INSTRUMENT FOR SOUNDSCAPE RECOGNITION, IDENTIFICATION AND EVALUATION (ISRIE): SIGNAL CLASSIFICATION

J Stammers Department of Electronics, University of York, York, YO10 5DD
D Chesmore Department of Electronics, University of York, York, YO10 5DD

1 INTRODUCTION

The ISRIE project is a collaboration between the universities of York and Newcastle, and ISVR in Southampton. Work at York is split between two projects; one focusing on signal separation and the other on identification. This paper describes the current work being undertaken on the latter of these two subjects and begins by briefly describing how the ISRIE project arose and its intended outcome. A review of the related literature is given which covers previous projects dealing with signal classification. There will follow a brief discussion of the types of sound ISRIE will aim to classify. The current techniques being investigated will be described along with preliminary results. Finally, conclusions are drawn and the proposed plans for the future of this research are presented.

1.1 ISRIE

The ISRIE project arose from the EPSRC Ideas Factory ‘A Noisy Future’. The proposed outcome of the project can be described briefly as an intelligent noise metering system able to determine the direction and source from which a sound originated. It will also be able to provide other details such as the time at which the sound occurred and how loud the sound was. If a number of these instruments are used as a network of sensors it should be possible to estimate the location from which a sound originated.

The primary motivation for such an instrument is to assist in urban noise level measurements, whether for research or for legislative purposes. At present detailed investigation of a sonic environment involves either attended monitoring or long-duration recordings and analysis of this data. Both of these methods are very time consuming and require the full attention of an individual and introduce the problem of subjectivity. An instrument such as that proposed by the ISRIE project would perform listening and evaluation in-the-field removing the requirement for an individual to be present or recording of large quantities of data. The instrument would also be capable of delivering a highly objective analysis of the soundscape. An example of where ISRIE could have been of use is an analysis of the occurrence of oversnow vehicles in Yellowstone National Park[1]. In this study audio samples were recorded in the field and then analysed in an office environment. Burson describes how the volume of playback for some recordings had to be increased by 10dB to approximate the audibility that would have been available in the field. ISRIE would remove both the need for level boosting and the need for time consuming analysis of large quantities of recorded audio.

1.1.1 ISRIE and Noise Legislation

It is thought that ISRIE will be of great benefit to those whose work is concerned with noise legislation, namely Planning Policy Guidance (PPG) 24 and BS 4142. PPG 24 is concerned with evaluating noise exposure to noise-sensitive developments and BS 4142 is concerned with rating industrial noise affecting both residential and industrial areas. When a noise complaint needs to be investigated ISRIE could be used in place of a person collecting data manually.

PPG 24 uses four noise exposure category (NEC) bands to describe a sound. Where a measurement is placed within these bands is dependant on the contribution of each of the four noise categories; road traffic, rail traffic, air traffic and mixed noise sources. Currently the contribution of each of these categories to determine where a measurement is placed is performed by a person and could therefore be quite subjective. ISRIE would be able to provide a purely objective NEC band recommendation. Analysis for BS 4142 would also benefit from ISRIE as it would remove the excessive labour in performing the numerous measurements required.

1.2 Sound Categories

As part of the signal classification research it was deemed necessary to develop an acoustic taxonomy into which sounds would be categorised. Initially it was thought that the quantity of categories would be very high (given how many different sounds can be observed in an urban environment). However, as a result of discussing analysis of PPG 24 and BS 4142 with the project partners from ISVR (Southampton) the categories given in Figure 1 have been decided upon as the most important to be identified.

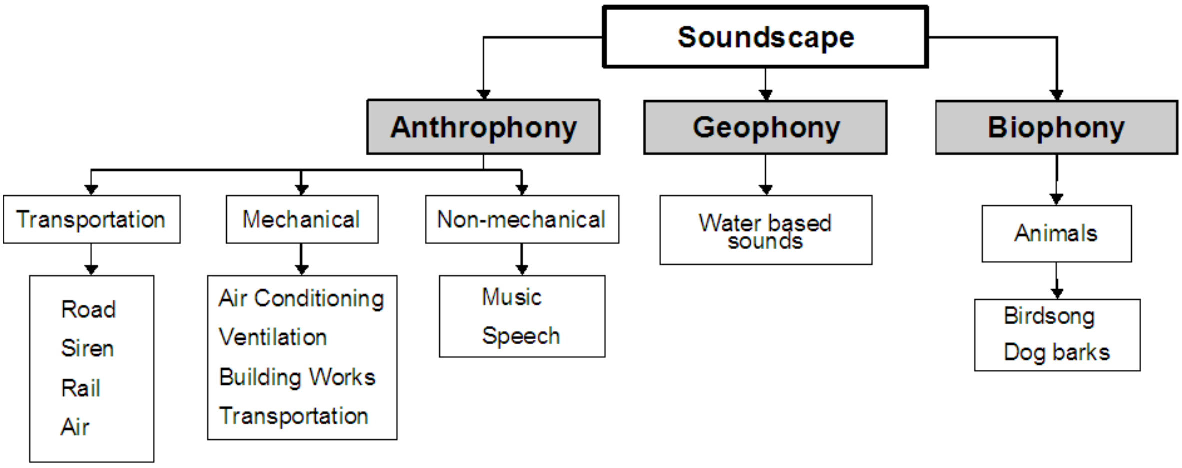

Figure 1: Acoustic taxonomy. The categories were derived from the sounds seen as having most influence on PPG 24 and BS 4142 measurements.

These sounds have been grouped into the main categories *anthrophony* (sounds related to human activities), *geophony* (sounds caused by nature), and *biophony* (sounds caused my animals). There are some sounds that may be seen as missing from this diagram. Under the Geophony category other natural based sounds, most of which are caused by the wind, could also be included. However, it was pointed out that when wind speeds are in excess of 5 m/s an acoustic consultant would often not perform a measurement as the noise induced on the sensor by the wind is too great. This situation will also apply to ISRIE so wind-induced sounds are not included. Other sounds considered missing from this diagram are not included because they are simply not loud enough to have an impact on a soundscape. The incidental sounds made by humans and animals are a good example of this. There is little in the literature regarding categorisation of urban audio signals. The reason for this may be that many sound classification projects focus on a small set of sounds or a particular species of animal. There are 2 good examples of categorisation in the literature. Raimbault and Dubois[2] use a tree-like structure which broadly splits the categories into *transportation/works* and *people presence*. These are then further subdivided to include some of the typical sounds found in an urban soudscape. It is somewhat strange to find the subcategories of *running water* and *birds singing* under *people presence* as these are not strictly due to the presence of humans. The other categorisation found in the literature is very similar to the idea generated at York[3] (Figure 1). The approach that Gage et al. use assigns specific frequency bands to each of anthrophony, geophony and biophony. This may cause a mis-categorisation because there are biophonic sounds that will fall below the 2.5 kHz low-end frequency of the biophony category.

2 CLASSIFICATION SYSTEMS

A classification system is typically a 2-stage process and consists of a feature extractor and a classifier[4]. The flow diagram in Figure 2 describes the basic structure of a classification system. Some systems also use post-processing after the classifier to correct for any errors in classification.

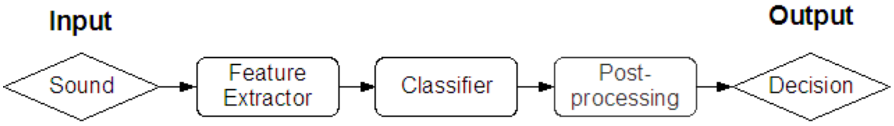

Figure 2: Flow diagram showing the basic structure of a classification system. Adapted from [5].

There are many feature extraction techniques and classifiers to choose from and the selection of each of these is usually based on the intended application and any prior knowledge. Table 1 gives examples of both feature extractors and classifiers found in the literature surrounding the topic of audio signal classification.

Feature Extractor	**Classifier**
Fourier transform (Fast/Short)	Multilayer Perceptron
Wavelet transform	Self-Organising Map
Wigner-Ville distribution	Learning Vector Quantisation
Time-Domain Signal Coding	Time Warping

Table 1: Some examples of feature extractors and classifiers often seen in the literature

2.1 Previous Studies of Sound Classification

There are many studies to be found in the literature which aim to perform classification of signals, whether they are audio signals or some other wave-based signals, such as an ECG signals. There are numerous studies which look at identifying different species of animal based on the sounds they emit, both incidental and deliberate sounds are considered. Some examples of animals identified are wood-boring insects[6], crickets[7], frogs[8], and birds[9]. Each of these studies show that it is possible to discriminate between very similar vocalisations which gives promise in the context of ISRIE as there will undoubtedly be similar sounds occurring in an urban environment. Cowling and Sitte[10] provide an excellent comparative examination of various feature extractors and classifiers for the recognition of environmental sounds. The aim of the study was to find which feature extractor-classifier pair provided the best classification results for a sound surveillance system. They found that using a continuous wavelet transform for feature extraction with a dynamic time warping classifier provided the best results (70% accuracy). In a novel approach to environmental sound classification[11] the goal is similar to that of ISRIE; to develop a system capable of giving an efficient representation of an acoustic environment. This is study is novel in that it uses a system based on genetic algorithms (GAs) to determine which features to extract from a sound. Using this method in combination with 2 different classifiers (a Gaussian mixture model and a k-Nearest Neighbour algorithm), classification results between 90% and 95% are achieved. No specific data is given but it is suspected that an approach using GAs will be computationally expensive and therefore not suited to ISRIE. A study looking at *sound textures*[12] uses a Self-Organising Map fed directly by the audio data without any feature extractor. The data is fed in 2^n $(1 < n < 4)$ samples at a time and the SOM produces a histogram output. The theory behind this approach is that for a new input signal the trained SOM will produce a similar distribution to the distribution of the signal for which it was trained. This approach works very reliably to determine if a test signal is the same or different to the training signal but it cannot classify the test signal into a category.

All of the studies discussed above show that it is possible to perform accurate classification with a variety of audio signals. These can either be animal vocalisations or general audio signals such as those which may be found in an urban environment. In the following section the current approach to the classification problem being studied at York is described.

3 CURRENT APPROACH

The current approach being adopted at York is to use Time-Domain Signal Coding (TDSC) feature extraction and a Self-Organising Map (SOM) classifier. TDSC has been chosen because it is very computationally inexpensive and has had excellent success rates when used for classification of species[6,13]. A SOM has been chosen for use as an initial classifier partly due to its ease of use but also because it is easily expandable. If the SOM receives a signal it has not seen before it could potentially create a new output unit (and hence a new output class) for that input so that if a similar signal is presented it will be classified accordingly. TDSC is discussed in more detail below.

3.1 Time Domain Signal Coding

Time Domain Signal Coding is a purely time-domain technique based on a speech compression method known as Time Encoded Speech[13]. In TDSC analysis signals are segmented using zero crossings of the time-domain waveform. The data between successive zero crossings is termed an *epoch*. Each epoch can then be described by its shape (S - the number of minima) and its duration in samples (D) and a whole signal can then be described by its D-S pair characteristics. Figure 3 shows a simple example of D-S characteristics.

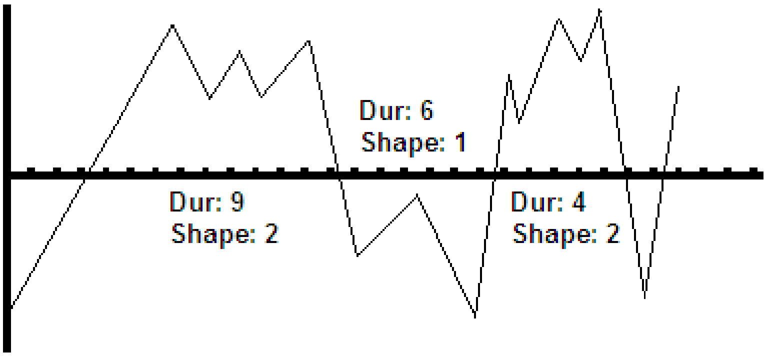

Figure 3: Simple example of D-S characteristics extracted using TDSC. The X-axis shows the sample intervals and the Y-axis represents amplitude.

In this example the first epoch has duration of 9 samples and a shape of 2. A signal analysed using TDSC will often have a large number of D-S pairs. These can be mapped onto a smaller symbol set (termed the *codebook*) to make processing easier. The codebook can then be used to generate either a 1- or 2-dimensional histogram describing the occurrence of D-S pairs. The 1-dimensional variant is called the *S-matrix* where the X-axis represents the code and the Y-axis the frequency of occurrence for each code.

3.1.1 Duration-Shape Distribution

In previous studies using TDSC the codebook has been generated manually to suit the application containing some 30 codes[6,13]. However, this is not possible for the application to the general sounds found in an urban environment (shown in Figure 1) due to the large variation in D-S distributions. To identify how much these distributions varied by, plots were generated of D versus S as shown in Figure 4. These plots were resized to show the areas containing the most information as the D-S distributions form very sparse matrices. The maximum D-S pairings found for each of these sounds were: A/C unit D=1468 S=165, Blackbird D=218 S=56, Digger D=434 S=68, and Human speech

D=1086 S=161. Using the speech recording as an example, there was 1 occurrence of the maximum D-S pair whereas the most frequently occurring D-S pair (D=2 S=0) was detected 9813 times. This illustrates that some sounds have anomalies present in the recordings which may not be characteristic of that sound.

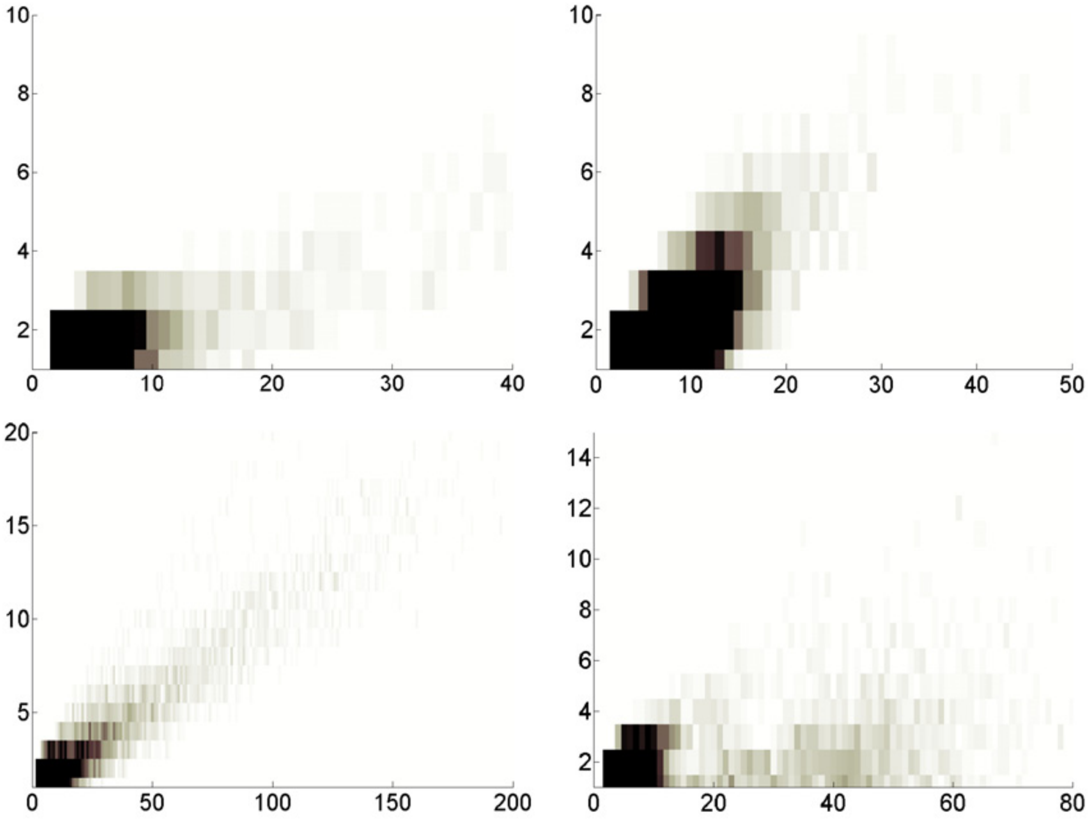

Figure 4: Distributions showing D (X-axis) versus S+1 (Y-axis). Clockwise from the top-left, the plots represent the following sounds: an air conditioning unit, a blackbird singing, human speech and a digger. The darker areas represent higher frequencies of occurrence for that D-S pair. *Recordings of the A/C unit and speech courtesy of ISVR, Southampton. All recordings sampled at 44.1 kHz.*

3.1.2 Generation of a general codebook

Based on the findings of plotting D-S distributions for a number of sounds it was decided that the maximum duration that would be useful to detect would be 1000 (using a sample rate of 44.1 kHz) and the maximum shape would be 75. This gave a very large total number of D-S pair combinations (>50,000). To reduce this number 40 duration ranges were devised based on the findings from the D-S distributions. The first 19 of these ranges had a range of 1 as most D-S data occurs within D<20. The subsequent ranges spanned durations of 10, 50 and 100 samples. After calculating these ranges the total number of D-S pair combinations was significantly reduced to 1700.

3.2 Classification of the Audio Data

Initial research has focussed on classifying audio recordings of the sounds given in Figure 1 using TDSC for feature extraction and a SOM to classify the resulting time domain data. Each audio sample was typically of 2 seconds in duration and analysed in 0.2 second frames. Each frame then had an S-matrix associated to it and these were used as the input to the SOM. Figure 5 illustrates this process.

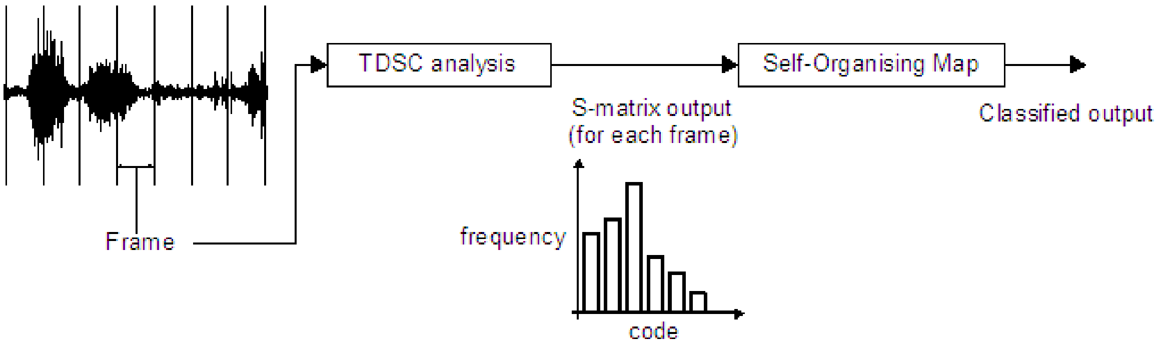

Figure 5: The classification process

During training of the SOM it was noticed that the winning unit was consistently either the highest or lowest numbered unit in the network. This result was not dependant on the input TDSC data; each input gave the same result. The SOM was tested with a smaller, controlled data set consisting of only one value and it performed as expected; the weights for one particular unit in the network approached the value of the input data after a number of training epochs. This showed that the SOM network was functioning as expected. Attention was then turned onto the S-matrices being produced by the TDSC algorithm. Figure 6 shows a plot of how the data is spread out through the codebook. It is quite clear to see that the arrays containing the code frequencies are very sparse. Analysis of the actual data contained in the arrays has shown that the number of cells with a value of zero averages 95%. This figure is similar for a variety of audio data typical to an urban environment. It is believed that this sparseness is the reason for the SOM continually training its weights toward zero and producing the same winning unit even for different initial audio data.

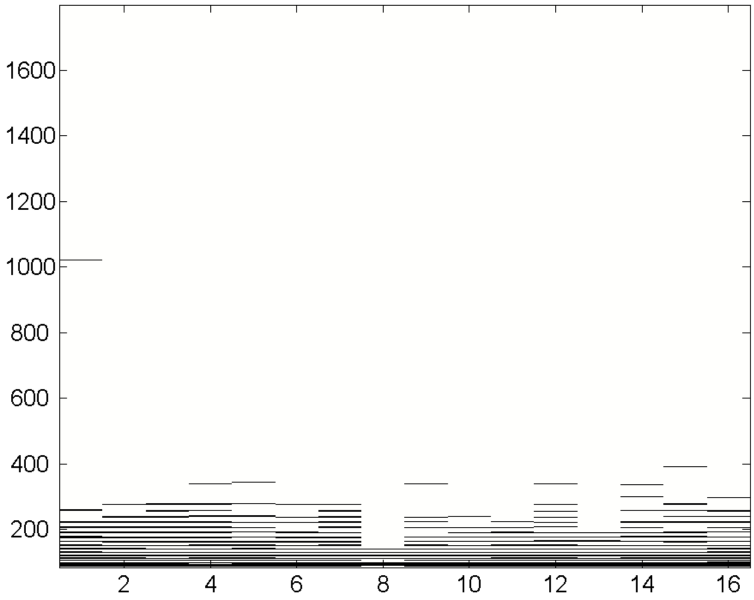

Figure 6: A plot showing the sparseness of the output for TDSC analysis for a recording of a blackbird. The X-axis shows the code number, the Y-axis shows the time frame number and the dark areas show which codes occur in each frame.

4 CONCLUSIONS AND FURTHER WORK

This paper has presented the current work on signal classification as part of the overall ISRIE project. What the ISRIE project is and what it is aiming to achieve has been described. Via discussions with project partners and analysis of PPG 24 and BS 4142 the set of sound categories presented in this paper has been defined to classify audio data into. Prior to discussing the current approach to signal classification a short description of how a classification system is typically constructed was given along with a brief review of previous research in the related areas.

The approach currently being used for signal classification has been described. This consisted of a time-domain signal coding (TDSC) feature extractor producing S-matrix data as an input for a self-organising map (SOM). It was clear to see from the distributions given in Figure 4 how D-S data varies from one sound to another showing that TDSC is a useful feature extractor. It is also clear to see from these distributions that the majority of D-S data is contained in the region D<25 S<10. It was shown that the general codebook generates very sparse arrays of data causing erroneous classification by the SOM. It is therefore necessary to review the arrangement of this general codebook to better describe the salient features of the D-S distributions in the region given above. Other methods of compressing the D-S information into a codebook to reduce the array sparseness will be investigated.

4.1 Further Work

Research will continue using the TDSC and SOM classification system described above. Other feature extraction methods are also due to be studied and tested, Wavelet transforms are of particular interest. A comparison of the classification accuracies produced will be made to see if any improvement can be had by using a different feature extractor.

It is intended that the classification process shown in Figure 5 will be expanded by using the output of the SOM as an input to a syntactic pattern recognition system (SPR - see [14] for an introduction to syntactic methods). The combination of a TDSC feature extractor and SOM classifier can produce an output class variation with repeating patterns. This is illustrated in Figure 7 which used a recording of a cricket as the input to the TDSC-SOM classifier. The change from class 1 to class 4 occurs in a repeating pattern throughout in line with the song of the cricket and the silence between chirps. It has been proposed that this sort of data could be used as the grammar in a syntactic pattern recognition system. Further analysis of the SOM output for various audio recordings will demonstrate if a SPR approach is a feasible solution in the context of the ISRIE project.

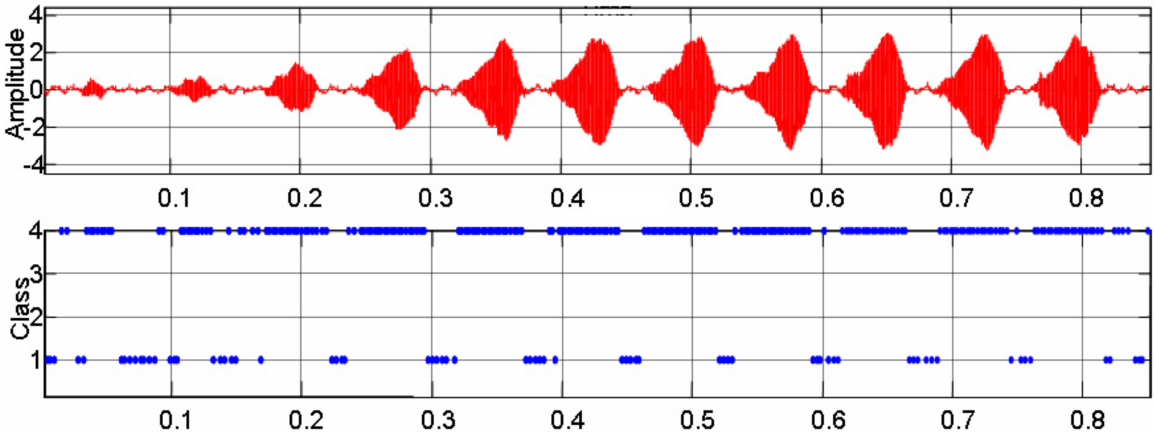

Figure 7: The top diagram shows the time-domain waveform of the cricket song. The lower diagram shows the pattern of the class output as the signal changes.

5 REFERENCES

1. S. Burson, Natural Soundscape Monitoring in Yellowstone National Park December 2005-March 2006, Grand Teton National Park Soundscape Program Report No. 200601 (2006)
2. M. Raimbault and D. Dubois, ‘Urban soundscapes: Experiences and knowledge’, Cities, Vol. 22, 339-350, (2005)
3. S.H. Gage, R. Maher and G. Sanchez, ‘EcoEARS: Ecological & Environmental Acoustic Remote Sensor – Application for Long-Term Monitoring and Assessment of Wildlife’, Technical Symposium & Workshop: Threatened, Endangered and At-Risk Species on DoD and Adjacent Lands, (2005)
4. R. Beale and T.O. Jackson, Neural Computing: An Introduction, 1st ed reprint, Hilger, (1998)
5. S. Allegro, M.C. Büchler and S. Launer, Automatic sound classification inspired by auditory scene analysis, Eurospeech (2001)
6. I. Farr and D. Chesmore, Automated bioacoustic detection and identification of wood-boring insects for quarantine screening and insect ecology, Proc. Institute of Acoustics, Vol. 29, Pt. 3 (2007)
7. C. Dietrich, G. Palm, and K. Riede, and F. Schwenker, ‘Classification of bioacoustic time series based on the combination of global and local decisions’, Pattern Recognition, 37(12) 2293-2305, (December 2004)
8. C. Lee, C. Chou, C. Han, Chin-Chuan and R. Huang, ‘Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis’, Pattern Recognition Letters, Vol. 27, 93-101, (2006)
9. A. Selin, J. Turunen and J.T. Tanttu, ‘Bird sound classification and recognition using wavelets’, Dissertationes Classis 4: Historia Naturalis, Vol. 47 (2006)
10. M. Cowling and R. Sitte, ‘Comparison of techniques for environmental sound recognition’, Pattern Recognition Letters, Vol. 24, 2895-2907, (2003)
11. B. Defréville, P. Roy, C. Rosin and F. Pachet, Automatic recognition of urban sound sources, Audio Engineering Society 120th Convention (2006)
12. M.J. Norris and S.L. Denham, ‘Sound texture detection using Self-Organizing Maps’, Center for Theoretical and Computational Neuroscience, University of Plymouth, UK, (November 2003)
13. D. Chesmore, ‘Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals’, Applied Acoustics, Vol. 62, 1359-1374 (2001)
14. H. Bunke (ed) and A. Sanfeliu (ed), Syntactic and structural pattern recognition – Theory and applications, World Scientific, 3-28, (1990)

6 ACKNOWLEDGEMENTS

The Instrument for Soundscape Recognition, Identification and Evaluation (ISRIE) project is funded by an Engineering and Physical Sciences Research Council grant (reference EP/E009581/1). The ISRIE project is a collaboration between 3 institutions and the discussions with Stuart Dyne (ISVR), Christos Karatsovis (ISVR), Gui Yun Tian (Newcastle) and Hidajat Atmoko (Newcastle) as well as the other members of the Applied Bio-Acoustics group at York have been extremely helpful in allowing this research to progress.