PATTERN RECOGNITION OF ENVIRONMENTAL SOUNDS USING TIME-FREQUENCY DISTRIBUTIONS

Ioannis Paraskevas[1] and Maria Rangoussi[2]

[1] Department of Technology of Informatics and Telecommunications,
Technological Education Institute of Kalamata / Branch of Sparta,
7, Kilkis str., Sparta, GR-23100, GREECE
paraskevas@env.aegean.gr

[2]Department of Electronics,
Technological Education Institute of Piraeus,
250, Thivon str., Aigaleo-Athens, GR-12244, GREECE
mariar@teipir.gr

Abstract

Environmental sounds are signals recorded in areas of environmental or ecological interest, that convey information as to the status, the inhabitation and the use / human activities of the area. An increasing research interest in this field has recently produced an increasing number of databases that contain environmental sounds. The need for automatic event classification within the recordings of such databases has accordingly grown in importance. In this paper, a novel method

for the automatic recognition of environmental sounds is presented. The signals tested are echolocation calls produced by different species of bats. In the proposed method, each signal is processed to yield a time-frequency distribution, as the basis for the feature extraction. Time-frequency distributions are then compressed by extracting appropriate features. The feature vectors formed are introduced to an Artificial Neural Network classifier, in order to obtain classification decisions for each sound / event. Experimental results obtained from the classification of the bats' echolocation calls verify that the proposed method is capable to discriminate the aforementioned family of environmental sounds. The potential of the proposed method to perform well for other classes of environmental sounds is based on its generic, signal independent nature.

Keywords

Signal processing, feature extraction, pattern recognition of environmental sounds, time-frequency distributions.

1. Introduction

The role of content-based classification has become increasingly important due to the increasing number of audio-visual databases (Wold et al., 1996). In most cases classification is based on features derived from the visual content of the database. However, although this approach seems successful, the rate of correct classification is increased when audio cues are also employed (Paraskevas and Chilton, 2004; Paraskevas et al., 2006). Moreover, there exist problem cases where visual information is not available and hence, sound is the only information source for event classification. In this paper, a new method is presented for the classification of

environmental sounds. The proposed method is based on features extracted from the time-frequency distributions of the signals (Paraskevas and Chilton, 2003).

In generally, the application of pattern recognition is divided into two stages, namely: i) feature extraction and ii) classification. The (correct) classification rate depends on how efficiently the feature vectors introduced to the classifier encapsulate the information content of the signals (Webb, 2002).

Features commonly used for the application of audio pattern recognition are:

- the audio signal energy function,
- the average zero-crossing rate,
- the fundamental frequency,
- the spectral peak track,
- the brightness,
- the bandwidth pitch frequency and
- the cepstral / Mel-cepstral coefficients

(Zhang and Kuo, 2001; Wold et al., 1996).

Subsets of the aforementioned features, typically used for speech / speaker recognition, have been applied ad hoc to sound classification problems / applications. The features, though, that are employed for speech / speaker recognition use apriori information, related to the human speech production model and hence, they are not appropriate or suitable for classes of sounds other than speech. Moreover, in existing research (Zhang and Kuo, 2001; Wold et al., 1996), the features employed are either temporal or frequency related; features extracted from time-frequency distributions of the signal are rarely utilized. Time-frequency distributions present the signal's frequency evolution in time. In the proposed method, the feature vectors are formed from

statistical features extracted from time-frequency distributions of the signal.

There exist two types of sound pattern recognition: the ‘coarse’ and the ‘fine’ one. The first aims to classify sounds that do not belong to the same family, so as to identify different sports, environmental events etc. based on their sound content, whereas the second aims to classify sounds that belong to the same family, so as to identify different species of bats, birds or different kinds of string instruments etc. based on the sounds they produce. Consequently, ‘fine’ classification is more demanding compared to ‘coarse’ classification as the features that have to be extracted need to accurately represent the characteristics of each class to the classifier.

In this work, experiments are carried out using a database of echolocation calls recorded from fourteen (14) species of bats that exist in the United Kingdom (UK). In the feature extraction part of the pattern recognition process, the statistical measurements extracted from the Fourier Magnitude Spectrogram and the Choi-Williams Distribution form the feature vectors. In the classification part these feature vectors are introduced to an Artificial Neural Network classifier. The novelty of the proposed method is the use of statistical features that encapsulate the evolution of the spectral content of the signal with time, for the application of pattern recognition of environmental sounds.

The remaining part of this paper is structured as follows: In Sections 2 and 3 the feature extraction and the classification parts of the pattern recognition process for this application are described and in Sections 4 and 5 are presented the results of the proposed method and the conclusions, respectively.

2. Time-Frequency Distributions and Statistical Feature Vectors

2.1. Time-Frequency Distributions

Probably the most popular time-frequency distribution is the Fourier Magnitude Spectrogram (FMS). The FMS evaluates the Fourier Transform of the signal using a sliding window and describes the evolution of the signal magnitude content in time (Proakis and Manolakis, 1992; Rabiner and Schafer, 1978). Figures 1a, 2a and 3a present the Time Domain signals and figures 1b, 2b and 3b the corresponding FMS of the echolocation call signals produced by three of the bat species examined in this work (see Section 4 for details).

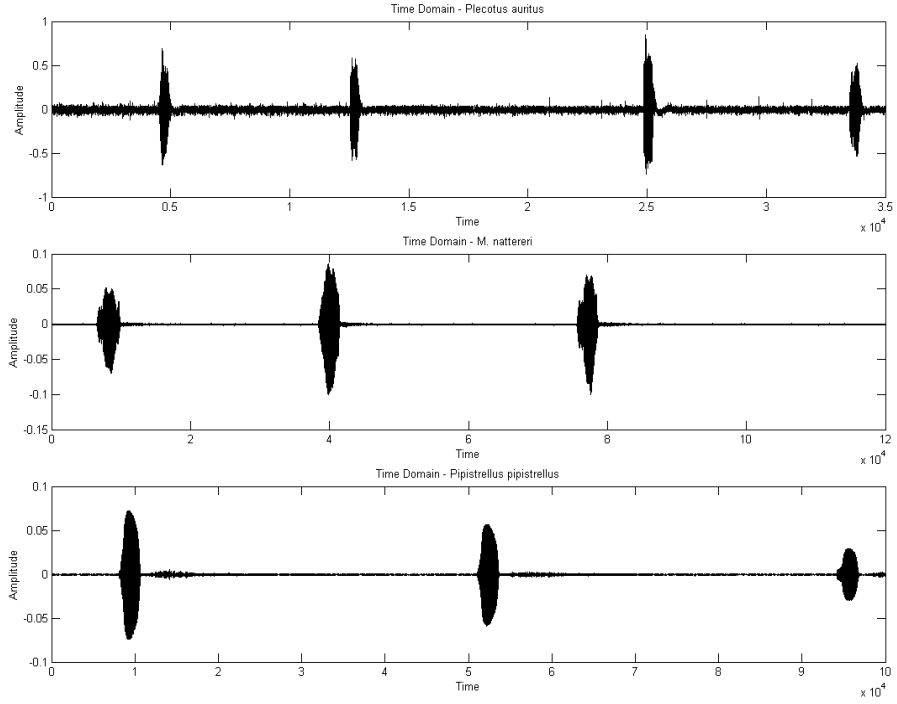

Figure 1a: Time Domain – Echolocation call of Plecotus auritus bat

Figure 2a: Time Domain – Echolocation call of M. nattereri bat

Figure 3a: Time Domain – Echolocation call of Pipistrellus pipistrellus bat

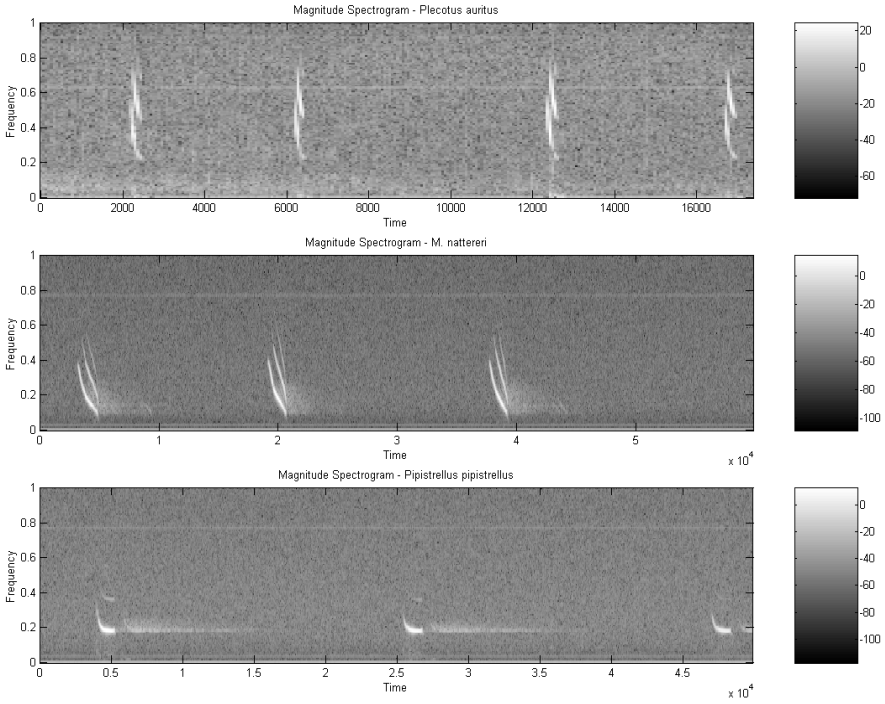

Figure 1b: Fourier Magnitude Spectrogram – Echolocation call of Plecotus auritus bat

Figure 2b: Fourier Magnitude Spectrogram – Echolocation call of M. nattereri bat

Figure 3b: Fourier Magnitude Spectrogram – Echolocation call of Pipistrellus pipistrellus bat

The other time-frequency signal representation employed in the proposed method is the Choi-Williams Distribution (CWD). The CWD is a relative of the Wigner-Ville Distribution in that they both belong to the Cohen's general class of time-frequency distributions (Cohen, 1989). Specifically, the Cohen's general class of time-frequency distributions is defined as:

$$P_{GEN}(t,f) = \frac{1}{4\pi^2}\iiint e^{-j\vartheta t - j2\pi f\tau + j\vartheta u}\phi(\vartheta,\tau)x^*(u-\frac{\tau}{2})x(u+\frac{\tau}{2})dud\tau d\vartheta \quad (1)$$

where: $\phi(\vartheta,\tau) = 1$ for the Wigner-Ville Distribution

or $\phi(\vartheta,\tau) = e^{-\frac{\vartheta^2\tau^2}{\sigma^2}}$ for the CWD.

Note that the complex-conjugate operator (*) is used, as the signals may be in their analytic

form, i.e. complex-valued.

The CWD is preferred, for our application, to the Wigner-Ville Distribution due to its cross-term reduction property (Nikias and Petropulu, 1993).

2.2. Statistical Feature Vectors

The information content of each time-frequency distribution has to be expressed in a more compact form - via feature vectors – in order to be introduced to the ANN in an efficient manner. Hence, after the time-frequency distributions are computed for each recording, a group of statistical features (Mood et al., 1974) is calculated from each distribution. Namely, the eight statistical measurements which form the feature vector for this application are:

i) Variance:

$$\text{Variance} = \varepsilon\left|x_i - \bar{x}\right|^2 \qquad (2)$$

where, $\bar{x}$ represents the sample mean.

ii) Skewness:

$$\text{Skewness} = \frac{\varepsilon(x_i - \bar{x})^3}{\sigma^3} \qquad (3)$$

where, σ represents the sample standard deviation.

iii) Kurtosis:

$$\text{Kurtosis} = \frac{\varepsilon(x_i - \bar{x})^4}{\sigma^4} \qquad (4)$$

where, σ again, represents the sample standard deviation.

iv) Inter-Quartile Range (I.Q.R.):

$$\text{I.Q.R.} = Q_{75} - Q_{25} \qquad (5)$$

i.e. the I.Q.R. is the difference between the 75^{th} and 25^{th} data percentile.

v) Median is defined as the central value of the ordered data set.

vi) Mean Absolute Deviation (M.A.D.):

$$M.A.D. = \varepsilon | x_i - \bar{x} | \qquad (6)$$

where $\bar{x}$ represents the sample mean.

vii) Range is the difference between the maximum and minimum values in the data set:

$$P = \max_i (x_i) - \min_i (x_i) \qquad (7)$$

viii) Log-Entropy:

$$E(x) = \sum_{i=1}^{N} \log(x_i^2) \qquad (8)$$

where log represents the logarithm and x_i represents the signal samples.

In all the statistical features defined above, the expected value operator $\varepsilon(.)$ is in practice estimated as time average along the N signal samples of a single recording, i.e., $\frac{1}{N}\sum_{i=1}^{N}(.)$ for a given signal recording producing a discrete-time signal x_i of length N samples.

These eight statistical measurements form the feature vector of each time-frequency distribution, calculated for each signal recording.

3. Artificial Neural Network Classifier

In this application, for the classification part of the pattern recognition process an Artificial Neural Network (ANN) is employed. The type of ANN selected is a Self-Organizing Map (SOM). Self-organizing networks detect regularities and correlations and adjust their future responses according to previous inputs (Kohonen, 2001). The neurons of the self-organizing networks recognize groups of similar vectors. Neurons that are physically close in the neuron layer tend to respond to input vectors that are similar to each other. These networks classify the input vectors into classes based on a competitive layer so as (a) to form subclasses of input vectors and then (b) to combine subclasses into final output classes (Kohonen, 1988).

4. Experiments and Results

In this Section, the experimental procedure and the results are presented for the pattern recognition of fourteen (14) species of bats, which exist in the UK, based on the echolocation calls they produce. These fourteen bat species namely are:

- Barbastella barbastellus,
- Eptesicus serotinus,
- Myotis bechsteinii,
- M. brandtii,
- M. daubentonii,
- M. mystacinus,
- M. nattereri,
- Nyctalus leisleri,
- N. noctula,
- Pipistrellus pipistrellus,
- P. pygmaeus,
- Plecotus auritus,
- Rhinolophus ferrumequinum and
- R. hipposideros.

The recordings of the echolocation calls employed for the experiments, are provided by (see Acknowledgement).

The proposed method is divided in two stages. The aim of the first stage is to classify the

aforementioned fourteen species in four groups and the aim of the second stage is to classify each one of the species within its group. Analytically, for the first stage, the feature vectors introduced to the ANN are formed from the statistical features (Subsection 2.2) extracted from the FMS of the signals (Parsons and Jones, 2000). Thus, the feature vectors introduced to the ANN classify the fourteen species in four groups:

1. Group A consists of the following five bat species:

   Barbastella barbastellus, Plecotus auritus, Eptesicus serotinus, Nyctalus leisleri and N. noctula,

2. Group B consists of the following five bat species:

   Myotis bechsteinii, M. brandtii, M. daubentonii, M. mystacinus and M. nattereri,

3. Group C consists of the following two bat species:

   P. pygmaeus and Pipistrellus pipistrellus,

4. Group D consists of the following two bat species:

   Rhinolophus ferrumequinum and R. hipposideros.

After the fourteen echolocation calls have been classified in these four groups based on their magnitude spectral characteristics, the same process is repeated for the second stage, using the CWD in order to classify each bat's echolocation call within its group. Consequently, for the second stage, the feature vectors introduced to the ANN, formed from the statistical features (Subsection 2.2) extracted from the CWD rather than the FMS, classify each bat's echolocation call within its group (figures 4a, 4b, 4c and 4d).

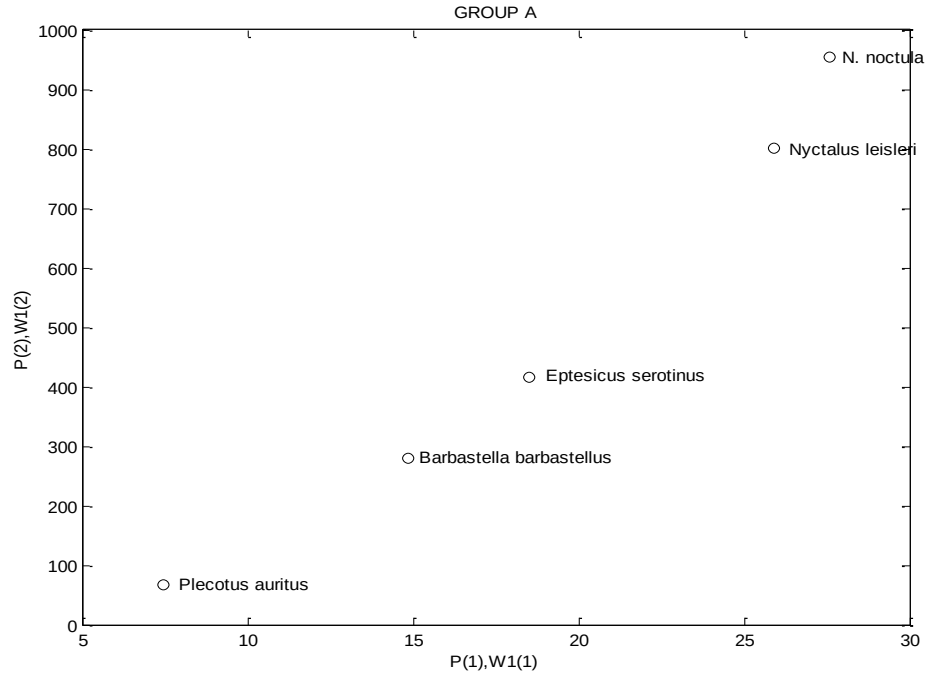

Figure 4a Classification of Group A of bats

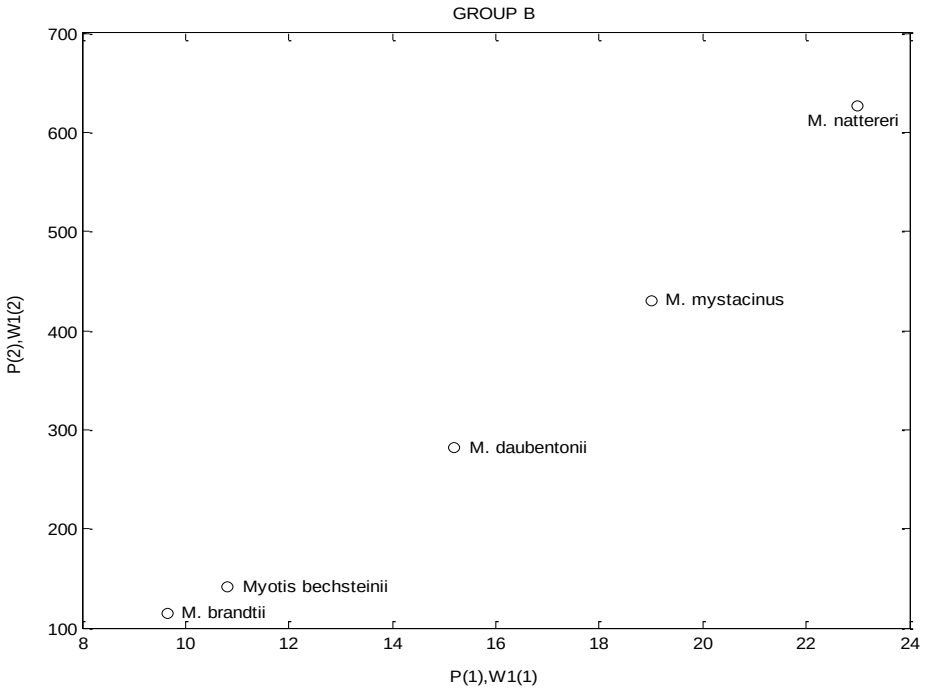

Figure 4b Classification of Group B of bats

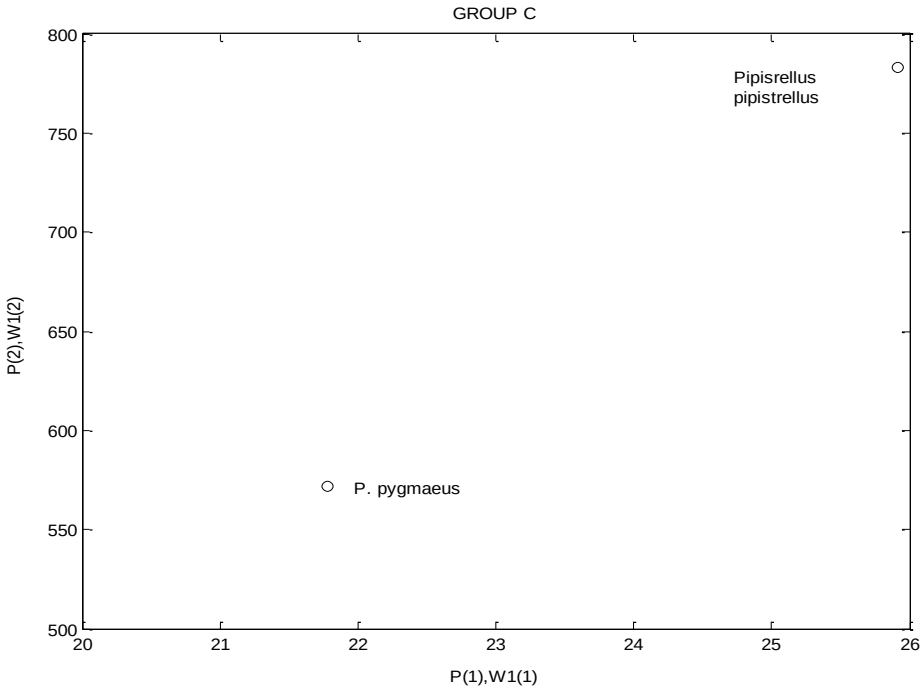

Figure 4c Classification of Group C of bats

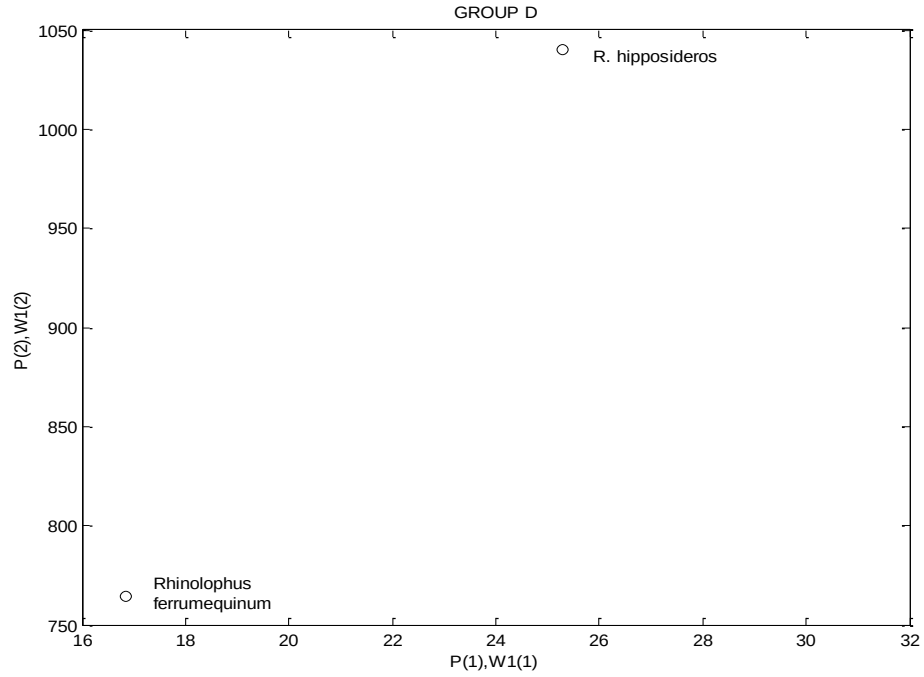

Figure 4d Classification of Group D of bats

Summarizing, the feature vectors formed by the feature extraction and classification method previously described, are distributed as shown in figures 4a, 4b, 4c and 4d. For the experiments, ten samples are employed of each species. Of each species within each group, only the central 'point' is shown in figures 4a, 4b, 4c and 4d, for simplicity. These central points are projected to a 1×2 – dimensional space for presentation purposes, although they are 1×8 – dimensional feature vectors, originally (Subsection 2.2). Hence, one can observe that the feature vectors employed are appropriate to discriminate the aforementioned classes as none of the central points coincide or appear to be close to each other.

5. Conclusions and Future Research

The aim of this research is to apply pattern recognition methods for the classification of environmental events based on their sound content. The database tested for this research consisted of the echolocation calls that produce fourteen species of bats that exist in the UK. These fourteen classes of sounds belong to the same family and consequently, the pattern recognition task is demanding due to the similarity among them. The experimental results showed that the novel method proposed, which employs statistical features extracted from time-frequency distributions, is appropriate for the classification of environmental events.

A relatively new research area where the classification of environmental sounds is applied, is the development of soundscapes (Krause, 2002). Soundscapes are maps which contain sound rather than morphological information. They are an important tool for the monitoring of certain areas of interest, such as NATURA 2000 areas (European Union's network of nature protection areas). The periodic comparison of the soundscapes of a certain area can provide useful ecological related observations.

Acknowledgement: We gratefully acknowledge the assistance of S. Parsons and G. Jones of School of Biological Sciences, University of Bristol for providing the echolocation call recordings employed in this research.

References

Cohen L. (1989). Time-frequency distributions - A review. *Proceedings of the IEEE*, 77(7): 941-980.

Kohonen T. (1988). Self-organization and Associative Memory. 2nd Edition Springer-Verlag Berlin-Heidelberg-New York.

Kohonen T. (2001). Self-organizing maps. 3rd Edition Springer-Verlag Berlin-Heidelberg-New York.

Krause B. (2002). Wild soundscapes: Discovering the voice of the natural world. Berkeley, California: Wilderness Press.

Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). Introduction to the theory of statistics, McGraw-Hill International, New York, chapter V.5.

Nikias C.L., and Petropulu A.P. (1993). Higher-order spectra analysis a nonlinear signal processing framework. PTR Prentice Hall, Englewood Cliffs, New Jersey 07632.

Paraskevas, I., and Chilton, E. (2003). Audio classification for retrieval from multimedia databases. *Proceedings of the EC-VIP-MC, 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications*, Zagreb, Croatia, pp.187–192.

Paraskevas, I., and Chilton, E. (2004). Combination of Magnitude and Phase Statistical Features for Audio Classification. *Acoustics Research Letters Online*, 5 (3), 111–117.

Paraskevas, I., Chilton, E., and Rangoussi, M. (2006). Audio classification using features derived from the Hartley transform. *Proceedings of the 13th Intl. Workshop on Systems, Signals and Image Processing (IWSSIP'2006)*, Budapest, Hungary, pp.309–312.

Parsons S., and Jones G. (2000). Acoustic Identification of twelve species of echolocation bat by discriminant function analysis and artificial neural networks. *The Journal of Experimental Biology*, 203, 2641-2656.

Proakis J.G., and Manolakis D.G. (1992). Digital Signal Processing Principles, Algorithms, and Applications. Macmillan Publishing Company.

Rabiner L., and Schafer R.W. (1978). Digital processing of speech signals. Prentice-Hall.

Webb A.R. (2002). Statistical Pattern Recognition. 2nd edition John Wiley & Sons.

Wold E., Blum T., Keislar D., and Wheaton J. (1996). Content-Based classification, search and retrieval of audio. *IEEE Multimedia Fall 1996*, pp. 27-36.

Zhang T., and Kuo C.C.J. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4.