NATURAL SOUNDSCAPES AND IDENTIFICATION OF ENVIRONMENTAL SOUNDS: A PATTERN RECOGNITION APPROACH

I. Paraskevas[1]*, S. M. Potirakis*[2] *and M. 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
ioannis.paraskevas21@gmail.com

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

ABSTRACT

‘Soundscapes’ are maps that depict the sound content of an area at a time interval. Sound features encapsulate information which can be combined with the visual features of a landscape, in order to produce useful ecological observations/data, for areas of environmental or ecological interest. These include monitoring of the wildlife, the inhabitation and the use/human activities of the area, as they evolve with time. In this paper, a method is proposed for the development of a soundscape – a procedure that requires a hierarchical, coarser-to-finer classification scheme for the environmental sounds. The proposed method is illustrated for echolocation calls produced by different species of bats. Time-frequency representations of the sound signals are obtained as a basis for feature extraction. Vectors of statistical features are classified by an Artificial Neural Network classifier. The experimental results verify the potential of the proposed method for classification of environmental sounds within a soundscape development task.

Index Terms— Acoustic ecology, soundscape, sound pattern recognition, time-frequency distributions, coarse / fine classification.

1. INTRODUCTION

‘Acoustic ecology’ is a term first introduced by R. Murray Schafer in 1994, [25]. In the research area of acoustic ecology significant is the contribution of the SEKI group, [26]. One of the objectives of their work is to examine whether environmental sound recordings convey useful information and, consequently, whether features extracted from these recordings can be employed as indicators for the health of biotopes and for the biotopes’ dynamic balance. Recently, the interest for acoustic ecology has increased due to the activities of the World Forum for Acoustic Ecology (WFAE), founded in Canada in 1993.

Research related to the environmental or ecological assessment of landscapes was originally focused on their visual content, e.g., the morphological characteristics of a biotope. However, acoustic ecology has shown that the sound content of the landscape can be employed as a valuable additional information stream in order to characterize or monitor areas of ecological interest, e.g., biotopes. Indeed, sound can provide an additional ecological indicator for such areas, for purposes that include monitoring of the wildlife or the various human activities and their evolution with time. After appropriate signal processing, the large amounts of information required to this end, originally in the raw form of sound recordings, can be presented in the concise yet meaningful form of a sound map or ‘soundscape’.

Soundscapes, [9], [10], are maps of a certain region that present the sound rather than the visual features of this region at a given ‘time instant’ – an appropriately defined short time interval, actually. They are useful tools for nature conservation, [28], because periodic comparison of soundscapes from a certain area – e.g., regions of the NATURA 2000 network (European Union’s network of nature protection areas) – results in significant ecological observations. In contrast to geographic maps that are rarely changed, soundscapes require regular updates because they vary significantly with time. It is interesting to note that natural soundscapes may include sounds that are not necessarily detected by the human ear; therefore,

978-1-4244-3298-1/09/$25.00 ©2009 IEEE

terms such as ‘acoustic’ or ‘audio’ are avoided in such context.

The development of a soundscape is a stepwise procedure including among others a pattern recognition and sound classification step, which is critical to the performance of the overall task. The role of content-based classification has become increasingly important due to the increasing number of audio-visual databases, [30]. 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, [19], [20]. Moreover, there exist situations where visual information is not available and so sound is the only information source for event classification.

2. OUTLINE OF THE PROPOSED METHOD

In this work, a pattern recognition method is proposed for the identification of environmental sounds recorded within an area and for the development of the corresponding soundscape.

The novelty of the proposed method is that it addresses these two tasks under a common framework, through a hierarchical pattern recognition approach.

1. The first step for the development of a soundscape is the collection of sound recordings by microphones that are placed at chosen locations over the whole area of interest. Microphone location for (optimal) coverage of the area is an interesting problem per se. The recorded sounds form the sound database.
2. In the second step, each sound recording is classified – and identified, if possible. Features are extracted by signal processing of the sound recordings, in order to feed the classifier selected for this step.
3. In the final step, the classified / identified recording is placed as a tag on a geographical map of the area, in order to produce the soundscape. Tags are color-coded and placed on a sequence of levels of progressive detail to visualize the soundsacpe on two dimensions (screen or paper).

Classification / identification of environmental sounds, mentioned in the second step above, along with the necessary feature extraction process, is critical for the performance of the overall method. Indeed, the (correct) classification rate depends on how efficiently the feature vectors encapsulate the information content of the signals, [4], [29]. For efficient feature extraction, the signals may be transformed to the frequency domain by, e.g., Fourier transform, [22], or Hartley transform, [1], or to the time-frequency domain, through a time-frequency representation, e.g., the spectrogram or one of the Cohen class members, [3]. The proposed method is based on features extracted from the time-frequency distributions of the signals, [17], [18] and specifically from the Choi-Williams distribution.

The feature vectors formed from the extracted features should introduce their information content to the classifier in a compact manner. Different kinds of classifiers, e.g., distance metric classifiers, Artificial Neural Networks (ANN) etc., can be employed depending on the intrinsic characteristics of the sound classes to be classified. The type of ANN selected here is a Self-Organizing Map (SOM). Self-organizing networks detect regularities and correlations and adjust their future responses according to previous inputs. The SOM architectural details and relative training algorithm are given in [8].

Classification / identification of environmental sounds, is addressed hierarchically here, in levels that proceed from ‘coarse’ to ‘fine’ classification.

- At the first level of the hierarchy, the aim is to group sounds into major classes, differentiated by their source and its characteristics. Three major classes are employed here, namely, human-related, geophysical-related and animal-related sounds, [7].
- At the second level, the aim is to further classify into subclasses sounds that belong to the same major class, e.g., to identify different species of animals within the major class of animal-related sounds, based on the sounds they produce, [21].
- At the third level, sounds within the same subclass are further classified into smaller groups, e.g., sounds produced by bats are classified into the existing families of this species.

It is thus seen that the task of soundscape development includes an inner, hierarchical, three-level pattern recognition task. The development of a soundscape becomes more complicated as the number of sound classes increases at any of these three levels.

Moreover, for a fixed number of classes and subclasses at all three levels, the ‘finer’ second and third levels are yet more demanding, both in terms of the features extracted and of the classifier employed, as compared to the ‘coarse’ first level. It is interesting to note that the majority of the research work for content-based sound pattern recognition belongs to the ‘coarse’ type ([30], [6], [14], [31]) rather than the ‘fine’ type ([17], [19], [20]).

3. FEATURES EXTRACTED FROM ENVIRONMENTAL SOUNDS

Feature extraction is critical for correct pattern recognition. For a general pattern recognition task, features can be extracted from:

- the time domain signal recordings, e.g., zero-crossings rate (ZCR), linear prediction coefficients (LPC), [13], etc.,
- the frequency domain transform of the signal, e.g., voice pitch, [23], cepstral coefficients, [2], bandwidth, and, finally,

- time-frequency representations of the signal, [3], [5], e.g., statistical features, [12], coefficients extracted from the magnitude spectrogram, [24], [20].

Features commonly used for audio pattern recognition include, [31], [30]:

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

Subsets of the aforementioned features, typically used for speech / speaker recognition, have been applied ad hoc to sound classification applications. The features, though, that are employed for speech / speaker recognition are selected on the basis of a-priori information, related to the human speech production model. Hence, they are not appropriate for classes of sounds other than speech. Moreover, in existing research, [31], [30], features employed are either temporal or spectral but scarcely ever time-frequency related.

Time-frequency distributions present the signal spectral content evolution in time. They provide a valid representation for environmental sound signals, because, in contrast to speech or audio, environmental sounds do not typically possess the stationarity (or even short-term stationarity) property that would allow the use of Fourier-type spectral analysis and representation.

The importance of the appropriate signal representation for an efficient feature extraction is illustrated in the following example, where a sound recording consists of two simultaneous sound events, namely, 'birds croaking' and 'waterfall'. Figures 1[a], 1[b] and 1[c] show the time domain signals of the recordings of:

- sound of birds croaking,
- sound of waterfall and
- the simultaneous occurrence of both sound events, i.e., sound of waterfall with birds croaking.

Figures 2[a], 2[b] and 2[c] show the magnitude spectrogram of the sounds presented in figures 1[a], 1[b] and 1[c], respectively. From figure 1[c], it is observed that when two (or more) sound events occur simultaneously, they are not distinguishable using the time domain signal representation. Yet, they become distinguishable when the magnitude spectrogram is employed, as in figure 2[c]. This observation advocates the use of joint time-frequency domain features, such as the magnitude spectrogram, in order to construct the feature vector for classification.

In the proposed method, the feature vectors are formed from statistical features extracted from time-frequency distributions of the sound signals, namely, from the Fourier Magnitude Spectrogram and the Choi-Williams Distribution.

3.1. Time-Frequency Distributions

The proposed method is tested on a set of sound signals from a database of echolocation calls from fourteen (14) species of bats that exist in the United Kingdom (UK).

Probably the most popular time-frequency distribution is the Fourier Magnitude Spectrogram (FMS). The FMS evaluates the Fourier Transform of the signal over a set of consecutive very short periods of time. FMS describes the evolution of the spectral magnitude content of the signal in time, [22], [24]. As an illustrative example, figures 3[a] and 3[b] present the time domain signals and figures 4[a] and 4[b] the corresponding FMS of the echolocation calls produced by two of the bat species examined here.

An alternative time-frequency representation examined as a candidate within the framework of 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, [3]. Specifically, the Cohen's general class of time-frequency distributions is defined as:

$$P_{GEN}(t,f) =$$

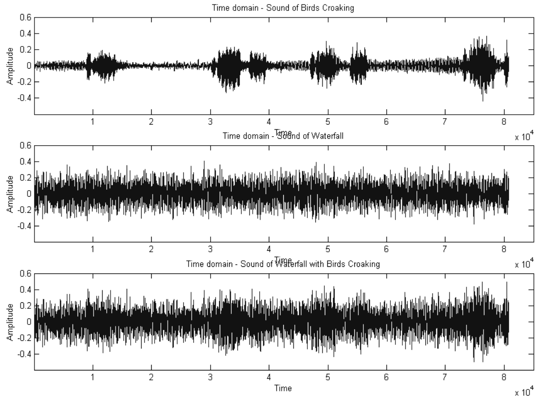
$$= \frac{1}{4\pi^2} \iiint e^{-j\vartheta t - j2\pi f\tau + j\vartheta u} \phi(\vartheta,\tau) x^*(u - \frac{\tau}{2}) x(u + \frac{\tau}{2}) du d\tau d\vartheta$$

(1)

where: x(u) denotes the windowed signal and t, f denote time and frequency, respectively.

Moreover, in equation (1) for $\phi(\vartheta,\tau) = 1$ yields the Wigner-Ville Distribution while for $\phi(\vartheta,\tau) = e^{-\frac{\vartheta^2\tau^2}{\sigma^2}}$ yields the CWD where ϑ and τ denote the Doppler and the delay, respectively and σ is an adjustable parameter which controls the suppression of the cross-terms and the frequency resolution. Note that the complex-conjugate operator (*) is used, as the signals may be in their analytic form, i.e. complex-valued. In practice summations replace integrals for discrete time signals, while appropriate pseudo-forms are calculated for finite length signal sequences. For the present application, the CWD is preferred to the Wigner-Ville Distribution due to its cross-terms reduction property, [16].

Summarizing, the FMS encapsulates only the magnitude content of the signal evolving in time, whereas the CWD encapsulates more complete information related to the signal [11]. Hence, as will be described in Section 4, the FMS is appropriate for the coarser classification task, whereas the CWD is more appropriate for the finer classification task.

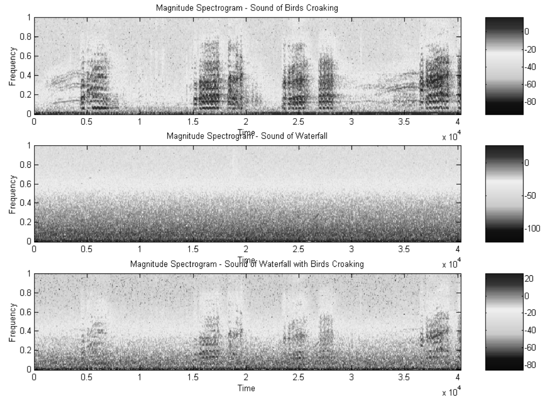


Time domain - Sound of Birds Croaking

Time domain - Sound of Waterfall

Time domain - Sound of Waterfall with Birds Croaking

Fig. 1. Time domain sound signals of – [a] Birds Croaking, [b] Waterfall, [c] Waterfall with Birds Croaking.

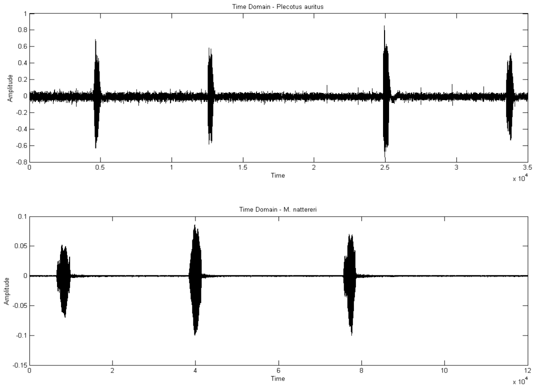


Magnitude Spectrogram - Sound of Birds Croaking

Magnitude Spectrogram - Sound of Waterfall

Magnitude Spectrogram - Sound of Waterfall with Birds Croaking

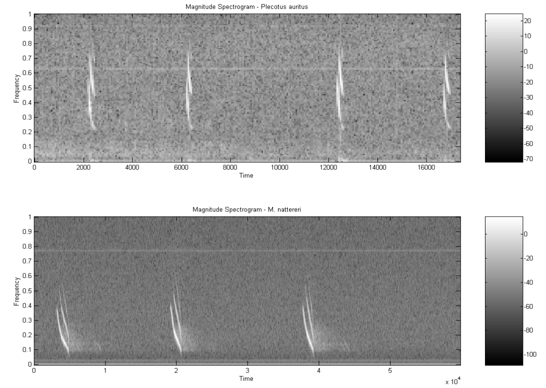
Fig. 2. Magnitude Spectrogram of the sounds in fig. 1, from [a] Birds Croaking, [b] Waterfall, [c] Waterfall with Birds Croaking.



Time Domain - Plecotus auritus

Time Domain - M. nattereri

Fig. 3. Time domain sound signals (echolocation calls) of - [a] the Plecotus auritus bat, [b] the M. nattereri bat.



Magnitude Spectrogram - Plecotus auritus

Magnitude Spectrogram - M. nattereri

Fig. 4. Magnitude Spectrogram of the sounds in fig. 3, from [a] the Plecotus auritus bat, [b] the M. nattereri bat.

3.2. Statistical Feature Vectors

Each sound recording is divided into equal-length frames (256 samples) with zero-padding of the last frame if necessary, windowed by a Hanning window and transformed to the frequency domain. The transformed frames are placed row-wise in a matrix and thus, the time-frequency distributions are formed (FMS and CWD). In order to introduce the matrix to an ANN for classification it has first to be vectorized and subsequently, the dimensionality of this vector has to be reduced for practical purposes. Therefore, statistical features are calculated from each time-frequency distribution vector and for each sound recording, in order to compress the information into a compact feature vector, [15].

The eight statistical features intuitively selected here to form the 8x1 feature vector for each time-frequency (t-f) distribution vector are:

1. Variance of the t-f distribution vector,
2. Skewness of the t-f distribution vector,
3. Kurtosis of the t-f distribution vector,
4. Inter-Quartile Range of the t-f distribution vector,
5. Median of the t-f distribution vector,
6. Mean Absolute Deviation of the t-f distr. vector,
7. Range of the t-f distribution vector, and
8. Log-Entropy of the t-f distribution vector.

Each class of sounds consists of ten recordings which form the training set.

4. EXPERIMENTS AND RESULTS

A pattern recognition experiment, at levels 2 and 3 of the hierarchical scheme described in Section 2, is set up for the case of fourteen (14) species of bats which exist in the UK, based on the echolocation calls they produce (see Acknowledgement for the provision of the recordings).

These fourteen bat species are:

1. Barbastella barbastellus,
2. Eptesicus serotinus,
3. Myotis bechsteinii,
4. M. brandtii,
5. M. daubentonii,
6. M. mystacinus,
7. M. nattereri,

8. Nyctalus leisleri,
9. N. noctula,
10. Pipistrellus pipistrellus,
11. P. pygmaeus,
12. Plecotus auritus,
13. Rhinolophus ferrumequinum and
14. R. hipposideros.

At level 2 classification, signals from all fourteen species are coarsely classified into four Groups defined by the biology of the species, namely:

- Group A: (1) Barbastella barbastellus, (2) Plecotus auritus, (3) Eptesicus serotinus, (4) Nyctalus leisleri and (5) N. noctula,
- Group B: (1) Myotis bechsteinii, (2) M. brandtii, (3) M. daubentonii, (4) M. mystacinus and (5) M. nattereri,
- Group C: (1) P. pygmaeus and (2) Pipistrellus pipistrellus,
- Group D: (1) Rhinolophus ferrumequinum and (2) R. hipposideros.

A SOM-type ANN is employed at this level. It is a Learning Vector Quantization (LVQ) structure with a linear layer of eight (8) nodes (feature vectors consist of eight features – Subsection 3.2) and a non-linear layer of four (4) nodes (subclasses) to produce four (4) outputs (Groups A, B, C and D).

The feature vectors presented to the ANN contain the statistical features of Subsection 3.2, extracted from the FMS of the signals. Classification results at this level are also provided in [21], however they are based on a different feature set. The classification results of level 2 obtained here (based on the Subsection 3.2 feature set) agree with the results in [21].

At level 3 classification, which takes place within each of the four Groups obtained from the previous level, every signal is classified as belonging to a specific bat species. Features extracted from the CWD rather than the FMS are employed here, as this is a finer classification task. Ten (10) sample sound recordings from each one of the fourteen species of bats form the training set of the ANN.

Four (4) different SOM-type ANNs are employed at this level, one for each Group A, B, C, D. The LVQ architecture varies as to the number of nodes of the second (non-linear) layer (five (5) for Groups A and B, two (2) for Groups C and D) whereas, the number of input nodes (linear layer) is eight (8) for all four (4) ANNs, according to the dimensionality of the feature vectors.

The results of this second step, in the form of class representatives projected onto a two-dimensional feature space for illustration purposes, are presented in figures 5[a] and 5[b] - due to space limitation only the results concerning Group A and Group B, are presented here. The circular points of figures 5[a] and 5[b] denote the relative location of the central value of each cluster (i.e. bat species) with respect to the other clusters, for Group A and Group B, respectively.

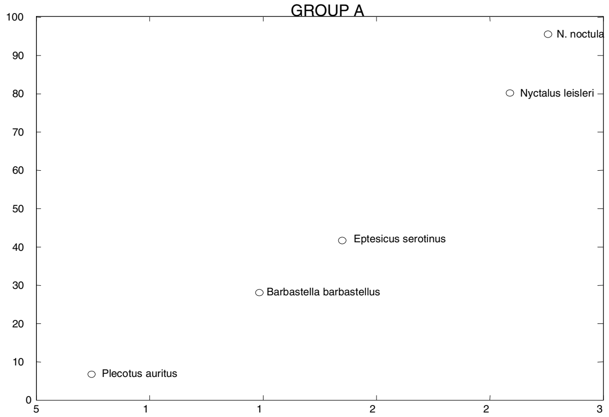


Fig. 5. [a] Level 3 classification of sounds within Group A of bats.

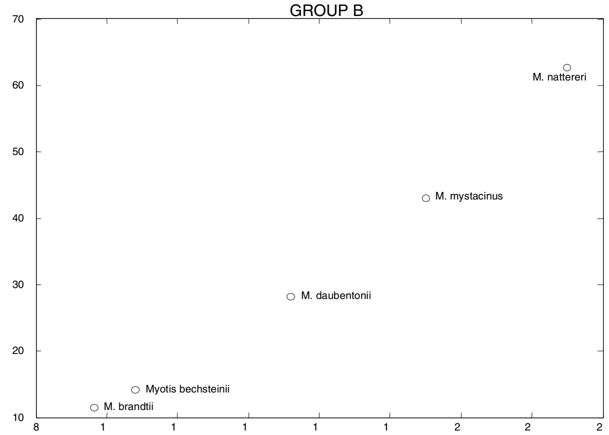


Fig. 5. [b] Level 3 classification of sounds within Group B of bats.

Class representatives are spaced apart in both levels (2 and 3) of classification, indicating that the proposed features, within the feature extraction and classification method adopted here, are suitable for coarse and fine classification tasks of environmental sounds.

Finally, it is important to mention that the aim of this work is to propose a method for the development of natural soundscapes; hence, the preliminary results presented are not yet suitable for comparison with other feature extraction and / or classification methods.

5. CONCLUSIONS AND FUTURE RESEARCH

This work focuses on the task of developing a natural soundscape based on a hierarchical scheme of coarser-to-finer classification of recorded environmental sounds. The proposed method is illustrated by an example, where the echolocation calls produced by fourteen different species of bats are recorded and classified. These classes of sounds belong to the same family and thus constitute a demanding pattern recognition task, due to the similarity they bear. Statistical features extracted from time-frequency representations of the bats' sounds are used to feed an Artificial Neural Network classifier. The experimental results show the potential of the proposed method for the classification of environmental sounds. As the sound clusters produced by the ANNs are not tested on a test set different than the training set, further experimentation is necessary in order to establish that this approach generalizes well for these and other types of sounds.

6. ACKNOWLEDGEMENT

Authors gratefully acknowledge the assistance of S. Parsons and G. Jones of the School of Biological Sciences, University of Bristol, UK, for providing the echolocation call recordings employed in this research [27].

7. REFERENCES

[1] R.N. Bracewell, *The Fourier Transform and Its Applications*, 2nd edition, McGraw-Hill, 1986.

[2] D.G. Childers, D.P. Skinner, R. Kemerait, "The cepstrum: a guide to processing.," *Proc. of IEEE*, pp. 1428–1443, 1967.

[3] L. Cohen, "Time-frequency distributions - A review," *Proceedings of the IEEE*, 77(7), pp. 941-980, 1989.

[4] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Ltd., 2000.

[5] S. Esmaili, S. Krishnan, K. Raahemifar, "Content based audio classification and retrieval using joint time-frequency analysis," *Proc. IEEE-ICASSP*, vol. 5, pp. 665-668, 2004.

[6] J. Foote, "Content-based retrieval of music and audio," *SPIE Multimedia Storage and Archiving Systems (II)*, vol.3229, pp. 138-147, 1997.

[7] S.H. Gage, B.M. Napoletano, M.C. Cooper, "Assessment of ecosystem biodiversity by acoustic diversity indices," *American Institute of Physics, 109, 2430*, 2001.

[8] T. Kohonen, *Self-organizing maps.* 3rd edition Springer-Verlag, Berlin-Heidelberg-New York, 2001.

[9] B. Krause, "Bioacoustics, habitat ambience in ecological balance," *Whole Earth Review*, 57, 1987.

[10] B. Krause B, *Wild soundscapes: Discovering the voice of the natural world*, Wilderness Press, Berkeley, California, 2002.

[11] V. F. Kravchenko, O. V. Lazorenko, V. I. Pustovoit, L. F. Chernogor, "Choi-Williams transform and atomic functions in digital signal processing," *Doklady Physics, MAIK Nauka / Interperiodica distributed exclusively by Springer Science*, vol. 52, number 4 / pp. 207-210, 2007.

[12] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," *Proc. IEEE-ICASSP*, vol: 6, pp. 3621 – 3624, 1998.

[13] J. Makhoul, "Linear Prediction: A tutorial review," *Proc. IEEE*, pp. 561-580, 1975.

[14] L. Mingchun, W. Chunru, "A study on content-based classification and retrieval of audio database," *Intl. Symposium on Database Engineering & Applications*, pp. 339 – 345, 2001.

[15] A.M. Mood, F.A. Graybill, D.C. Boes, *Introduction to the theory of statistics*, McGraw-Hill International, New York, chapter V.5, 1974.

[16] C.L. Nikias C.L., A.P. Petropulu, *Higher-order spectra analysis a nonlinear signal processing framework*, PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[17] I. Paraskevas, E. Chilton, "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, 2003.

[18] E. Chilton, I. Paraskevas, "Audio fine classification using the statistical analysis of acoustic images," *145th Meeting of the Acoustical Society of America*, 113, 2271, Nashville, USA, 2003.

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

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

[21] S. Parsons, G. Jones, "Acoustic Identification of twelve species of echolocation bat by discriminant function analysis and artificial neural networks," *The Journal of Experimental Biology*, vol. 203, pp. 2641-2656, 2000.

[22] J.G. Proakis, D.G. Manolakis, *Digital Signal Processing Principles, Algorithms, and Applications,* Macmillan Publishing Company, 1992.

[23] L. Rabiner, M. Cheng A. Rosenberg, C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans ASSP*, vol. 24(5), pp. 399–418, 1976.

[24] L. Rabiner, R.W. Schafer, *Digital processing of speech signals*. Prentice-Hall, Englewood Cliffs, New Jersey , 1978.

[25] R.M. Schafer, *The soundscape: Our sonic environment and the tuning of the world,* Destiny Books, Rochester, Vermont, 1994.

[26] SEKI Group web site: http://envirosonic.cevl.msu.edu/seki, 2008.

[27] The Bats of Britain web site:
http://www.bio.bris.ac.uk/research/bats/britishbats/index.htm

[28] M.G. Turner, R.H. Gardner, R.V. O'Neill, *Landscape ecology in theory and practice: Pattern and process,* Springer-Verlag, New York, Inc., 2001.

[29] A.R. Webb, *Statistical Pattern Recognition*, 2nd edition John Wiley & Sons, 2002.

[30] E. Wold, T. Blum, D. Keislar, J. Wheaton, Content-Based classification, search and retrieval of audio, *IEEE Trans on Multimedia*, pp. 27-36, 1996.

[31] T. Zhang, C.C.J. Kuo, Audio content analysis for online audiovisual data segmentation and classification, *IEEE Trans ASSP*, vol. 9, no. 4, 2001.