euronoise

Instrument for Soundscape Recognition, Identification and Evaluation (ISRIE): Signal Classification

Jon Stammers and David Chesmore

University of York, Department of Electronics, Heslington, YO10 5DD York, UK
js185@york.ac.uk

ISRIE is a collaborative project between the universities of York and Newcastle and ISVR in Southampton. The work being undertaken at York is in its second year and focuses on signal separation and classification. Developing novel methods for classifying urban and other sounds into distinct categories (such as transportation, industrial, human, animal, etc.) is the focus of the work detailed in this paper. The classification system will initially consist of 2 main parts: a feature extractor and a classifier. Results from this basic system will be presented and a discussion given on how the system will be expanded. It is envisaged that eventually the system will use some form of syntactic pattern recognition to perform the identification of individual sounds.

1 Introduction

The ISRIE project arose from the EPSRC Ideas Factory 'A Noisy Future'. The proposed outcome of the project can be briefly described 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 in a sensor network it should be possible to estimate the location of the sound source. Such an instrument would be a useful tool in urban noise level measurements, either for research or for legislative purposes. This can be a very time consuming process when performed manually and can also be subjective if soundscape content is also being examined. More details on the ISRIE project and its application to legislative procedures can be found in [1].

1.1 Sound Categories

The signal classification part of the ISRIE project has identified the key sounds that are to be recognised within an urban soundscape. These are shown in Fig. 1. The decision to focus on these sounds for the final system was based on discussions with project partners and current noise legislation in the UK.

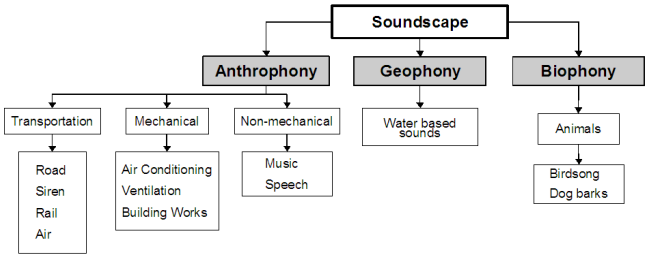

Fig. 1 The relationship between the key sounds to be identified.

The soundscape has been split into 3 main categories: *Anthrophony*, relating to sounds made by humans; *Geophony*, naturally occuring sounds; and *Biophony*, sounds made by animals. Only the most prevalent sounds that would be heard in an urban soundscape have been included. So under the category of *biophony* only birdsong and the bark of a dog have been included because other animal sounds are unlikely to exceed the background noise level of an urban environment. Other sounds that could be included under *geophony* are likely to be caused by the interaction between an object and the wind (a tree, for example). These sounds are not included in the diagram because it was pointed out that when wind speeds exceed 5 m/s an acoustic measurement is unlikely to be taken because of the noise induced on the sensor by the wind. A similar method of breaking down the soundscape has been seen in [2]. The approach taken by Gage et al. applied frequency divisions to separate the 3 main categories. This could lead to mis-classification as not all sounds found in each of the categories will necessarily adhere to the frequency bands.

2 Classification Systems

Classification systems typically consist of 2 main components – a feature extractor and a classifier [3]. The role of the feature extractor is to reduce the complexity of the data being input to the classifier to optimise the classifying process [4]. There are many examples in the literature of classification systems for the analysis and classification of both audio and other wave-based signals. The range of techniques used and applications vary considerably from wavelet feature extraction and multi-layer perceptron network classifiers for human bowel-sound monitoring [5] to time-domain and Mel-frequency techniques for species identification [6,7,8]. The area of environmental sound analysis has also had a lot of development. Cowling and Sitte [9] provide an excellent overview of techniques as applied to a sonic security system. Their research found that a continuous wavelet transform feature extractor coupled with a dynamic time warping classifier gave the highest recognition accuracy (70%). A novel approach to environmental sound recognition is found in the work of Defréville et al. [10]. Their work focussed on using genetic algorithms to find problem-specific features for each individual signal. The results of this method are promising (~90% accuracy) but the signal processing techniques discovered are very complex.

To date, of the many feature extractor and classifier methods available Time-Domain Signal Coding for feature extraction and a Self-Organising Map have been implemented to make up the classification system.

Time-Domain Signal Coding (TDSC) is a feature extraction technique which focuses purely on the time-domain representation of an audio signal. The waveform is seperated into epochs (the signal data between two consecutive zero crossings) and each of these are analysed in terms of shape (S) and duration (D). The shape of an epoch is determined by how many positive or negative minima it contains and the duration is simply the length of the epoch in samples. Further details of how TDSC is performed can be found in [6,7]. TDSC has previously been used for monitoring of machinery and heart sound analysis [6]. In its application to species recognition TDSC has achieved 100% classification accuracy for 13 different Cricket species.

3 Application of TDSC

It was mentioned above that a TDSC feature extractor was coupled with a Self-Organising Map (SOM) classifier for the initial system development. Details of SOMs and their implementation can be found in [11]. The main focus of this work has so far been to produce an output from the TDSC algorithm which is suitable for classification by the SOM. The duration-shape (D-S) information gathered by TDSC is typically organised into a codebook representing a range D-S combinations. The S-matrix is an array of data which associates a frequncy of occurrence to each of these combinations. In previous studies using TDSC the codebook has been manually designed for the application, typically giving ~30 codes. To generate a suitable codebook for containing urban sound data distributions were produced of D-S combinations for various sounds. Fig. 2 shows an example of one such distribution.

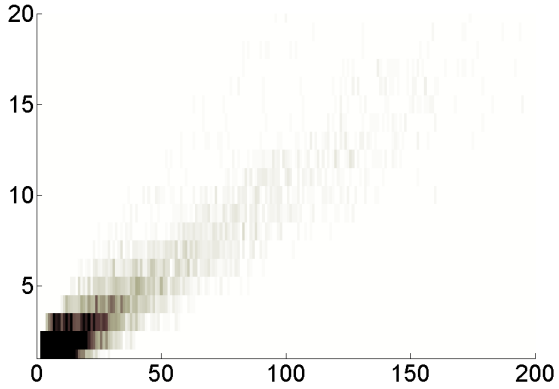

Fig. 2: D-S distribution for a recording of a building site digger. The x-axis represents D and the y-axis represents S.

The maximum D-S pairing found was for an air conditioning unit with D=1468 and S=165. Based on this and other results it was decided to limit D to 1000 and S to 75. Using all possible combinations of these D and S ranges would give a codebook with a very large order of magnitude (>50,000). This number was reduced to 1700 by deviding duration ranges to fit the data into.

Fig. 3 shows the initial classification system used. The audio signal was broken into frames and each of these were analysed seperately using the TDSC algorithms. The TDSC output data for each frame was then classified by the SOM to give a class output for each frame.

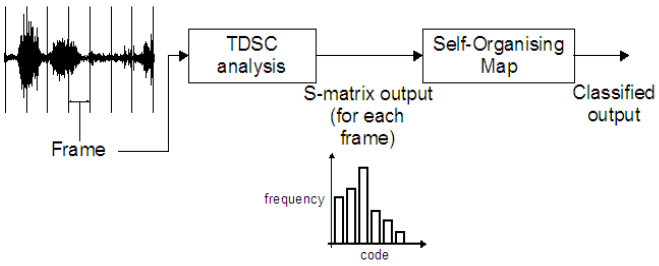

Fig. 3: The arrangement of the initial classification system.

Initial results using the codebook discussed above were disappointing. The SOM consistently gave the same winning units for all sounds. Upon analysis of the TDSC output it was found that the S-matrices had an average sparseness of 95%. For this reason the SOM was struggling to differentiate between the S-matrices generated for audibly different signals.

Further inspection of D-S distributions (Fig. 2) showed that further reduction of the maximum D and S values was possible without sacrificing significant amounts of data. A codebook with a size of 340 was achieved using D=150 and S=15. Results using this codebook are presented below using recordings of sounds found in an urban setting and recordings of some Cicadas.

3.1 Cicada Classification

High quality recordings of 3 different species of Cicada were made available to test the system. It was decided to experiment with these recordings for two reasons: a) the Cicadas are difficult to differentiate by ear so it would be a good test to see if the system could; and b) there is interest in developing a real-time system for the identification of different Cicada species. A total of 24 recordings were used – 3 in which the species were known (training set) and 21 unknown (for testing). The framelength used in the TDSC analysis was 0.2 seconds. It was decided to use a 10 unit SOM for classification.

Fig. 4 shows a plot of the class outputs for each frame of the known recordings. It is clear to see from this plot that a very simple decision rule (perhaps based on an LVQ method) would allow seperation of the 3 different species of Cicada. *Flamatus* appears only in classes 1-3, *japonicus* in classes 4-6 and *biahamatus* dominates classes 8-10. Fig. 4 also shows the class outputs for one of the unknown recordings. By visual inspection it is clear that this particular recording would be placed in the *bihamatus* category. Overall, the system comprising of a TDSC feature extractor and a SOM classifier achieved a classification accuracy of 95%.

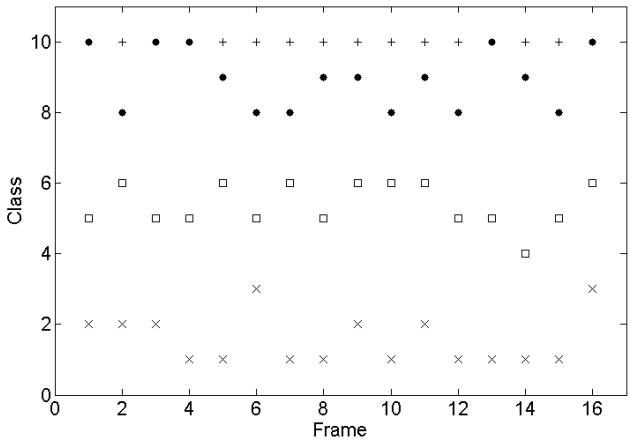

Fig. 4: Frame-wise representation of SOM class output. The different classes are *bihamatus* (●), *japonicus* (□), *flamatus* (x) and test data (+).

3.2 Urban sound classification

Using the same system as that described above, classification of urban sounds was experimented with. The recordings used were of some of the sounds given in Fig. 1

(air conditioning unit, single motor vehicle, birdsong and building works). There were not as many recordings of each of these sounds available as there were for the Cicadas but the resulting data discussed below is still useful.

The recordings of each type of sound were seperated into 6 second sections (of the available data this provided 2 or 3 different recordings for testing). Initially a framelength of 0.1 seconds was chosen. As little is known of the significance of framelength in this application, 0.1 seconds was chosen as a starting value. The same theory applies to the number of output classes chosen for the SOM . In this instance 40 classes were used.

A plot of class output for each frame of a building site recording is shown in Fig. 5. The prominent sound in this recording was a large caterpillar-track-driven digger interspersed with some road noise.

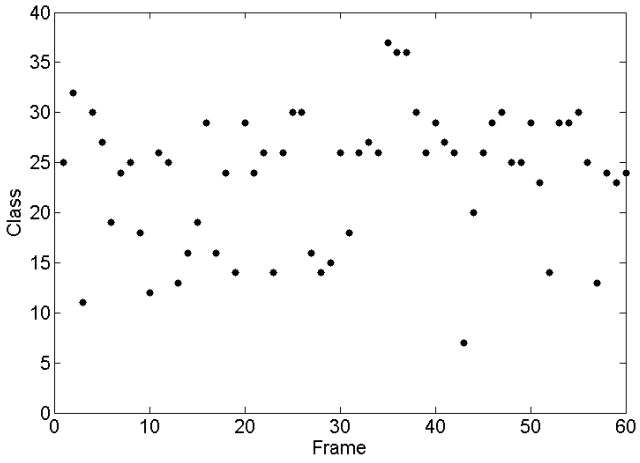

Fig. 5: Framewise SOM class output for a 6 second building site recording.

It is clear to see from this plot that there is no clear banding of class output as there was for the Cicada recordings. Plots for the other sounds listed above produced very similar results, i.e. no obvious class dominance. These disappointing results influenced the decision to start looking at how framelength affected the SOM class output plots. Reducing the framelength had the effect of increasing the apparent lack of structure seen in Fig. 5. Increasing the framelength to 0.5 seconds produced plots that were more promising. Fig. 6 shows the result of using the longer framelength with the same recording as used to produce Fig. 5 and another recording of a similar soundscape.

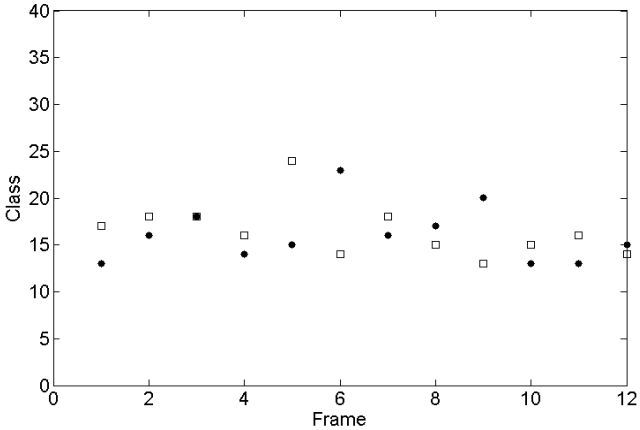

Fig. 6: SOM class output for 2 similar building site recordings; (●) uses the same audio data as that in Fig. 5 and (□) is shown for comparitive purposes.

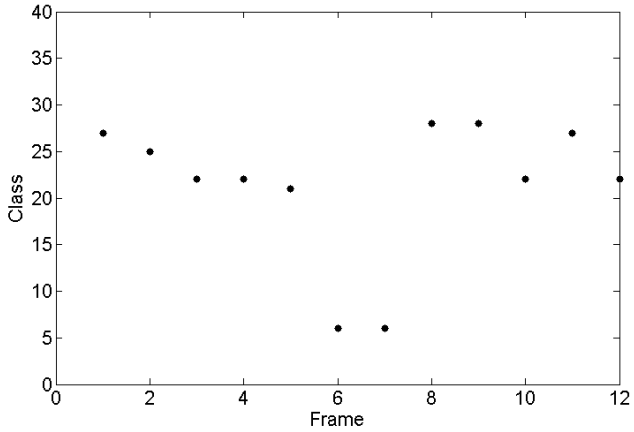

Fig. 7: SOM class output for a blackbird recording.

Using a longer framelength does seem to have a positive affect on the SOM output. Visual analysis of Fig. 6 shows that both of the building site recordings mostly produce outputs in the 12-20 class region. Figure 7 shows the class outputs for a blackbird recording and is included for comparison to the building site output. The class range for the blackbird recording is mostly 20-27, different ot that of the building site.

Converting the class output data for the building site into a histogram shows a distinct tendency for the class range stated (see Fig. 8). Similar SOM output histograms were achieved for the other urban sounds when analysed using a framelength of 0.5 seconds. The Cicada recordings produced results in line with those discussed in Section 3.1 showing that increasing the framelength to 0.5 seconds does not have an adverse affect on their classification.

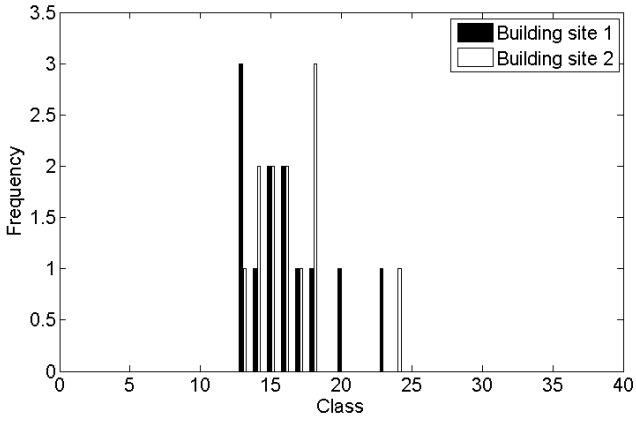

Fig. 8: Histogram of SOM class output for 2 building site recordings using a framelength of 0.5s for TDSC analysis.

3.3 Further Classification Work

The above findings are encouraging for the development of a system able to distinguish between various urban sounds (and species specific sounds). Further work in this area will initially focus on retrieval and analysis of more urban sound recordings. This will enable validation of the above results and to see if this approach shows promise for a broader range of audio data. SOM output class histograms for the new recordings will also be generated to discover any trends that may be present. It could also be possible to use

these histograms as an input to another classifier to see if a clear distinction can be found.

One direction for development of the classification system which is of particular interest would be to implement syntactic methods. Syntactic pattern recognition (SPR) involves breaking data into its basic building blocks, known as pattern primitives, and devising a grammar for a data set [12]. In the case of speech analysis (an area in which SPR is often used) the pattern primitives and grammar are fairly obvious. To use syntactic techniques for urban sound classification the pattern primitives will have to be devised based on the data available. From the plots given in Figures 4-6 there seem to be two options for pattern primitives; either the actual class outputs and how these follow each other, or applying trends to how the classes change with time. Once some pattern primitives have been decided upon a suitable classifier is then required to analyse these. A hidden Markov model (HMM) classifier could be the solution. HMMs are based on a state machine structure where the transition from the current state to the next has a probability associated with it [13]. A HMM will need training like any other classifier to determine the state transition probabilities. There are other possibilities for SPR classifiers but HMMs will be considered in the first instance because previous studies have shown they can be used for classification of everyday sonic environments [14].

The further work described above expands on the system structure shown in Fig. 3 by making the TDSC feature extractor and SOM classifier combination a preprocessor for further classification.

4 Conclusions

This paper has discussed the current work on signal classification for the ISRIE project. A system has been described consisting of a Time-Domain Signal Coding feature extractor and a Self-Organising Map classifier. A suitable TDSC codebook has been developed for use with urban audio signals and the current version of the codebook has improved significantly on the original. The effect of framelength on the SOM output classes has been investigated. From the results given a framelength of 0.5 seconds has shown the best results so far. Increasing the framelength further may have the effect of averaging the results too much and there will be no discernible difference between sources.

Suggestions for further work have been made which investigate the potential for using syntactic methods in the classification process and how the current implementation can be improved upon. The current TDSC/SOM combination will become a preprocessing unit for any expanded system that is developed.

Acknowledgments

The Instrument for Soundscape Recognition, Identification and Evaluation (ISRIE) project is funded by the Engineering and Physical Sciences Research Council (EP/E009581/1). The authors would like to thank the ISRIE project collaborators (Stuart Dyne and Christos Karatsovis, ISVR Southampton and Gui Yun Tian and Hidajat Atmoko, University of Newcastle) and other members of the Applied Bioacoustics group at York for their discussions and input relating to this work.

References

[1] C.Karatsovis, S. Dyne, "Instrument for soundscape recognition, identification and evaluation: an overview and potential use in legislative applications", *Proceeding of the Institute of Acoustics*, 602-608 (2008)

[2] S.H. Gage, R. Maher, G. Snachez, "EcoEARS – Application for Long-Term Monitoring and Assesment of Wildlife", In *Technical Symposium & Workshop: Threatened, Endangered and At-Risk Species on DoD and Adjacent Lands* (2005)

[3] R. Beale, T.O. Jackson, "Neural Computing: An Introduction", 1st ed. Reprint, Hilger (1998)

[4] N. Beltran, M. Duarte-Mermoud, M. Bustos, S. Salah, E. Loyola, A. Peña-Neira, J. Jalocha, "Feature extraction and classification of chilean wines", *Journal of Food Engineering* 75, 1-10 (2005)

[5] C. Dimoulas, G. Kalliris, G. Papanikolaou, V. Petridis, A. Kalampakas, "Bowel-sound pattern analysis using wavelets and neural networks with application to long-term, unsupervised, gastrointestinal motility monitoring", *Expert Systems with Applications* 34, 26-31 (2008)

[6] D. Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals", *Applied Acoustics* 62, 1359-1374 (2001)

[7] I. Farr, D. Chesmore, "Automated bioacoustic detection and identification of wood-boring insects for quarantine screening and insect ecology", *Proceedings of the Institute of Acoustics* 29, Pt. 3 (2007)

[8] C.H. Lee, C.H Chou, C.C. Han, R.Z. Huang, "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis", *Pattern Recognition Letters* 27, 93-101 (2006)

[9] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition", *Pattern Recognition Letters* 24, 2895-2907 (2003)

[10] B. Defréville, P. Roy, C, Rosin, F. Pachet, "Automatic recognition of urban sound sources", *Audio Engineering Society 120th Convention* (2006)

[11] F.M. Ham and I. Kostanic, "Principles of Neurocomputing for Science & Engineering", McGraw-Hill (2001)

[12] K.S. Fu, "Syntactic Methods for Pattern Recognition", Academic Press (1974)

[13] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE* 77, No. 2 (1989)

[14] L. Ma, B. Milner, D. Smith, "Acoustic Environment Classification", *ACM Transactions on Speech and Language Processing* 3, No. 2 (2006)