Noise Robust Audio Surveillance using Reduced Spectrogram Image Feature and One-Against-All SVM

Roneel V Sharan* and Tom J Moir

School of Engineering, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand

Email: roneel.sharan@aut.ac.nz, tom.moir@aut.ac.nz

*Corresponding Author

Abstract

This paper builds on the technique of feature extraction from the spectrogram image of sound signals for automatic sound recognition. The spectrogram image is divided into blocks and statistical distributions are extracted from each block as features. However, when compared to related work, we reduce the dimensionality of the feature vector using mean and standard deviation values along the row and column of the blocks without compromising the classification accuracy. We demonstrate the technique in an audio surveillance application and evaluate the performance using four common multiclass support vector machine (SVM) classification techniques, one-against-all, one-against-one, decision directed acyclic graph, and adaptive directed acyclic graph. Experimentation was carried out using an audio database with 10 sound classes, each containing multiple subclasses with intraclass diversity and interclass similarity in terms of signal properties. Under noisy conditions, the proposed reduced spectrogram image feature (RSIF) produced significantly better classification accuracy than the conventional log compressed mel-frequency cepstral coefficients (MFCCs) and marginally better classification accuracy than linear MFCCs, which does not utilize any compression. The linear spectrogram image representations for feature extraction and the one-against-all multiclass SVM classification method were found to be the most noise robust. In addition, significantly improved results were obtained under noisy conditions when the RSIF is combined with linear MFCCs.

Keywords: audio surveillance, noise robust, sound recognition, reduced spectrogram image feature, support vector machines

1.0 Introduction

Unlike speech recognition, which has been a highly researched area over the past few decades, research in sound recognition, a closely related area, is relatively new. Sound recognition can cover a wide range of applications. Some of these include content-based audio classification such as for application in multimedia [1] or more specifically in music genre classification [2] and musical instrument sound classification [3], hearing aid [4], environmental sound recognition [5], audio surveillance [6], and respiratory sound classification [7, 8].

While most initial work in sound recognition concentrated on content-based audio classification for multimedia applications, sound event recognition has also gained attention in recent years. Surveillance and security systems are common applications for such work such as security monitoring in a room [9] and medical telemonitoring [10]. While these are examples of standalone audio surveillance systems, which is the aim of this work, audio and video surveillance systems could also be integrated for a more holistic approach to the development of surveillance systems. Video surveillance systems have been around for many years but they have limitations such as limited field of view, challenging external conditions, and relatively expensive computation and data storage. An automatic sound recognition

(ASR) system could be used to complement a video-based surveillance system such as in public transports [11] and surveillance in banks [12].

Sound recognition is essentially a pattern recognition problem and most of the techniques involved are inspired from speech recognition. The two key components underlying the robustness of an ASR system are feature selection and choice of classifier. Mel-frequency cepstral coefficients (MFCCs) have been used a baseline feature in many sound recognition systems and are often complemented with other features for improved performance. However, conventional MFCCs, which applies log compression to the filter bank energies before computing the cepstrum coefficients, has been shown to give poor performance in the presence of noise [13-15]. In a recent work, though, statistical moments derived as features from the spectrogram image of sound signals, referred as the spectrogram image feature (SIF), was shown to give relatively good results under noisy conditions in sound event recognition [14].

In this work, which is a continuation of our earlier work in [16, 17], we use features derived from the spectrogram image of sound signals for classification of sounds in an audio surveillance application. In addition, we use this feature in combination with linear MFCCs, a special case of root compressed MFCCs which has shown to be more noise robust than the conventional MFCCs in speech recognition [15]. We test the robustness of the proposed feature set at different noise levels and different noise environments using support vector machines (SVMs) for classification.

The remaining of this paper is organized as follows. Section 2 covers related work in sound recognition and, in particular, audio surveillance. Section 3 is on feature extraction and feature vector formation. In section 4, we give an overview of SVMs and the multiclass classification techniques for SVMs. Experimental results and discussions are given in section 5 followed by conclusion and recommendations in section 6.

2.0 Related Work

One of the early work in the area of content-based audio classification and retrieval, which also found commercial success, was called Muscle Fish [1] (www.muslcefish.com). It used nearest neighbor (NN) method of classification using low-level features such as loudness, pitch, brightness, and bandwidth. This research was extended in [18] where the nearest feature line (NFL) method of classification and the introduction of MFCCs as features showed superiority when compared to [1]. Two other similar work but using SVMs for classification can be found in [19, 20].

Environmental sound recognition is another area of sound recognition. It poses a greater challenge when compared to most other sound recognition applications since an environmental sound can comprise of a number of different sounds within the environment which can be present in different combinations at any given time. An example of environmental sound recognition can be found in [5] where fourteen environment types are considered. A combination of MFCCs and matching pursuit (MP) [21] features gives the highest accuracy at 83.9% using Gaussian mixture model (GMM) for classification.

The techniques in audio surveillance systems are similar to content-based audio classification and environmental sound recognition. In [22], a *scream* and *gunshot* detection and localizing system is presented using MFCCs and other temporal, spectral, spectral distribution, and correlation-based features. In [23], features based on pitch range (PR) and MFCCs are proposed. The audio database has four abnormal events: *glass breaking*, *dog barking*, *scream*, and *gunshot*; and three normal events: *engine noise*, *rain*, and *restaurant noise*. SVM, radial basis function (RBF) neural networks, and NN classifiers are experimented with but best results are achieved using SVM for classification and the combined feature set. SVMs are also used for classification of seven audio events: *screaming*, *crying*, *speech (male)*, *speech (female)*, *laughing*, *knocking*, and *explosion* in [24]. Both time and frequency domain features were studied but only MFCCs and its derivatives were found to be more useful.

One of the more comprehensive piece of work in audio surveillance is given in [6] where large feature dimensions were considered for classification using one-class SVM (1-SVM) [25]. The features considered in this work are divided into time-domain features: zero-crossing rate (ZCR) and short-time energy (STE); frequency-domain features: spectral centroid (SC) and spectral roll-off (SR); linear prediction, perceptual linear prediction, and cepstral features: linear prediction cepstral coefficients (LPCCs), perceptual linear prediction (PLP), and MFCCs; and wavelet-based features, derived from wavelet coefficients. The sound database consists of 1015 sound files from 9 classes: *human screams*, *gunshots*, *glass breaking*, *explosions*, *door slams*, *dog barks*, *phone rings*, *children voices*, and *machines*. The highest classification accuracy achieved is 96.89% under clean conditions and 93.33%, 89.22%, 82.80%, and 72.89% with the best performing feature set at 20dB, 10dB, 5dB, and 0dB signal-to-noise ratio (SNR), respectively, with 70% of clean data used for training and the remaining for testing.

In [14], a slightly different approach is taken where features extracted from spectrogram image of sound signals are used for sound event recognition. The spectrogram images are partitioned into blocks and second and third central moments are computed in each block as features. For experimentation, 60 sound categories are used to give a selection of collision, action, and characteristics sounds. Each class has 80 files of which 50 files are randomly selected for training and 30 files for testing. Four noise types: *speech babble*, *destroyer room control*, *factory floor 1*, and *jet cockpit 1* from NOISEX-92 [26] database are added at 20dB, 10dB, and 0dB SNR to test the robustness of the proposed method. The best results for training with clean signals were between 74-77% at 0dB SNR for the four noise types.

However, literature using this unique approach is limited, especially with an application in the presence of noise. Spectrogram derived features were used in a hearing aid application in [4]. While more than thirty features were extracted, eleven features were chosen through correlation analysis for classifying four classes: *speech*, *speech in noise*, *noise*, and *classical music*. The original image is in grayscale but binary images are also created for feature extraction. In [27], features derived from spectrogram image texture analysis are used in music genre recognition. For environmental sound recognition in [28], spectrogram derived features were shown to give higher results than MFCCs, linear prediction coefficients (LPC), and MP. In addition, log-Gabor filtered spectrogram images are used for feature extraction in [29].

In this work, we largely follow the SIF technique proposed in [14] for automatic sound recognition in an audio surveillance application. However, we propose a method to reduce the SIF dimension using the mean and standard deviation of the extracted features along the rows and columns of the blocks. We refer this as the reduced spectrogram image feature (RSIF). In addition, concatenating two or more set of features for improved classification accuracy is a common practice in ASR systems. Conventional MFCCs have been shown to produce good results under clean conditions and are often combined with other features for improved performance, such as in [5, 6, 23]. However, the log compression used in conventional MFCCs has been shown to reduce its performance in the presence of noise. As such, root compressed MFCCs are proposed in [15]. We propose to combine the RSIF with linear MFCCs, which uses the upper limit of the root value, to potentially achieve even greater classification accuracy than the individual features.

The approach taken towards the problem of audio surveillance is similar to [6] where a sound class has a number of different sound events. For example, shots fired from a rifle, shotgun, and machine gun are examples of different sound events, which would be treated as three different sound classes as per the approach in [14], but are taken as a single sound class, such as gunshots, in [6]. In some cases, the signal properties of subclasses in one particular class are similar to the subclasses in other classes and different to subclasses in its own class. This creates interclass similarity and intraclass diversity, increasing the complexity of the problem as a result.

A total of 10 classes are selected to show the robustness of the proposed method. This is more than most other work in the area of audio surveillance such as seven classes in [12, 23, 24], and nine classes in [6]. It can generally be said that the classification accuracy decreases with an increase in the number of classes as summarized in [5] in relation to the problem of environmental sound recognition.

As far as the choice of the classifier is concerned, in most sound recognition systems, such as [6, 19, 20, 23, 24], SVM has been preferred. In [20], the performance of SVM was found to be better than KNN which was in turn shown to give better results than GMM. Some other literature where the performance of SVM has been compared against other classifiers but with similar conclusions can be found in [23, 30, 31].

SVM is a relatively new classifier which has especially been found to give good results when using low training data. Initially intended as a binary classifier, a number of methods have since been developed to use SVMs for multiclass classification. The most common technique in solving the multiclass problem is to reduce it into multiple binary classification problems. Four of the widely used methods based on this approach are: one-against-all (OAA), one-against-one (OAO), decision directed acyclic graph (DDAG), and adaptive directed acyclic graph (ADAG).

The performance of the multiclass SVM classification methods have been compared in a number of literature such as in [32, 33]. The difference in the classification accuracy in most cases is minimal and, as such, the preference of one technique over the others is largely based on faster training and evaluation times. However, most such analysis are limited to clean conditions and it is unclear which approach is more suitable for classification under noisy conditions. In this work, we compare the performance of OAA, OAO, DDAG, and ADAG multiclass SVM classification methods under different noise conditions and SNR. We evaluate the performance of each method using its classification accuracy and also compare the training and evaluation times.

3.0 Feature Extraction

3.1 Reduced Spectrogram Image Feature

3.1.1 Grayscale Spectrogram

To generate the grayscale spectrogram, firstly, the DFT is applied to the windowed signal as

$$X(k,t) = \sum_{n=0}^{N-1} x(n)w(n)e^{\frac{-2\pi ikn}{N}}, \qquad k = 0,\ldots,N-1 \tag{1}$$

where N is the window length, $x(n)$ is the time-domain signal, $X(k,t)$ is the k^{th} harmonic corresponding to the frequency $f(k) = kF_s/N$ for the t^{th} frame, F_s is the sampling frequency, and $w(n)$ is the window function.

The linear and log values are then obtained as

$$S_{Linear}(k,t) = |X(k,t)| \tag{2}$$

$$S_{Log}(k,t) = \log|X(k,t)|. \tag{3}$$

These values are normalized in the range [0,1] which gives the grayscale image intensity values. The normalization is given as

$$G(k,t) = \frac{S(k,t) - \min(S)}{\max(S) - \min(S)}. \tag{4}$$

Feature extraction using cepstrogram images wasn't considered in this work since this approach was shown to give relatively poor results in [14].

3.1.2 Color Mapping

The linear and log grayscale intensity values are then quantized and mapped onto the red, green, and blue (RGB) monochrome components which is a generalization of the pseudo-color mapping procedure as mentioned in [14]. There are many color spaces which can be used for this purpose and in this work one of the common color spaces, the HSV color space, was used. The mapping of the grayscale image to the monochrome image can be given as

$$Q_c(k,t) = f_c(G(k,t)) \qquad \forall c \in (c_1, c_2, \ldots c_N) \tag{5}$$

where Q_c is a monochrome image (R, G, or B), c is the quantization regions, and f is a nonlinear mapping function.

The linear and log spectrogram images for a sound signal from *construction* sound class are given in Figure 1.

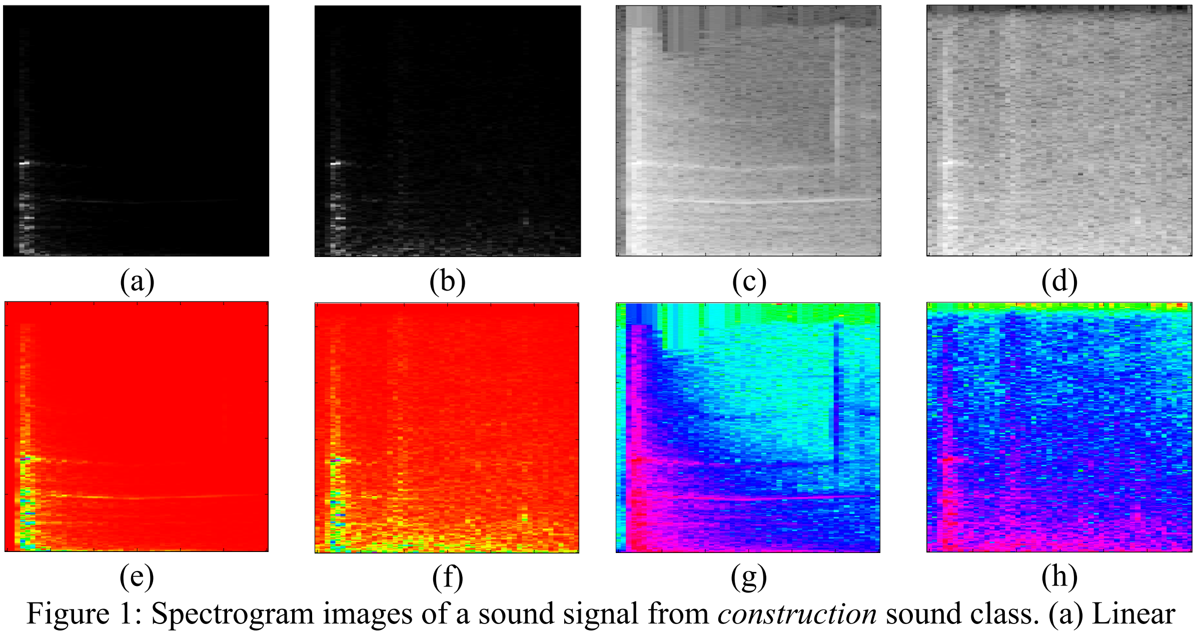

Figure 1: Spectrogram images of a sound signal from *construction* sound class. (a) Linear grayscale image under clean conditions, (b) linear grayscale image at 0dB SNR with factory noise, (c) log grayscale image under clean conditions, (d) log grayscale image at 0dB SNR with factory noise, (e) linear quantized image under clean conditions, (f) linear quantized image at 0dB SNR with factory noise, (g) log quantized image under clean conditions, and (h) log quantized image at 0dB SNR with factory noise

3.1.3 Feature Extraction and Representation

The spectrogram image is essentially a matrix of data formed by stacking the spectrum values from each frame side-by-side as depicted in Figure 2 for a grayscale image. Each image is then divided into blocks and central moments are computed as features. The v^{th} central moment for any given block of image can be determined as

$$m_v = \frac{1}{N_s} \sum_{s=1}^{N_s} \left(I_s - \mu \right)^v \tag{6}$$

where N_s is the sample size or the number of pixels in the block, I_s is the intensity value of the s^{th} sample in the block, and μ is the mean intensity value of the block.

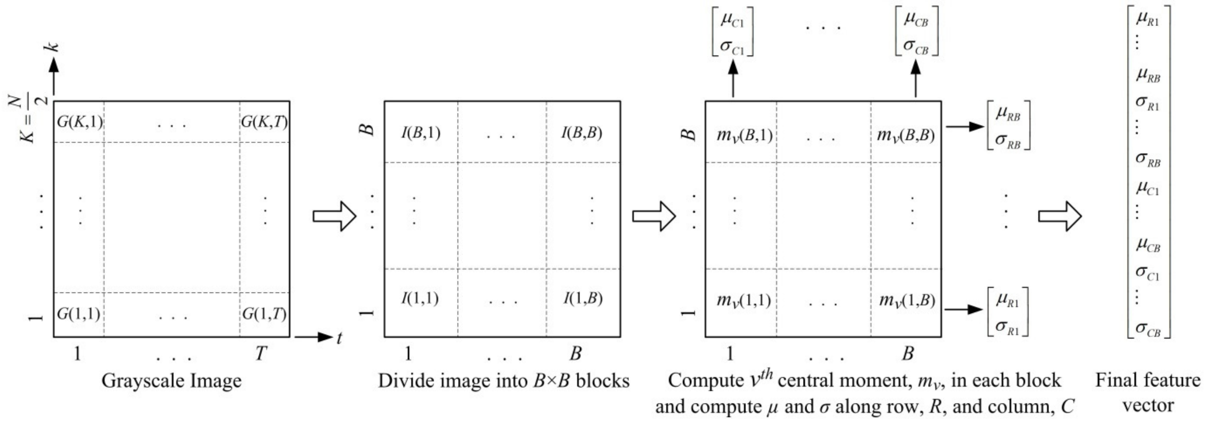

Figure 2: Proposed RSIF data representation. Note that $I(b,b)$ is a matrix of image intensity values for the block in the b^{th} row and b^{th} column, $m_v(b,b)$ is the v^{th} central moment for the block in the b^{th} row and b^{th} column, and μ_{Rb}, σ_{Rb} and μ_{Cb}, σ_{Cb} are the mean and standard deviation of the extracted feature for the blocks in the b^{th} row and b^{th} column, respectively, $b = 1, 2, \ldots, B$.

An important consideration after feature extraction is feature vector representation. We considered two approaches for feature data representation. The first approach is same as in [14] where raw feature data from all the blocks are concatenated to form the final feature vector for each signal. If the number of blocks along the row and column of the spectrogram image is same and is given as B, the dimension of the final feature vector using this approach is B^2. While this gives a reasonable size feature dimension for small values of B, the feature vector dimension can become extremely large as the number of blocks increases. In [14], the images are divided into 9×9 blocks. For the grayscale spectrograms, the final feature dimension with two features, second and third central moments, computed in each block is $9 \times 9 \times 2 = 162$. For the quantized images, with three quantization regions, this increases to $9 \times 9 \times 2 \times 3 = 486$.

The feature data representation method that we propose is to concatenate the mean and standard deviation of the central moment values along the row and column of the image blocks as depicted in Figure 2. This gives a feature vector dimension of $B \times 4$. While this approach gives a higher feature dimension than the approach in [14] for $B < 4$, it gives a lower feature dimension for $B > 4$. Using the case of 9×9 blocks once again, the feature dimension is $9 \times 4 \times 2 = 72$ for the grayscale spectrogram and $9 \times 4 \times 2 \times 3 = 216$ for the quantized spectrogram, which is 2.25 times smaller than in [14]. However, the preference of one feature data representation method over another is largely dependent on the classification accuracy which is discussed in section 5.5.

3.2 MFCCs

MFCCs are computed using the DFT power coefficients. The power coefficients are firstly filtered using a triangular filter bank. In conventional MFCCs, the filter bank energies are log compressed before applying discrete cosine transform (DCT). The root compressed MFCCs are computed in a similar manner but root compression is applied to the filter bank energies instead of log compression. Root compressed MFCCs can be determined as [15]

$$MFCC_d = \sqrt{\frac{2}{F}} \sum_{f=1}^{F} E_f^{\gamma} \cos\left(\frac{\pi d}{F}(f-0.5)\right), \qquad d = 1,2,\ldots,L \tag{7}$$

where L is the order of the cepstrum, E_f is the output of the f^{th} filter bank, $f = 1, 2, \ldots, F$, and γ is the root value used to compress the filter bank energies, $0 < \gamma \leq 1$. When $\gamma = 1$, the filter bank energies are uncompressed and we refer this as linear MFCCs.

3.3 Other Features

Other time and frequency domain features that we have considered in this work, such as zero-crossing rate (ZCR), short-time energy (STE), sub-band energy (SBE), spectral centroid (SC) or brightness, bandwidth (BW), spectral roll-off (SR), are as defined [16].

4.0 Support Vector Machine

4.1 Basic Theory

A support vector machine determines the optimal hyperplane to maximize the distance between any two given classes. It has been well described in many literature, such as in [34-36], and is summarized here. Starting with a case of linearly separable dataset, consider a set of l training samples belonging to two classes, a positive class and a negative class, given as $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i \in R^d$ is a d-dimensional feature vector representing the i^{th} training sample, and $y_i \in \{-1, +1\}$ is the class label of $\mathbf{x}_i$. There can be many possible hyperplanes but the two classes can be said to be optimally separated by the hyperplane if the separation distance, or margin, between the closest vector, known as support vectors, to the hyperplane is maximal.

Any hyperplane in the feature space can be described by the equation $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w} \in R^d$ is a normal vector to the hyperplane and b is a constant. Selecting two hyperplanes, $\mathbf{w} \cdot \mathbf{x} + b = +1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$ such that the data points are separated with no data between them in the margin region, the aim then is to maximize the distance between them. The distance between these two hyperplanes is given as $^{2}/_{||\mathbf{w}||}$, therefore, $||\mathbf{w}||$ has to be minimized. To prevent the data points from falling into the margin, the following constraints are added: $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$. For mathematical convenience, and without altering the solution, $||\mathbf{w}||$ is substituted with $½||\mathbf{w}||^2$ which becomes a quadratic programming problem. The optimization problem can be solved under the given constraints by the saddle point of the Lagrange functional. For ease of computation, the primal problem is transformed to a dual problem using classical Lagrangian duality which gives the solution

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \tag{8}$$

where α_i are the non-negative Lagrange multipliers. The $\mathbf{x}_i$ for which $\alpha_i > 0$ are called the support vectors which lie exactly on the margin satisfying $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1,\ i = 1, 2, \ldots l$. The offset can then be determined as

$$b = y_i - \mathbf{w} \cdot \mathbf{x}_i \tag{9}$$

using any support vector or averaged over all support vectors.

However, there is no such hyperplane for linearly nonseparable problems to classify every training sample correctly. In such a case, the optimization can be generalized by introducing the concept of *soft margin* [35] implying a hyperplane separating most but not all the points. Introducing non-negative slack variables ξ_i which measure the degree of misclassification of data $\mathbf{x}_i$ and a penalty function $\sum_i \xi_i$, the optimization is a trade-off between a large margin and a small error penalty. The optimization problem can be solved as before and the solution is similar to the separable case except for a modification to the Lagrange multipliers: $0 \leq \alpha_i \leq C, i = 1, 2, \dots l$, where C is a penalty or tuning parameter to balance the margin and training error.

In applications where linear SVM does not give satisfactory results, nonlinear SVM is suggested which aims to map the input vector $\mathbf{x}$ to a higher dimensional space $\mathbf{z}$ through some nonlinear mapping $\phi(\mathbf{x})$ chosen *a priori* to construct an optimal hyperplane. The *kernel trick* [34] is applied to create the nonlinear classifier where the dot product is replaced by a nonlinear kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ which computes the inner product of the vectors $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$.

The typical kernel functions are: polynomial, $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^r$ where *r* is the degree of the polynomial; Gaussian RBF, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\right)$, where $\sigma > 0$ is the width of the Gaussian function; and multilayer perception, $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(a_1(\mathbf{x}_i \cdot \mathbf{x}_j) - a_2)$, where a_1 and a_2 are two given parameters known as *scale* and *offset* respectively.

The classifier for a given kernel function with the optimal separating hyperplane is then given as

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{10}$$

4.2 Multiclass Classification

OAA is probably the earliest of the multiclass SVM classification techniques [36, 37]. For an M-class problem, M binary SVM classifiers are constructed and evaluated where each classifier separates one class from all the other classes combined. That is, the i^{th} classifier is trained with all the training samples from the i^{th} class as positive labels and all the remaining samples as negatives labels. During classification, a sample $\mathbf{x}$ is classified in the class with the largest value of the decision function

$$f(\mathbf{x}) = \arg \max_{i=1,2,\dots,M} \left(\mathbf{w}^i \cdot \phi(\mathbf{x}) + b^i\right). \tag{11}$$

The disadvantage of OAA-SVM is the high mismatch in the training samples between the positive and negative classes while some literature [38, 39] also shows that the training and evaluation times can be high.

The OAO approach distinguishes between every pair of classes and classification is done using a max-wins voting strategy [40]. For an M-class problem, OAO-SVM constructs and evaluates $M(M-1)/2$ classifiers where each SVM is trained on samples from two classes at a time, that is, using training samples from the i^{th} and j^{th} class. During classification, the class label of a test sample is predicted as

$$f\left(\mathbf{x}\right) = \arg \max_{i=1,2,\ldots,M} \sum_{j=1,j\neq i}^{M} \operatorname{sgn}\left(\mathbf{w}^{ij} \cdot \phi\left(\mathbf{x}\right) + b^{ij}\right). \tag{12}$$

While OAO-SVM has much more uniform training samples in the positive and negative classes when compared to OAA-SVM, its disadvantage is the inefficiency of classifying data because the number of SVM classifiers grows super linearly with an increase in the number of classes.

DDAG [38] and ADAG [41] are also based on classification between pair of classes but utilize a decision tree structure in the testing phase. Similar to OAO-SVM, $M(M-1)/2$ nodes are created during the training phase but only $M-1$ nodes are evaluated during testing.

5.0 Experimental Evaluation

A description of the sound database used in this work is given first followed by an overview of the noise conditions and experimental setup. We then present results using MFCCs and the RSIF. The final set of results use a combination of features.

5.1 Description of the Sound Database

The sound database consists of 10 classes mostly taken from the Real World Computing Partnership (RWCP) Sound Scene database in Real Acoustic Environment [42] and the BBC Sound Effects library [43]. All signals in the database have 16-bit resolution and a sampling frequency of 44100 Hz. A summary of the selected sound classes, total number of sound files, and total duration is shown in Table I.

Table I: Overview of sound classes

	Class	Number of Subclasses	Total Number of Files	Total Duration (s)
Alarms	A	6	180	83.4533
Children Voices	B	6	180	131.9286
Construction	C	3	90	26.2251
Dog Barking	D	3	84	22.3042
Footsteps	E	6	171	24.0566
Glass Breaking	F	2	60	107.3296
Gunshots	G	3	84	8.9500
Horn	H	3	66	27.4115
Machines	I	3	90	56.8423
Phone Rings	J	6	138	119.7996
Total			1143	608.3008

The choice of sound classes is similar to that of other audio surveillance applications, [6] in particular. *Alarm* sounds in the database include car alarms, electronic alarms, and siren. *Children voices* include children crying and screaming. *Construction* sounds are sawing, metal hammering, and pneumatic drilling. The *footstep* sounds include those from metal and wooden stairs and on pavement. The three types of *machine* sounds are machine hum, motor, and warble. The *phone rings* class includes cellphone and telephone ringtones.

The database has both harmonic and impulsive sounds and an irregular number of sound files which are important in testing out the robustness of the system. It is also important to have some degree of intraclass diversity and interclass similarity for this purpose and this is demonstrated using K-means clustering [44]. The centroid of each of the subclasses was determined and these were grouped into 10 clusters using K-means clustering algorithm. The results for these are shown in Table II where A_B and A_A show the subclasses in class A before and after applying K-means clustering, respectively.

As an example, there are six type of *alarm* sounds (class A_B) which have been labeled as $A_1, A_2, \dots, A_6$. However, after applying K-means clustering, the six subclasses fall in five different clusters: A_1 and A_2 in class A_A, A_5 in class B_A, A_4 in class C_A, A_3 in class D_A, and A_6 in class H_A. This means that only A_1 and A_2 have similar signal properties. There are three subclasses in *construction* and all fall in different clusters, B_A, C_A, and G_A, unlike the subclasses from *dog barking*, *glass breaking*, and *horn* which all fall in the same cluster, D_A, F_A, and H_A, respectively, but have been combined with subclasses from other classes.

Table II: Demonstration of intraclass diversity and interclass similarity using K-means clustering

Normal Cluster		After K-means Clustering	
Class	Subclasses	Class	Subclasses
A_B	A_1 A_2 A_3 A_4 A_5 A_6	A_A	A_1 A_2 J_2 J_3 J_4 J_6
B_B	B_1 B_2 B_3 B_4 B_5 B_6	B_A	A_5 B_1 B_2 B_5 C_1 J_1
C_B	C_1 C_2 C_3	C_A	A_4 C_2 I_1
D_B	D_1 D_2 D_3	D_A	A_3 B_3 B_4 B_6 D_1 D_2 D_3
E_B	E_1 E_2 E_3 E_4 E_5 E_6	E_A	E_3 E_4 E_5 E_6
F_B	F_1 F_2	F_A	F_1 F_2 G_3
G_B	G_1 G_2 G_3	G_A	C_3 G_1 G_2
H_B	H_1 H_2 H_3	H_A	A_6 H_1 H_2 H_3 J_5
I_B	I_1 I_2 I_3	I_A	I_2
J_B	J_1 J_2 J_3 J_4 J_5 J_6	J_A	E_1 E_2 I_3

5.2 Noise Conditions

The performance of the different features and classification methods are investigated under three different noise environments taken from the NOISEX-92 database [26]: *speech babble*, *factory floor 1*, and *destroyer control room*. The signals are resampled at 44100 Hz and the performance is measured in clean conditions and at 20dB, 10dB, 5dB, and 0dB SNR.

5.3 Experimental Setup

For all experiments, features are extracted from a Hamming window of 512 points (11.61 ms) with 50% overlap. The system is trained with two-third of the clean samples with the remaining one-third samples used for testing under clean conditions and with the addition of noise.

Classification accuracy, given in percentage as *number of correctly classified test samples* divided by the *total number of test samples*, is compared using four multiclass SVM classification techniques, OAA, OAO, DDAG, and ADAG. All results are reported using a nonlinear SVM with a Gaussian RBF kernel as it was found to give the best results during preliminary experiments. The penalty parameter C and σ for the Gaussian RBF kernel were tuned using cross validation. For DDAG and ADAG, the class order list in alphabetical order was used. Results using KNN classification with Euclidean distance measure are also presented for comparison.

5.4 Results with MFCCs

With MFCCs, the feature vector for each frame is 39-dimensional: 13 MFCCs, using a 20-filterbank system, plus deltas and accelerations. The overall feature vector dimension for a signal is $39 \times T$, where T is the total number of frames in the sound signal, which is different in each case depending on the length of the signal. After data normalization, the feature vector is represented by concatenating the mean and standard deviation for each feature which seems to be the most commonly used technique as seen in [14, 18, 19]. As such, the final feature vector is 78-dimensional. We also experimented with the feature data representation technique given in [6] but found it to be less effective.

We experimented with both log compressed MFCCs and root compressed MFCCs. For root compressed MFCCs, various values of γ were experimented with and best results were obtained for values of γ closer to 1. However, linear MFCCs, $\gamma = 1$, were found to be more effective when combined with other features, results for which are presented in section 5.6. As such, only results using log MFCCs and linear MFCCs are presented here.

The classification accuracy with log and linear MFCCs is given in Table III. For log MFCCs, the classification accuracy in clean conditions is 98.43% for the SVM methods and 96.85% for KNN. However, the classification accuracy reduces greatly with the addition of noise, especially at 10dB, 5dB, and 0dB SNR with the highest classification accuracy at 73.14%, 57.57%, and 43.31%, respectively, all of which are using OAA-SVM classification. With linear MFCCs, significant improvement can be seen in the classification accuracy at 10dB, 5dB, and 0dB SNR using all classification methods.

Table III: Classification accuracy using MFCCs

Classification Method	Log MFCCs						Linear MFCCs					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
OAA-SVM	98.43	92.83	73.14	57.57	43.31	**73.05**	99.21	93.53	86.09	70.87	47.16	79.37
OAO-SVM	98.43	92.04	63.52	47.51	33.42	66.98	98.69	94.84	83.38	64.57	43.31	76.96
DDAG-SVM	98.43	93.09	61.07	45.67	33.25	66.30	98.69	94.75	83.03	64.30	42.26	76.61
ADAG-SVM	98.43	93.00	63.60	49.26	35.70	68.00	98.69	95.36	83.55	64.92	42.52	77.01
KNN	96.85	91.25	58.79	43.83	32.20	64.58	97.38	92.56	83.90	73.32	57.83	**81.00**

In both set of results, the OAA-SVM classification method gives the best average classification accuracy of the four multiclass SVM classification methods. While there isn't a significant difference in the classification accuracy using the four methods in clean conditions and at 20dB SNR, the OAA-SVM classification method performs much better than the other methods at 10dB, 5dB, and 0dB SNR. It also outperforms KNN in the case of log MFCCs but KNN gives slightly better performance with linear MFCCs.

5.5 Results with RSIF

We now present the results using the RSIF. With the RSIF, the spectrogram image is divided into 9×9 blocks and second and third central moments were computed in each block. We experimented with $3 \times 3, 5 \times 5,$ and 7×7 blocks as well but best results were obtained with 9×9 blocks. It was seen that the classification accuracy increased with an increase in the number of blocks but 9×9 was the maximum that could be experimented with due to limitations in the length of the sound signal and the size of the spectrogram image as a result.

The classification accuracy using RSIF for the grayscale and quantized spectrograms is given in Table IV(a) and Table IV(b), respectively. There is a significant improvement in the average classification accuracy when compared with log MFCCs, from 73.05% to 81.66% with the linear quantized spectrogram. Also, the most improved results are under noisy conditions at 10dB, 5dB, and 0dB SNR. At 81.08%, the average classification accuracy using the linear grayscale spectrogram is only slightly below the linear quantized spectrogram despite a three times smaller feature dimension. The OAA-SVM classification method once again gives the best overall results in all cases except with the log grayscale spectrogram, where all the multiclass SVM classification methods give comparable performance. As with linear MFCCs, which performs only marginally below the linear spectrogram methods, the

KNN classification method was more effective with the linear spectrograms and only slightly below the OAA-SVM classification method in terms of overall performance.

Table IV(a): Classification accuracy using RSIF – grayscale spectrograms

Classification Method	Linear Grayscale Spectrogram						Log Grayscale Spectrogram					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
OAA-SVM	92.13	92.04	89.33	78.57	53.37	**81.08**	96.06	72.00	50.48	38.93	31.32	57.76
OAO-SVM	92.13	86.70	82.33	72.79	48.29	76.45	97.11	73.05	50.22	38.93	30.80	58.02
DDAG-SVM	91.86	87.40	82.50	66.67	44.97	74.68	97.38	72.62	48.91	38.15	30.62	57.53
ADAG-SVM	90.81	86.70	82.15	71.92	48.29	75.98	97.90	74.45	50.57	39.55	31.50	**58.79**
KNN	87.93	87.23	83.90	78.48	56.61	78.83	93.44	64.57	39.28	30.27	24.23	50.36

Table IV(b): Classification accuracy using RSIF – quantized spectrograms

Classification Method	Linear Quantized Spectrogram						Log Quantized Spectrogram					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
OAA-SVM	95.80	95.28	88.89	72.79	55.56	**81.66**	99.74	84.08	59.93	51.01	40.59	**67.07**
OAO-SVM	94.75	92.83	81.10	63.52	41.12	74.66	99.21	83.90	57.22	46.54	36.92	64.76
DDAG-SVM	94.23	93.44	82.41	63.34	38.76	74.44	99.21	83.64	56.08	45.93	36.66	64.30
ADAG-SVM	95.01	94.93	85.04	65.18	44.09	76.85	99.21	84.25	56.26	46.19	36.92	64.57
KNN	93.70	93.61	92.13	75.85	47.16	80.49	97.64	80.66	57.83	44.88	34.65	63.13

A time-frequency image represents two-dimensional data which makes it more useful for feature extraction when compared to the one-dimensional data available in time-domain and frequency-domain representation of the signal on its own. The log grayscale approach gives the highest classification accuracy in clean conditions which can be expected since taking log power reveals the details in the low power frequencies unlike the linear grayscale approach where only the dominant power frequencies are shown. This can be visualized in the linear grayscale and log grayscale images in Figure 1(a) and (c), respectively. However, the performance of the two representations reverses with the addition of noise. The noise is more diffuse than the sound signal and its power affects most of the frequencies in the log grayscale image as shown in Figure 1(d). For the linear representation, the strong peaks of the sound are larger than the noise and remain largely unaffected with the addition of noise as can be seen in Figure 1(b).

In [14], mapping the grayscale spectrograms to a higher dimensional space greatly improved the results. In our work, the increase in the average classification accuracy is minimal for the linear quantized spectrogram but greatly improved with the log quantized spectrogram. However, it still does not match the performance of the linear representations. More on the effect of quantization can be found in [14].

We also experimented with the SIF data representation method proposed in [14], which has been summarized in section 3.1.3. The results are given in Table V using OAA-SVM classification. While the average classification accuracy is slightly higher for the log representations when compared to the proposed RSIF data representation technique, the results with the linear representations are lower. The proposed RSIF method has the added advantage of a feature vector which is 2.25 times smaller in dimension, as explained in section 3.1.3. As such, the proposed method can be said to be much more effective for its dimension.

Table V: Classification accuracy using SIF data representation as in [14]

Spectrogram Representation	Clean	20dB	10dB	5dB	0dB	Average
Linear Grayscale	91.60	91.34	88.80	67.19	40.51	75.89
Log Grayscale	93.70	70.60	54.59	44.36	36.22	59.90
Linear Quantized	93.70	93.70	84.69	68.07	45.14	77.06
Log Quantized	98.95	87.58	64.04	51.53	41.21	68.66

5.6 Results with Combined Features

In the next experiment, we combine log and linear MFCCs with the RSIF to form the final feature vector for a sound signal. For the RSIF, we use the linear grayscale representation only since this approach was shown to give the best results in preliminary experiments. The feature vector dimension for the combined feature set is $(39 \times 2) + (9 \times 4 \times 2) = 150$.

The classification accuracy with the combined features is given in Table VI. In both set of results, OAA-SVM classification method gives the highest average classification accuracy and also the most noise robust performance. In addition, at 10dB, 5dB, and 0dB SNR, the combination of linear MFCCs and linear grayscale spectrogram features performs significantly better than log MFCCs in combination with linear grayscale spectrogram features for all classification methods.

Table VI: Classification accuracy with MFCCs + RSIF (linear grayscale)

Classification Method	Log MFCC + Linear Grayscale						Linear MFCC + Linear Grayscale					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
OAA-SVM	97.38	95.98	90.11	80.49	57.13	**84.22**	97.11	96.06	93.61	89.15	70.95	**89.38**
OAO-SVM	97.11	92.21	76.12	63.25	42.87	74.31	96.33	91.16	85.04	76.38	56.87	81.15
DDAG-SVM	97.11	93.53	79.88	64.13	45.23	75.98	95.80	91.60	88.80	80.23	53.28	81.94
ADAG-SVM	97.11	89.76	69.73	57.39	42.69	71.34	95.01	87.49	76.90	70.08	50.13	75.92
KNN	96.85	91.34	58.79	43.92	32.28	64.64	97.38	92.56	83.90	73.40	57.92	81.03

We also experimented with the inclusion of various time and frequency domain features mentioned in section 3.3 but only SBE, ZCR, and STE was shown to give any improvement in the classification accuracy. The classification accuracy values with the inclusion of these three features are given in Table VII.

Table VII: Classification accuracy with linear MFCCs + RSIF (linear grayscale) + ZCR + STE + SBE

Classification Method	Clean	20dB	10dB	5dB	0dB	Average
OAA-SVM	98.16	96.41	94.23	90.81	71.83	**90.29**
OAO-SVM	96.33	92.39	85.04	76.47	57.04	81.45
DDAG-SVM	96.06	93.44	88.80	81.01	53.72	82.61
ADAG-SVM	95.28	89.50	76.99	71.30	51.62	76.94
KNN	97.38	92.48	81.98	73.40	58.88	80.82

OAA-SVM classification method once again gives the best average results and is seen to be more noise robust than the other classification methods. The better performance of the OAA classification method over the other methods under noisy conditions could be explained in terms of its decision function. In OAA classification, the class corresponding to the largest margin is declared the winner indicating a high confidence level in the decision. However, in the other three multiclass SVM classification methods, the final decision is based on classification between pair of classes. The class even with the slightest of margin wins and gets a vote in the case of OAO classification method or proceeds to the next round as in the case of DDAG and ADAG classification methods. The hyperplane between classes has been determined using clean samples only and with the addition of noise, there could be more overlapping of data points meaning the hyperplane is no longer an optimal one. As such, chances of error with the OAO, DDAG, and ADAG methods are increased more than the OAA method.

Multi-conditional training provides a more optimal hyperplane since it is trained with noise manipulated samples as well as clean samples. This explains the comparable classification accuracy obtained using OAO, DDAG, and ADAG methods as per the results using multi-conditional training in [16, 17]. However, multi-conditional training increases the training time and also makes the classifier noise dependent. In addition, while KNN classification was seen to be very effective with linear MFCCs and linear spectrogram image features, it is seen to be ineffective in classification of combined feature sets.

Furthermore, for features extracted from the linear spectrograms, which have been shown to be more noise robust than the log spectrograms, the results achieved using the proposed RSIF method of data representation are better than the SIF method of data representation given in [14]. We also achieved significant improvement in the performance under low SNRs using a combination of linear MFCCs and RSIF (linear grayscale). The inclusion of some time and frequency domain features also gave a marginal improvement in the performance with classification accuracy values of 98.16%, 96.41%, 94.23%, 90.81%, and 71.83% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNR, respectively. Except at 0dB SNR, all these values are higher than in [6], a related work the results for which are summarized in section 2. In this work, we considered 10 sound classes while 9 sound classes were used in [6]. We have used the same sound and noise databases but they used some hand recorded signals as well. While 6 of the 10 classes in our work are same as in [6], the choice of subclasses and selection of sound files may not necessarily be the same. As such, it cannot be conclusively said that our approach is better unless we have exactly the same experimental conditions.

In Table VIII and IX, we present the confusion matrices under clean conditions and at 0dB SNR, respectively, with OAA-SVM classification method for the results obtained using the combined feature set in Table VII. The confusion matrix allows the observation of the degree of confusion between the different classes which gives a better understanding of the classification performance when compared to the overall classification accuracy results presented so far. The rows of the confusion matrix denote the sound classes that we want to classify and the columns denote the classified results. The values are given in percentage as *number of correctly (or incorrectly) classified samples* divided by *number of test samples in the class*.

As an example, for the confusion matrix under clean conditions given in Table VIII, 98.33% of test samples from *children voices* were correctly classified while the remaining 1.67% were misclassified as *footsteps*. It can be said that there is only one-sided confusion between *children voices* and *footsteps* because test samples from *children voices* are misclassified into *footsteps* but not vice-versa. In fact, all the misclassifications in this case have one-sided confusion.

Table VIII: Confusion matrix for test samples under clean conditions with the combined feature set and OAA-SVM classification

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	100.00	0	0	0	0	0	0	0	0	0
Children voices	0	98.33	0.00	0	1.67	0	0	0	0	0
Construction	0	3.33	96.67	0	0	0	0	0	0	0
Dog barking	0	0	0	100.00	0	0	0	0	0	0
Footsteps	0	0	0	0	100.00	0	0	0	0	0
Glass breaking	0	5.00	0	0	0	95.00	0	0	0	0
Gunshots	0	0	0	0	0	0	100.00	0	0	0
Horn	0	0	0	0	0	0	0	100.00	0	0
Machines	0	0	0	0	0	0	0	0	100.00	0
Phone Rings	0	8.70	0	0	0	0	0	0	0	91.30
Overall Classification Accuracy = 98.16%										

While only four classes have misclassifications under clean conditions, all classes have misclassifications when tested with samples at 0dB SNR as per confusion matrix given in Table IX. Most of the misclassifications are into *children voices* which is consistent with the results using clean samples. In fact, all classes now have misclassifications into *children voices*. However, *alarms*, *dog barking*, *gunshots*, and *horn* are the worst performing classes with classification accuracy of less than 60%, largely due to the high misclassifications into *children voices*.

Table IX: Confusion matrix for test samples at 0dB SNR with the combined feature set and OAA-SVM classification

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	48.33	31.67	1.67	0	1.67	0	0	0	10.00	6.67
Children voices	0	83.33	5.56	0	0.56	4.44	0	0	6.11	0
Construction	0	7.78	92.22	0	0	0	0	0	0	0
Dog barking	0	44.05	3.57	52.38	0	0	0	0	0	0
Footsteps	0	4.68	15.20	0	80.12	0	0	0	0	0
Glass breaking	0	3.33	0.00	0	0	96.67	0	0	0	0
Gunshots	0	14.29	21.43	0	3.57	0	58.33	0	2.38	0
Horn	0	30.30	4.55	0	1.52	3.03	0	59.09	1.52	0
Machines	0	11.11	1.11	0	0	2.22	0	0	85.56	0
Phone Rings	0	10.87	5.07	0	0	0	0	0	13.77	70.29
Overall Classification Accuracy = 71.83%										

Finally, we compare the training and evaluation time for the four multiclass SVM classification methods. These are provided in Table X for the combined feature set, the results for which are given in Table VII. The OAO, DDAG, and ADAG approaches have the same training procedure and time. The training time for OAA is about 19% higher than these three classification methods.

The DDAG and ADAG classification methods have approximately the same evaluation time and are the fastest. Using the ADAG evaluation time as basis, OAA classification method takes about 35% longer while OAO classification method takes a significantly greater time, about 386% longer. The significantly higher evaluation time for the OAO classification method can be expected since it requires the evaluation for 45 classifiers per test sample when compared to only 9 classifiers for DDAG and ADAG classification methods. As such, ideally, the OAO approach should take 400% longer time to evaluate.

Table X: Comparison of training and evaluation time of the multiclass SVM classification methods with the combined feature set

Classification Method	Training Time (s)	No. of Classifiers Evaluated per Test Sample ($M = 10$)	Total Testing Time (s)
OAA-SVM	0.498	10	28.223
OAO-SVM	0.420	45	101.403
DDAG-SVM	0.420	9	20.944
ADAG-SVM	0.420	9	20.876

6.0 Conclusion

The overall classification accuracy of the proposed feature set which combines linear MFCCs and the RSIF (linear grayscale) produces much better results under noisy conditions when compared to the individual features. In general, the OAA multiclass SVM classification approach was seen to give the best overall classification accuracy together with being more noise robust. However, the training time of this classification method is slightly longer than the other multiclass SVM classification methods and it also has a slightly longer evaluation time when compared to DDAG and ADAG classification methods.

While the proposed method is noise independent, that is, it does not require multi-conditional training, it is yet to be tested under other noise types. It will also be interesting to test the proposed method with other sound databases and with increasing number of sound classes.

Author Biography

Roneel V Sharan completed the B.E.Tech. and M.Sc. degrees in electrical and electronic engineering from the University of the South Pacific (USP), Fiji, in 2003 and 2006, respectively. He was a Graduate Assistant at USP from 2004 to 2005 and, later, an Assistant Lecturer. He is currently pursuing Ph.D. degree at the Auckland University of Technology (AUT), New Zealand, under an AUT doctoral scholarship. His current research interests are signal processing, pattern recognition, and image processing. He is a member of IEEE.

Tom J Moir was born in Dundee Scotland. He was sponsored by GEC Industrial Controls Ltd, Rugby Warwickshire UK from 1976 to 1979 during his B.Sc in control engineering which he was awarded in 1979. In 1983 he received the degree of Ph.D for work on self-tuning filters and controllers. From 1982 to 1983 he was with the Industrial Control unit University of Strathclyde, Scotland. From 1983 to 1999 he was a lecturer then senior lecturer at Paisley College/University of Paisley, Scotland. Moving to Auckland, New Zealand in

2000, he was with Massey University for 10 years at the Institute of Information and Mathematical Sciences followed by the School of Engineering and Advanced Technology. He moved to Auckland University of Technology in 2010 as an Associate Professor in the School of Engineering where he works in the area of signal processing and automatic control engineering. He has authored over 100 publications in these fields and is chairman of the Signals and Systems group. He is the holder of one US patent on amplitude-locked loop circuits. Dr. Moir is an IET member and member of FEANI and IPENZ.

References

[1] E. Wold, T. Blum, D. Keislar, J. Wheaten, Content-based classification, search, and retrieval of audio, IEEE MultiMedia, 3 (3) (1996) 27-36.

[2] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing, 10 (5) (2002) 293-302.

[3] A.A. Wieczorkowska, Z.W. Ras, Z. Xin, R. Lewis, Multi-way hierarchic classification of musical instrument sounds, in: International Conference on Multimedia and Ubiquitous Engineering (MUE '07), 2007, pp. 897-902.

[4] K. Abe, H. Sakaue, T. Okuno, K. Terada, Sound classification for hearing aids using time-frequency images, in: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim), 2011, pp. 719-724.

[5] S. Chu, S. Narayanan, C.C.J. Kuo, Environmental sound recognition with time-frequency audio features, IEEE Transactions on Audio, Speech, and Language Processing, 17 (6) (2009) 1142-1158.

[6] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, Using one-class SVMs and wavelets for audio surveillance, IEEE Transactions on Information Forensics and Security, 3 (4) (2008) 763-775.

[7] F. Jin, F. Sattar, D.Y.T. Goh, New approaches for spectro-temporal feature extraction with applications to respiratory sound classification, Neurocomputing, 123 (2014) 362-371.

[8] B. Lei, S.A. Rahman, I. Song, Content-based classification of breath sound with enhanced features, Neurocomputing, 141 (2014) 139-147.

[9] A.R. Abu-El-Quran, R.A. Goubran, A.D.C. Chan, Security monitoring using microphone arrays and audio classification, IEEE Transactions on Instrumentation and Measurement, 55 (4) (2006) 1025-1032.

[10] D. Istrate, E. Castelli, M. Vacher, L. Besacier, J.F. Serignat, Information extraction from sound for medical telemonitoring, IEEE Transactions on Information Technology in Biomedicine, 10 (2) (2006) 264-274.

[11] V.T. Vu, F. Bremond, G. Davini, M. Thonnat, P. Quoc-Cuong, N. Allezard, P. Sayd, J.L. Rouas, S. Ambellouis, A. Flancquart, Audio-video event recognition system for public transport security, in: The Institution of Engineering and Technology Conference on Crime and Security, 2006, 2006, pp. 414-419.

[12] J. Kotus, K. Lopatka, A. Czyżewski, G. Bogdanis, Audio-visual surveillance system for application in bank operating room, in: A. Dziech, A. Czyżewski (Eds.) Multimedia Communications, Services and Security, Springer Berlin Heidelberg, 2013, pp. 107-120.

[13] X. Zhang, Y. Li, Environmental sound recognition using double-level energy detection, Journal of Signal and Information Processing, 4 (3B) (2013) 19-24.

[14] J. Dennis, H.D. Tran, H. Li, Spectrogram image feature for sound event classification in mismatched conditions, IEEE Signal Processing Letters, 18 (2) (2011) 130-133.

[15] R. Sarikaya, J.H. Hansen, Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition, in: EUROSPEECH-2001, Aalborg, Denmark, 2001, pp. 687-690.

[16] R.V. Sharan, T.J. Moir, Comparison of multiclass SVM classification techniques in an audio surveillance application under mismatched conditions, in: Proceedings of the 19th International Conference on Digital Signal Processing (DSP 2014), Hong Kong, 2014, pp. 83-88.

[17] R.V. Sharan, T.J. Moir, Audio surveillance under noisy conditions using time-frequency image feature, in: Proceedings of the 19th International Conference on Digital Signal Processing (DSP 2014), Hong Kong, 2014, pp. 130-135.

[18] S.Z. Li, Content-based audio classification and retrieval using the nearest feature line method, IEEE Transactions on Speech and Audio Processing, 8 (5) (2000) 619-625.

[19] G. Guo, S.Z. Li, Content-based audio classification and retrieval by support vector machines, IEEE Transactions on Neural Networks, 14 (1) (2003) 209-215.

[20] L. Lu, H.-J. Zhang, S.Z. Li, Content-based audio classification and segmentation by using support vector machines, Multimedia Systems, 8 (6) (2003) 482-492.

[21] S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Transactions on Signal Processing, 41 (12) (1993) 3397-3415.

[22] G. Valenzise, L. Gerosa, M. Tagliasacchi, E. Antonacci, A. Sarti, Scream and gunshot detection and localization for audio-surveillance systems, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007), 2007, pp. 21-26.

[23] B. Uzkent, B.D. Barkana, H. Cevikalp, Non-speech environmental sound classification using SVMs with a new set of features, International Journal of Innovative Computing, Information and Control, 8 (5(B)) (2012) 3511-3524.

[24] W. Huang, S. Lau, T. Tan, L. Li, L. Wyse, Audio events classification using hierarchical structure, in: Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia, Singapore, 2003, pp. 1299-1303.

[25] B. Schölkopf, A.J. Smola, Learning with kernels, MIT Press, Cambridge, MA, 2002.

[26] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication, 12 (3) (1993) 247-251.

[27] Y.M.G. Costa, L.S. Oliveira, A.L. Koericb, F. Gouyon, Music genre recognition using spectrograms, in: 18th International Conference on Systems, Signals and Image Processing (IWSSIP), 2011, pp. 1-4.

[28] P. Khunarsal, C. Lursinsap, T. Raicharoen, Very short time environmental sound classification based on spectrogram pattern matching, Information Sciences, 243 (2013) 57-74.

[29] S. Souli, Z. Lachiri, Multiclass support vector machines for environmental sounds classification using log-Gabor filters, World Academy of Science, Engineering and Technology, 68 (2012) 1185-1189.

[30] S. Kolozali, M. Barthet, G. Fazekas, M. Sandler, Automatic ontology generation for musical instruments based on audio analysis, IEEE Transactions on Audio, Speech, and Language Processing, 21 (10) (2013) 2207-2220.

[31] S. Essid, G. Richard, B. David, Musical instrument recognition by pairwise classification strategies, IEEE Transactions on Audio, Speech, and Language Processing, 14 (4) (2006) 1401-1412.

[32] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, 13 (2) (2002) 415-425.

[33] N. Seo, A comparison of multi-class support vector machine methods for face recognition, The University of Maryland, Research Report, 6 Dec 2007.

[34] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA, 1992, pp. 144-152.

[35] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning, 20 (3) (1995) 273-297.

[36] V.N. Vapnik, Statistical learning theory, Wiley, New York, 1998.

[37] L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, L.D. Jackel, Y. LeCun, U.A. Muller, E. Sackinger, P. Simard, V. Vapnik, Comparison of classifier methods: a case study in handwritten digit recognition, in: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image Processing, 1994, pp. 77-82.

[38] J.C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, in: S.A. Solla, T.K. Leen, K.-R. Müller (Eds.) Advances in Neural Information Processing Systems 12 (NIPS-99), MIT Press, Cambridge MA, 2000, pp. 547-553.

[39] G. Madzarov, D. Gjorgjevikj, Evaluation of distance measures for multi-class classification in binary SVM decision tree, in: L. Rutkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, J. Zurada (Eds.) Artificial Intelligence and Soft Computing, Springer, Berlin Heidelberg, 2010, pp. 437-444.

[40] U.H.G. Kreßel, Pairwise classification and support vector machines, in: B. Schölkopf, C.J.C. Burges, A.J. Smola (Eds.) Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 255-268.

[41] B. Kijsirikul, N. Ussivakul, S. Meknavin, Adaptive directed acyclic graphs for multiclass classification, in: M. Ishizuka, A. Sattar (Eds.) PRICAI 2002: Trends in Artificial Intelligence, Springer, Berlin Heidelberg, 2002, pp. 158-168.

[42] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, in: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 2000, pp. 965–968.

[43] BBC Sound Effects Library. Available: http://www.leonardosoft.com

[44] S. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory, 28 (2) (1982) 129-137.