Edinburgh, Scotland
EURONOISE 2009
October 26-28

Instrument for soundscape recognition, identification and evaluation (ISRIE): technology and practical uses

Oliver Bunting
Jon Stammers
David Chesmore
University of York, YO10 5DD, UK

Omar Bouzid
Gui Yun Tian
University of Newcastle upon Tyne, NE1 7RU, UK

Christos Karatsovis
Stuart Dyne
ISVR Consulting, University of Southampton, SO17 1BJ, UK

ABSTRACT

Technological advancements in microelectronics and continuing research into signal characterisation and classification techniques have lead to promising results in developing an advanced sound meter. This instrument would be capable of characterising a sound field in terms of the relative contributions of the different noise sources. This paper provides an overview of this collaborative project, due for completion in October 2009, and the milestones that have been reached. In particular, the consideration and implementation of sensors and systems, the signal processing algorithms of source identification and classification, and the potential uses of the instrument in specific noise assessments in the UK are discussed.

1. INTRODUCTION

The collaborative work of three Universities; Newcastle upon Tyne, York and Southampton, has led to promising results in the development of an advanced sound meter that could provide a powerful measurement platform for many applications ranging from environmental noise assessments to the recording and evaluation of a variety of soundscapes.

Partners at the University of Newcastle upon Tyne have developed a multi-sensor technique for localising sound sources. In their particular method, the commercially available SoundField microphone probes have been used for 2D and 3D sound source localisation. Also, known beamforming techniques have briefly been investigated as an alternative technique for source localisation. Partners at the University of York have made use of a single SoundField microphone probe instead for developing a single-sensor technique for source localisation, separation and signal classification. Finally, partners at the University of Southampton have investigated the potential uses of ISRIE in existing noise legislation, planning and guidance and have also liaised with a wide range of stakeholders that could directly benefit from the use of such an advanced sound instrument.

2. ACOUSTIC SOURCE LOCALISATION

Over the course of the ISRIE project the co-authors at Newcastle University implemented an acoustic localisation system that is capable of locating a single sound source using at least three omni-directional microphones (i.e. 2D linear arrays) in a reverberant indoor environment with high accuracy for angle detection and small errors for distance estimation[1]. Sound source localisation in a 3D environment has been achieved by utilising the commercially available SoundField probes.

Figure 1 shows the use of three acoustic sensors in the context of a sound localisation system.

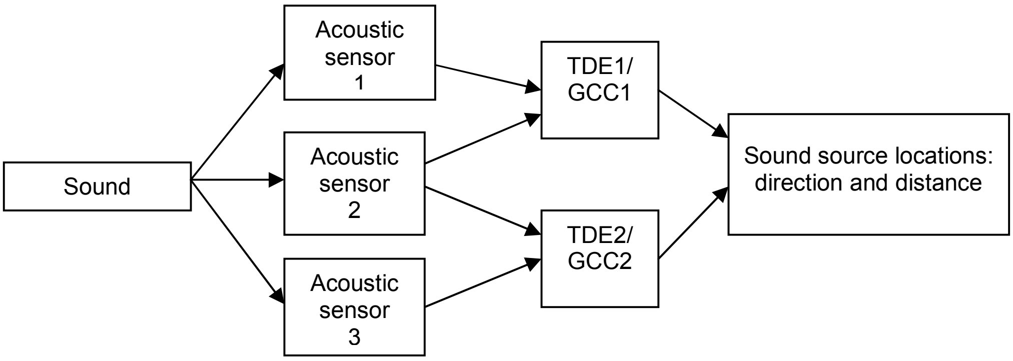

Figure 1: Three-microphone array system for acoustic monitoring[1].

The three acoustic sensors (omni-directional or 3D SoundField microphones) capture the sound simultaneously and the Time Delay Estimation (TDE) is extracted from any two sound signals from the three sensors using the Generalized Cross-Correlation (GCC). This method would ultimately derive sound source direction and distance through triangulation and geometric parameters. The three microphones are positioned in a straight line and the sides of the triangles formed by the source and each microphone represent the directional propagation paths from the source to each microphone. The direction of each propagation path is determined from the time differences between the signals arriving at the microphones. GCC is used to increase robustness to the adverse effects of early reflections and reverberation.

A. The 3 SoundField Microphone Method

Three SoundField SPS422B microphones were arranged in a straight line in order to achieve source localisation in a 3D environment[1]. Each microphone output is formed into a special signal format, the B-format, where four channels represent the velocity component in the three Cartesian directions; X (front-back), Y (left-right), Z (above-below) and one omni-directional signal, W, representing the pressure component. These signals are then fed into a PC for post-processing.

The Y and Z channel will generally be the same due the linear arrangement of the probes. The 2D configuration can be used for tilt and yaw estimation of sound direction in 3D. The X and W were therefore used for estimation in the experiment. With this arrangement, it has been possible to locate a single sound source in a reverberant indoor environment with an accuracy of 1° for angle detection and errors less than 4% for distance estimation. A rearrangement of the soundfield array in the Z Cartesian direction was tested in order to provide estimates of yaw instead of azimuth angles. The W and Z microphone outputs were used for the estimation and the results were similar. The SoundField probes could therefore potentially be used in a commercial source localisation system, where the sensitivity of these microphones to sounds arriving from different directions will be applied to source localisation in planes other than that defined by the line of the array.

B. Beamforming Techniques

In the literature, beamforming is another suggested technique that has extensively been used in developing instruments for soundscape recognition, identification and sound source localisation[2, 3]. The beamforming technique is a technique that searches for a peak (or peaks) by achieving a full directional scan in order to determine the source(s) direction(s) from this (or these) peak(s). This can be achieved by delaying and summing the acoustic emitted signals to minimise the noise effects and enhancing (or maximising) the amplitude of the point (or direction) that represents the location of the sound source[2, 3]. The sound source can be considered to be in the near-field if the wavefront is modelled as spherical, whereas it is considered to be in the far-field if it is assumed to be planar[3]. The consequences of these assumptions are that in the near-field both the range and Direction of Arrival (DOA) can be computed, whereas in the far-field, only the DOA can be estimated due to computational costs[3]. Li[3] designed a flexible broad-band beamformer using nested Concentric Ring Array (CRA) that can be divided into sub arrays, where each sub array can cover a specified operating range. In our study, the acoustic camera, which mainly includes a microphone array of Star 36 sensors[4], a data-reader device, a laptop computer and the "NoiseImage" software[4], has been used for the investigation on flexible beamforming techniques and instrument validation. The data from this study is currently under investigation.

3. SOURCE SEPARATION

The task of automated recognition of audio signals is made considerably more complex by multiple sources being present in the audio recording, with a consequent reduction in recognition accuracy rates. To provide enhanced recognition accuracy, ISRIE employs a source separation algorithm prior to the recognition stages. The separation method developed for ISRIE is based on the assumption of W-disjoint orthogonality. That is, audio sources are sparse in a time-frequency domain. The sensor used is a Soundfield ST350, a B-format coincident microphone array[5, 6] that offers a more portable microphone system over the SPS422B.

A. Model

Consider a 3-dimensional coincident array comprising of 3 orthogonal sets of figure-of-eight microphones and an omni-directional microphone at the centre of the array. Given the location of the sources, the B-format mixture of signals in the anechoic case can be expressed as:

$$\begin{vmatrix} w(t) \\ x(t) \\ y(t) \\ z(t) \end{vmatrix} = \begin{vmatrix} 1/\sqrt{2} & \dots & 1/\sqrt{2} \\ \cos(\theta_1)\cos(\lambda_1) & \dots & \cos(\theta_N)\cos(\lambda_N) \\ \sin(\theta_1)\cos(\lambda_1) & \dots & \sin(\theta_N)\cos(\lambda_N) \\ \sin(\lambda_1) & \dots & \sin(\lambda_N) \end{vmatrix} \begin{vmatrix} s_1(t) \\ . \\ . \\ . \\ s_N(t) \end{vmatrix} \qquad (1)$$

where x, y, z are the mixtures observed on the Cartesian axis, w is the mixture observed by the omni-directional sensor, and θ, λ are the azimuth and elevation for the direction of arrival of a particular source.

B. Assumptions

Separation of a given mixture is subject to two conditions on the source mixture being met. These are W-disjoint orthogonality[7] and radial sparsity. These are described formally below.

W-disjoint Orthogonality

Two sources s_i and s_j are W-disjoint orthogonal if the following condition is met.

$$S_i(\omega,\tau)S_j(\omega,\tau) = 0 \qquad \forall\ i \neq j, \omega, \tau \qquad (2)$$

where $S(\omega,\tau)$ represents the time-frequency domain transformation of $s(t)$.

Radial Sparsity

This a condition placed on the geographical location of the sources. Each source must have a unique direction of arrival at the sensor.

$$(\theta_i, \lambda_i) \neq (\theta_j, \lambda_j) \qquad \forall\ i \neq j \qquad (3)$$

C. Direction of Arrival (DOA) Calculation

Provided the above conditions have been met, the DOA of the B-format audio signal can be calculated in the time-frequency domain using a method from Directional Audio Coding Scheme (DirAC)[8, 9].

$$\vec{D}(\omega,\tau) = -\Re\left(W^*(\omega,\tau) * \left(\vec{e}_x X(\omega,\tau) + \vec{e}_y Y(\omega,\tau) + \vec{e}_z Z(\omega,\tau)\right)\right) \forall\ \omega,\tau \qquad (4)$$

where $\vec{e}_x$, $\vec{e}_y$ and $\vec{e}_z$ are unit vectors along the Cartesian axes.

D. Source Location Estimation

Using the calculated DOA vectors, it is possible to perform source localisation using a variety of techniques. Perhaps the simplest is to construct a histogram over an arbitrary time period, and look for peaks. This method, along with another clustering method based on self-learning neural networks, has been looked at to perform this task.

E. Demixing

For each source location, which is denoted E_i, M_i describes a bit mask in the time-frequency domain for each source.

$$M_i(\omega,\tau) = \begin{cases} 1 & \Big| \quad \arccos\left(\frac{\vec{E}_i}{\left|\vec{E}_i\right|} \cdot \frac{\vec{D}}{\left|\vec{D}\right|} \right) \le \delta \\ 0 & \Big| \quad otherwise \end{cases} \qquad \forall i \tag{5}$$

where δ provides a user defined angular margin around the source location.

The sources can then be recovered by using the mask to filter W in the time-frequency domain.

$$\hat{S}_i = M_i(\omega,\tau) * W(\omega,\tau) \tag{6}$$

from which $\hat{s}_i$ can be gained by performing an inverse time frequency transformation.

F. Results

Table 1 shows the results from a signal separation experiment.

Table 1: Results from a signal separation experiment.

Speaker	Performance Measure				Location	
	Signal-to-Interference Ratio (SIR) in mixture	SIR after masking	SIR gain	Preserved Signal Ratio (PSR) after masking	azimuth	elevation
1	-0.17 dB	12.14 dB	12.32 dB	12.32 dB	120	0
2	-2.96 dB	12.30 dB	15.27 dB	15.27 dB	280	10
3	-6.81 dB	10.89 dB	17.70 dB	17.70 dB	340	20

The separation algorithm was tested on a mixture of three male speakers reading passages from a novel. Each speaker was recorded independently under anechoic conditions, and the mixture created by the summation of the three B-format recordings. The recordings were performed in this manner to allow analytical comparison of the separated speakers with the original recording. Speakers one and two show much higher Preserved Signal Ratio (PSR) results compared to speaker three. This is perhaps unsurprising, considering that speaker three has an initial Signal-to-Interference Ratio (SIR) of −6.81 dB. All the speakers are intelligible on listening, although there is an appreciable level of crackling on speaker three. The SIR gain for all speakers shows excellent results, showing high suppression of the interfering speakers, with an average improvement in SIR of 15 dB. These results compare well to those listed for mixtures of two speakers[10].

As far as the validity of the assumptions, W-disjoint orthogonality has been shown to be a valid assumption for speech signals. Acoustic niche theory also suggests an evolutionary pressure for this to be the case in the animal kingdom. However, the authors concede that in the general case, the assumptions are not guaranteed to hold true. Further investigations into the applicability of these assumptions to a range of situations need to be performed.

4. SIGNAL CLASSIFICATION

ISRIE will also perform the classification of the separated audio signals which are provided by the signal separation as discussed previously. The output of the classification algorithms will advise the user of ISRIE which category of sounds a particular signal belongs to. It is assumed that the input signal to the classification system contains only one sound source.

A. Sound Categories

A taxonomy of sound categories has been devised specifically for the purpose of ISRIE. Figure 2 illustrates these categories.

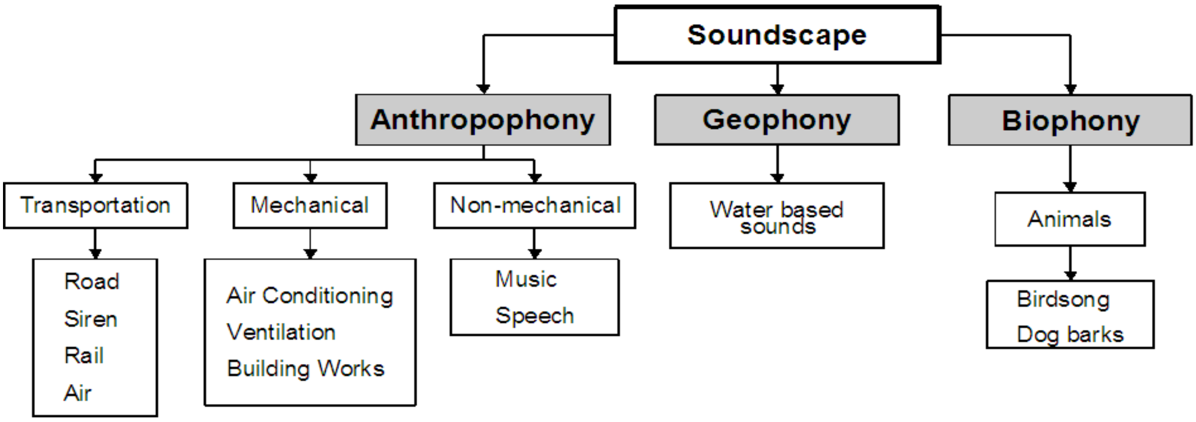

Figure 2: Urban soundscape categories.

Initially, the soundscape is split into three main categories. *Anthropophony* relates to sounds made or caused by human activity, *biophony* sounds are those made by animals, and *geophony* encompasses sounds not caused by either of the above.

B. Classification using Time-Domain Signal Coding

A typical classification system consists of two components: a feature extractor and a classifier[11]. There is sometimes a third component to provide some pre- or post-processing either at the input or output to the system. The data that is to be classified will be passed into the feature extractor whose role it is to reduce the complexity of the data before it reaches the classifier[11] thus optimising the classification process. A good overview of a selection of these techniques can be found in the comparison made by Cowling and Sitte[12].

The feature extraction method that has been used for data reduction in ISRIE is known as Time-Domain Signal Coding (TDSC). This is a purely time-domain analysis method which has previously shown to be successful in the identification of wood-boring insects[13] and in the classification of different Orthoptera[14]. The data produced by the TDSC algorithm describes a waveform by the number of samples (duration - D) and number of minima (shape – S) contained within each epoch (signal between 2 consecutive zero crossings) of the waveform. The D-S information is stored for a given frame of the waveform by means of a codebook. After a signal has been analysed using TDSC, each code within the codebook will have a number of occurrences associated with it to describe its D-S characteristics. It is this frequency information, the S-matrix, which is then used for classification. A more detailed explanation of how TDSC was developed and the other features it can extract from the full bandwidth signal is given by Chesmore[14]. Figure 3 shows how the TDSC analysis fits into the classification system.

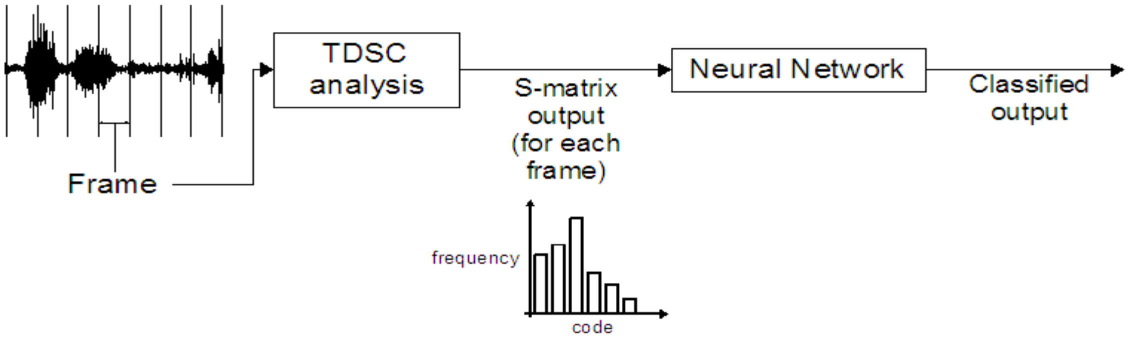

Figure 3: Proposed classification system. The S-matrices for each frame of the waveform are classified individually.

It was decided that a neural network approach in the classification would be adopted. Initially, an unsupervised Self-Organising Map (SOM) network was used but this struggled to differentiate between the test pieces of audio data. Significant improvements in classification were gained by introducing supervised learning into the system. A Learning Vector Quantisation (LVQ) network was implemented using the LVQ1 learning rule[15, 16]. Eight different categories of sounds were placed into 4 groups: group 1 contained air traffic, air conditioning and ventilation units, and building works; group 2 contained road and rail traffic; group 3 contained birdsong and also recordings of crickets; and group 4 contained some speech examples. The grouping of the sounds was chosen based on how consistent the signal was throughout the duration of the recording. After training was completed using a training set of 40 recordings, the network was tested using a 30-second test audio file which combined audio from each of the 4 groups. Network accuracy for each of the individual groups was poor for all but group 1 (88%). However, when combined results were observed for how well the system could recognise non-bioacoustic audio (groups 1 and 2), the accuracy rose to 93%. This shows that it is possible to perform an initial classification using the relatively simple methods discussed above. Work is now focused on developing the system further to incorporate classifiers to differentiate between the various bioacoustic and non-bioacoustic signals. Feed-forward neural networks with backpropagation training are being experimented with and are showing positive initial results.

5. APPLICATIONS

The uses of ISRIE could range from assisting acoustic consultants and planners in making the right decision on the most appropriate control measures in a project where noise concerns may arise, through to assisting soundscape artists and sound engineers with the recording of isolated sound events for either artistic reasons or for the subjective evaluation of different soundscapes. The usefulness of ISRIE in environmental noise impact assessments, such as PPG 24[17], BS 4142[18] and noise nuisance applications have previously been discussed[19]. Over the course of this research project, different stakeholders have also been interviewed in order to assess what measurement parameters would be required from such an instrument to log and what would be the additional benefits from the use of such an instrument.

A. BS 4142

In BS 4142 assessments, ISRIE could potentially be used to obtain the specific noise level L_{Aeq} of a source and the background noise level L_{A90} without requiring the need to measure these descriptors separately. The instrument would offer individual logged values of these two environmental noise level descriptors in order to establish the arithmetic difference between the intruding mechanical noise level and the typical background noise level without the presence of any mechanical plant or industrial noise. Also, in practice, there are instances where it is not possible to obtain separate measurements of these two descriptors, because either the mechanical source cannot be turned off in order to measure the background noise level, or the mechanical noise cannot accurately be quantified at the receptor's location due to interference from other sources, such as transportation related noise. ISRIE would be capable of deriving these parameters through its discrimination and classification algorithms as discussed above.

B. PPG 24

In PPG 24 assessments, the existing environmental noise levels are established over a 24-hour measurement period, when planning a new housing development. The measurements are normally unmanned for economic reasons since they cover such an extensive measurement period. Firstly, it is apparent that in mixed soundscapes, where for example there is almost an equal contribution of railway and road traffic noise, it is difficult to quantify the contributing noise sources, or even determine which is the dominant noise source. Therefore, it is not always feasible to establish the most representative noise source category in which the noise environment should be assessed in. ISRIE would be useful in obtaining these individual contributions in L_{Aeq} terms in order to decide which is the prominent noise source in that specific environment. Secondly, ISRIE would automatically log and classify individual events that exceed a certain criterion, such as 82 dB $L_{A,max,S}$ and assess whether these transient events are intrusive sources of noise, e.g. mechanical, or non-intrusive, e.g. birdsong or sounds from other animal life. This type of automated assessment is not possible with the use of current technology since the noise survey is normally unmanned and these individual transient events can only be evaluated and assessed at the post-processing stage.

C. Noise Nuisance

Environmental Health Officers (EHOs) of Local Authorities in the UK would make use of an advanced sound instrument for various reasons. Firstly, ISRIE would enable them to investigate complex noise complaints in the case where it is not clear which mechanical plant noise source affects the complainant's house in a highly built-up area. Secondly, the problem of low frequency noise, potentially originating from tunneling or drilling works, can be an issue for some residents in a community. These noise complaints can be difficult to assess with the current technology of sound level meters and ISRIE's characterisation capability would work well in these types of problem where the source is of tonal character. Thirdly, ISRIE would aid in monitoring the noise from music events and assist EHOs in reaching decisions upon the licensing of commercial premises that may give rise to noise complaints.

D. Other Engineering Consultancy Problems

The use of a conventional sound level might not be adequate in some cases since there can be interference from other noisy equipment when trying to quantify a particular noise source in an industrial area. There are also instances, where the noise of certain installations, such as electrical transformers, cannot easily be quantified because either these installations are near sources of transportation noise, e.g. motorways, or because there are other electro/mechanical installations nearby that may contribute to the overall measured level. Also, as part of the Land Compensation Act, difficulties can arise when trying to establish only the road traffic components at houses that are situated miles away from a newly constructed or modified road. ISRIE would be capable of solely measuring the traffic noise components from the remaining background noise, something that is not possible with the current sound level meters. Similar measurement problems can arise when trying to quantify noise solely emanating from racing tracks that might affect nearby communities.

E. Soundscape Recordings

Recordings of soundscapes is developing in many applications ranging from creating archived sound recordings of a variety of animal sounds through to the recordings of any other types of soundscape for recreating experiences in art installations, museums and galleries. The need for carrying out recordings of sounds in isolation is important in many applications. At the moment, in order to separate different sounds, noise suppression techniques are used in order to filter out the remaining sound, or the recording is delayed until the level of the intrusive noise has dropped to such a level that it is not significantly contributing to the overall level. ISRIE would be useful in recording these sounds as isolated events and hence providing a reference instrument for sound recording.

F. Future Policy

ISRIE could enable planners to consider the balance between ‘positive’, e.g. natural sounds and ‘negative’ sounds, e.g. mechanical-like sounds in a mixed sound environment as part of a regeneration plan for improving the quality of life in urban agglomerations or assist in the design of new spaces of personal enjoyment and recreation in metropolitan cities. The first step would be to establish which types of sound are considered ‘wanted’ and ‘unwanted’ in that environment. Then, ISRIE would be used as an instrument to establish the current percentage of wanted and unwanted sounds through its source discrimination and classification algorithms as presented above. Finally, the management of these sounds would involve standard noise abatement techniques along with the potential introduction of more wanted sounds. In the end, ISRIE could be used to assess whether the desired ‘mix’ of wanted and unwanted sounds was achieved.

5. CONCLUSIONS

The need of a network sensor system with the development of algorithms and techniques for automatically characterising sounds in a complex sound environment is more evident than ever before. This paper has presented a number of suggested measurement platforms for the measurement of sounds along with promising techniques for signal separation and classification. The use of ISRIE could ultimately revolutionalise the way we currently perceive soundscapes and could affect the way we measure, assess and record sounds in the future.

ACKNOWLEDGMENTS

University of Southampton, York and Newcastle gratefully acknowledge the Engineering and Physical Sciences Research Council (EPSRC) for sponsoring this research work.

REFERENCES

1. H. Atmoko, T. Gui Yun and B. Fazenda, "Accurate sound source localization in a reverberant environment using multiple acoustic sensors", Meas. Sci. Technol. Feb. 2008, pp. 1-10.
2. Terence Betlehem, "Acoustic Signal Processing Algorithms for Reverberant Environments", PhD Thesis, Department of Information Engineering, School of Information Sciences and Engineering, Australian National University, Nov. 2005.
3. Yunhong Li, "Broadband Beamforming and Direction Finding Using Concentric Ring Array", PhD Thesis, the Faculty of the Graduate School, University of Missouri-Columbia, Jul. 2005.
4. Acoustic sound source localisation: Download: Acoustic Camera: Applications and System Overview (PDF), Available at: http://www.acoustic-camera.com/pdfs/ac_brochure2009.pdf, Accessed: May 2009.
5. Michael Gerzon. Periphony: With-height sound reproduction. Journal Audio Eng. Soc., 21(1), pp2–10, 1973.
6. Michael Gerzon. The design of precisely coincident microphone arrays for stereo and surround sound. In Proc. 50th Convention of the Audio Eng. Soc., 1975.
7. S. Rickard and Z. Yilmaz. On the approximate w-disjoint orthogonality of speech. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '02), 1, pp 529–532, 2002.
8. J. Merinaa and V. Pulkki. Spatial impulse response rendering. In Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFx'04), pp 139–144, October 2004.
9. Ville Pulkki. Spatial sound reproduction with directional audio coding (DirAC). Journal Audio Eng. Soc., 55(6), June 2007.
10. O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. IEEE Journal. Sig. Proc., 52(7), pp1830–1847, 2004.
11. R. Beale and T.O. Jackson, Neural Computing: An Introduction, Hilger 1998.
12. M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition", *Pattern Recognition Letters* 24, pp. 2895-2907 (2003).
13. I. Farr and E. D. Chesmore, "Automated bioacoustic detection and identification of wood-boring insects for quarantine screening and insect ecology", in *Proceedings of the Institute of Acoustics* 29, Pt. 3, pp. 201-208 (2007).
14. E.D. Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals", *Applied Acoustics* 62, pp. 1359-1374 (2001).
15. T. Kohonen, "Improved Versions of Learning Vector Quantization", *International Joint Conference on Neural Networks* 1, pp. 545-550 (1990).
16. H. Demuth and M. Beale, Neural Network Toolbox User's Guide, The MathWorks, Inc. 2001.
17. Planning Policy Guidance 24: Planning and noise, Department of the Environment, 1994.
18. BS 4142: 1997: Method for rating industrial noise affecting mixed residential and industrial areas, BSI.
19. C. Karatsovis and S J C Dyne, "Instrument for soundscape recognition, identification and evaluation: an overview and potential use in legislative applications", in *Proceedings of the Institute of Acoustics,* 2008, Vol. 30, Pt.2.