarithm of the spectrum. The Mel-cepstrum is the cepstrum computed on the Mel-bands instead of the Fourier spectrum. The use of Mel spectrum allows better to take better into account the mid-frequencies part of the signal. The MFCC are the coefficients of the Mel cepstrum. Total 13 coefficients are stored for each frame.

Short-Time Energy: The short-time energy of an audio signal is defined as

$$E_n = 1/N \sum_m | x(m)\, w(n-m)|^2 \quad (5)$$

Where x(m) is the discrete time audio signal, n is the time index of the short time energy and w(m) is a rectangle window. The short time energy function shows the amplitude variation over time.

Spectral Centroid : The spectral centroid is the balancing point of the spectrum. It is a measure of spectral shape and is often associated with the notion of spectral brightness. The spectral centroid can be calculated as

N N

$$C = \sum_{n=1} M_t[n].n \Big/ \sum_{n=1} M_t[n], \quad (6)$$

Where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n.

Spectral Spread: Following the previous definition, we define the spectral spread as the spread of the spectrum around its mean value i.e. the variance of the above defined distribution

$$\sigma^2 = \int (x-\mu)^2 . p(x)\, \delta x \quad (7)$$

Spectral skewness: The skewness gives a measure of the asymmetry of a distribution around its mean value. It is computed from the 3rd order movement:

$$m_3 = \int (x-\mu)^3 . p(x)\, \delta x \quad (8)$$

The skewness is then: $\gamma_1 = m_3 / \sigma^3$

The skewness SK describes the degree of asymmetry of the distribution.

SK = 0 indicates a symmetric distribution

SK < 0 indicates more energy on the right,

SK > 0 indicates the more energy on the left.

Spectral kurtosis: The kurtosis gives a measure of the distribution around its mean value. It is computed from the 4th order movement:

$$m_4 = \int (x-\mu)^4 . p(x)\, \delta x \quad (9)$$

The kurtosis is then: $\gamma_2 = m_4 / \sigma^4$

The kurtosis K indicates the peakedness/flatness of the distribution.

K = 3 indicates a normal distribution

K < 3 indicates a flatter distribution,

K > 3 indicates a peaker distribution.

Spectral slope: The spectral slope represents the amount of the spectral amplitude. It is computed by linear regression of the spectral amplitude.

Spectral decrease: The spectral decrease also represents the amount of decreasing of the spectral amplitude. This formulation comes from perceptual studies, it is supposed to be more correlated to human perception.

Spectral-rolloff: The spectral roll-off point is the frequency so that 95% of the signal energy is contained below this frequency. It is correlated somehow to the harmonic/noise cutting frequency.

TABLE II. FEATURE DESCRIPTION

S.No.	Feature
1	Fundamental frequency – f0
2	Spectral centroid
3	Spectral skewness
4	Spectral rolloff
5	Spectral spread
6	Spectral kurtosis
7	Spectral slope
8	Spectral decrease
9	Spectral variation
10 - 13	Spectral flatness (4 values)
14 - 17	Spectral crest (4 values)
18 - 29	Auto-correlation coefficients (12 values)
30 - 42	MFCC (13 values)
43 - 55	Delta-MFCC (13 values)
56 - 68	Delta-Delta-MFCC (13 values)
69	Perceptual spectral centroid
70	Perceptual spectral spread
71	Perceptual spectral skewness
72	Perceptual spectral rolloff
73	Perceptual spectral kurtosis
74	Perceptual spectral slope
75	Perceptual spectral decrease
76	Perceptual spectral variation
77 – 79	Perceptual spectral Tristimulus (3 values)
80	Perceptual spectral Odd-to-Even Ratio
81	Zero Crossing Rate
82	Max-Energy FFT Bin
83	Log-Attack Time
84	Temporal centroid
85	Energy of the signal

The feature set of the proposed system contains 30 distinct features (85 features values). The complete list of features is enumerated in Table 2. Most of the features for the proposed system have been chosen from MPEG-7 Low-level audio descriptors [3]. Rests of the features are conventional and experimented by researchers proven with good results. There is no publicly available reference database of animal sounds. Animal vocalization sound files for this experiment were obtained from an internet search. All of the digital sound files were then converted into wav files and used for the experiment. The system has been trained with 300 test sets and has been tested with a different set of 300 test signals.

All the features except Temporal centroid and Log-attack time are frame based and are extracted from the segmented frame of 1024 samples. Temporal features are extracted from the time samples directly. Spectral features are computed from Short-Time Fourier Transform (STFT) of the signal. Classification methods are sensitive to the scale of the features, especially in relation to each other. Normalization of the feature set takes care of this. The max of training set was used to select normalization parameters, and these adjustments were applied to all feature sets uniformly.

IV. FEATURE SELECTION

Feature selection is the process of removing features from the feature set which are less important with respect to the classification task to be performed. Feature selection will also be useful to reduce the processing power required for the classifier and to improve the classification accuracy as well.

Feature selection algorithms can be categorized into two types - filter methods and wrapper methods. Filter methods are much faster than wrapper methods and therefore are better suited to high dimensional data sets. The minimal Redundancy-Maximal Relevance (mRMR) is one of the fastest feature selection algorithms belonging to the category of filter method. Peng et al [2] investigated the significance of the 'minimal Redundancy- Maximal Relevance' (mRMR) feature selection algorithm and compared with many practical feature selection algorithms. The results showed better classification accuracy using mRMR feature selection. For a detailed comparison of mRMR with other feature selection algorithms see [2]. In this paper we have investigated the impact of the addition of mRMR feature selection algorithm towards animal vocalization classification. For this purpose we evaluated several cases for animal sound classification to select a compact set of features

with better classification accuracy. We used an extensive collection of 85 features in the experiments of study. The feature set was later trimmed to 60 features by applying the mRMR feature selection algorithm. Results of both the experiments are presented and analyzed.

TABLE III. TEST RESULTS OF INDIVIDUAL CLASSIFIERS

Animal	KNN		MLP	
	No FS	*FS*	*No FS*	*FS*
Bear	65.7%	68.7%	80.7%	60.0%
Eagle	76.7%	81.0%	98.3%	98.3%
Elephant	45.0%	48.3%	55.0%	60.0%
Gorilla	54.3%	64.3%	30.3%	62.3%
Lion	35.7%	41.7%	45.3%	44.7%
Wolf	82.7%	80.7%	70.0%	88.0%
Overall	**60.0%**	**64.1%**	**63.2%**	**68.9%**

TABLE IV. TEST RESULTS OF CLASSIFIER FUSION (KNN + MLP)

Animal	Sum-Based		Confidence-based	
	No FS	*FS*	*No FS*	*FS*
Bear	62.0%	66.3%	63.3%	66.7%
Eagle	96.0%	97.7%	96.3%	98.7%
Elephant	62.7%	56.3%	65.0%	56.7%
Gorilla	53.7%	66.3%	56.3%	64.3%
Lion	43.7%	47.0%	42.7%	45.3%
Wolf	80.0%	88.0%	82.7%	87.0%
Overall	**66.3%**	**70.3%**	**67.7%**	**69.8%**

V. CLASSIFIERS AND CLASSIFIER FUSION

A. Classifiers

This section gives a brief overview of the classification methods and the parameters used. Two different classifiers were used: K-Nearest Neighbor and Multi-Layer Perceptron classifiers.

1) K-Nearest Neighbor

In k-Nearest Neighbor classification, the training dataset is used to classify each member of a testing dataset. The algorithm could be illustrated as follows:

- For each case in the set to be classified, locate the k nearest neighbors of the training dataset. A Euclidean Distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
- Examine the k nearest neighbors. Assign the category to which most of these neighbors belong to, to the case being examined.
- Repeat this procedure for the remaining cases in the target set.

The k-Nearest Neighbor method is intuitively a very attractive method. A disadvantage of this method is its large computing power requirement, since for classifying an object, its distance to all the objects in the learning set has to be calculated.

There are three significant ways to fine-tune the performance of the k-NN classifier:

- Modify the distance function.
- Rescale the features.
- Change k, the number of neighbors conferred in each classification.

In our proposed system, the kNN classifier with k=11 has been chosen experimentally.

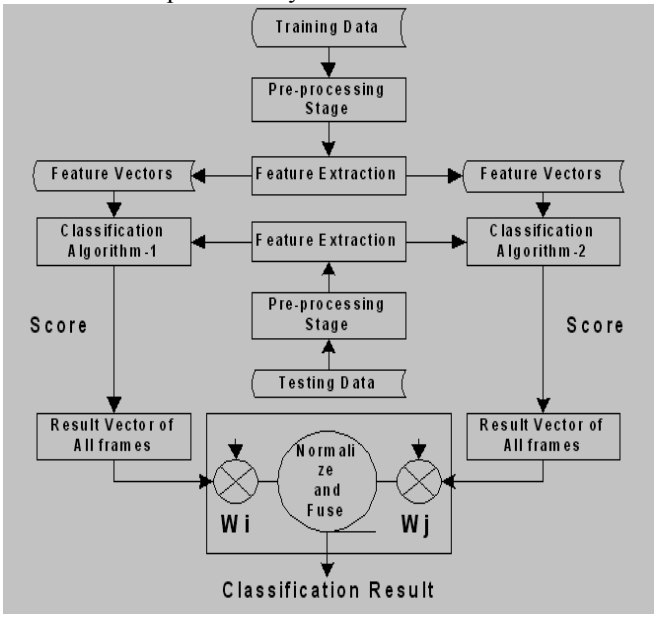

Figure 1. Classifier Fusion

2) Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) is a feed forward neural network. Multilayer perceptrons are networks with several layers of interconnected neurons, consisting of input neurons, output neurons and hidden neurons. A feed-forward neural network is one in which the neurons do not form a directed cycle. That is, a neuron in layer i - 1 is connected to every neuron in layer i, but to no other neurons in layer i - 1. Learning in multilayer neural networks is done using a technique called back propagation. Back propagation is a gradient descent search algorithm and can suffer from both slow convergence time and getting trapped within local minima. Extensive research has been conducted into improving back propagation algorithms. One of such optimization algorithm is scaled conjugate gradient descent. In our proposed system the MLP network was trained with back propagation using scaled conjugate gradient optimization.

B. Classifier Fusion

Classifier fusion has been shown to be beneficial theoretically and practically in terms of improving system accuracy. A number of methods have been developed for classifier fusion. There are two different types of classifier fusion techniques used widely. The first method operates on classifiers and put an emphasis on a development of the classifier structure. It does not do anything with classifiers outputs until combination process finds single best classifier or a selected group of classifiers and only then its outputs are taken as a final decision or for further processing. Another method operates mainly on classifiers outputs, and effectively the combination of classifiers outputs is

calculated.

In this paper we have chosen the method operating on classifiers outputs produced by individual classifiers. A diagrammatic representation of the classifier fusion method used is shown in Figure 1. The details of the experimental protocol are explained in the following.

The classifier fusion has been done at decision level. We employ the Sum-based and Confidence–based integration strategies to combine two classifiers k-NN and MLP.

The sum-Base fusion can be represented as

$$Q_j(x) = \sum_{i=1}^{M} C_{ij}(x) \qquad (10)$$

Where i is the index of the i^{th} classifier, $C_{ij}(x)$ is a measure of how probable it is that the classifier C_i will have a correct output given instrument x. In this case the errors in the confidences are averaged out by the summation.

Another probability-based combination method is the Confidence-based fusion. It is selecting the classifier that is most confident of itself. It can be described as follows

$$Q_j(x) = \max_j \{ C_{ij}(x) \} \qquad (11)$$

VI. Classification and Retrieval: Results

Our classification results are displayed in Table3 and Table4. In each column, the classification correctness of the wild animal vocalizations over different classification algorithm is presented. The rows of the table show how each animal sound is recognized. Column "No FS" is the results for the original feature vectors with no feature selection algorithm employed. Column "FS" is the results for compact set of features, minimized using mRMR feature selection algorithm. A significant observation that can be made is that the addition of the feature selection step has significantly improved the performance accuracy of the system. An average improvement of almost 5% is obtained.

Table 4 contains the recognition result of the classifier fusion experiment. The highest accuracy was obtained with KNN+MLP classifier (Sum-based fusion) using the Minimal Redundancy-Maximum Relevance-based Features Selection. Altogether, 70.3% of the test sounds of the six wild animals were recognized correctly while using both mRMR feature selection and Classifier fusion techniques.

VII. Conclusions and Future Work

Our aim was to study how wild animal vocalizations can be recognized and classified efficiently. Actually, the identification and extraction of features being representative for a distinct animal vocalization generally is a great challenge. The mRMR feature selection was selected for its ability to analyze and decorrelate the feature set. The presented results encourage further research on the underlying concepts. Results from experiments conducted also show that the two factors, feature selection and classifier fusion affects classification accuracy. The proposed research is still in its early stages. We shall continue developing the system for recognizing more wild animal vocalizations with extensive sound collection. We will study the reasons why specific animal sound produce high recognition errors and how to better differentiate these animal sounds.

References

[1] V. B. Deecke and V. M. Janik, "Automated categorization of bioacoustic signals: avoiding perceptual pitfalls," Journal of the Acoustical Society of America, vol. 119, no. 1, pp. 645-653, 2006.

[2] Peng, H.C., Long, F., and Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226-1238, 2005.

[3] MPEG-7 Overview (version 10) - ISO/IEC JTC1/SC29/WG11, October 2004.

[4] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," IEEE Transactions on Circuits and Systems, 2000.

[5] James Theiler, "Estimating fractal dimension", Journal of Optical Society of America, Vol. 7, No. 6, (page: 1055 - 1073) June,1990.

[6] Petros Maragos and Alexandros Potamianos, " Fractal dimensions of speech sounds: Computation and application to automatic speech recognition", Journal of Acoustical Society of America, March 1999.

[7] Shiu Yin Yuen, Chun Ki Fong, Kwok Leung Chan, Yiu Wah Leung, " Fractal dimension estimation and noise filtering using Hough transform", Journal of signal processing, Elsevier, Feb'2004.

[8] Gunasekaran, S. Revathy, K. "Fractal dimension analysis of audio signals for Indian musical instrument recognition". ICALIP.2008, 7-9 July 2008, ISBN: 978-1-4244-1723-0 (page: 257-261).

[9] Gunasekaran, S. Revathy, K, "Recognition of Indian Musical Instruments with Multi-Classifier Fusion," ICCEE, pp.847-851, 2008 International Conference on Computer and Electrical Engineering, 2008.

[10] Lie Lu, Hong-Jiang Zhang and Hao Jiang, "Content Analysis for Audio Classification and Segmentation" - IEEE Transactions on Speech and Audio Processing, VOL. 10, NO. 7, OCTOBER 2002.

[11] Mitrovic, D. Zeppelzauer, M. Breiteneder, C, "Discrimination and retrieval of animal sounds" , Multi-Media Modelling Conference Proceedings, 2006 ISBN: 1-4244-0028-7, 2006.

[12] Benzeghiba, M.F., Bourlard, H., 2003. "Hybrid HMM/ANN and GMM combination for user-customized password speaker verification". In: Proc. ICASSP, vol. 2. pp. 225-228.

[13] Xu, L.; Krzyzak, A.; Suen, C.Y . "Methods of combining multiple classifiers and- their applications to handwriting recognition". IEEE Trans. Systems Man, Cyb., SMC. v22 i3. 418-435.