Subband Time-Frequency Image Texture Features for Robust Audio Surveillance

Roneel V. Sharan, *Member, IEEE*, and Tom J. Moir

***Abstract*—In this paper, we utilize time-frequency image representations of sound signals for feature extraction in an audio surveillance application. Starting with the conventional spectrogram images, we consider a new feature which is based on image texture analysis. It utilizes the gray-level co-occurrence matrix (GLCM), which captures the distribution of co-occurring values at a given offset. We refer this as the spectrogram image texture feature (SITF). Texture analysis is carried out in subbands and experimented on a sound database containing 10 classes with each sound class containing multiple subclasses. The proposed feature was seen to be more noise robust than two commonly used cepstral features, mel-frequency cepstral coefficients (MFCCs) and gammatone cepstral coefficients (GTCCs), the spectrogram image feature (SIF), where central moments are extracted as features, and a variation of SIF with reduced feature dimension (RSIF). In addition, we achieved significant improvement in classification accuracy for the three time-frequency image features by utilizing a gammatone filter-based time-frequency image, referred as cochleagram image, for feature extraction instead of the spectrogram image. A combination of cepstral and cochleagram image features also gave improvement in the classification performance.**

***Index Terms*—Audio surveillance, cochleagram, gammatone filter, gray-level co-occurrence matrix, spectrogram, support vector machines**

I. Introduction

AUTOMATIC sound recognition has many applications. Some of these include music genre classification [1], audio surveillance [2, 3], computer keystroke sound recognition for user identification [4], and biometric identification using heart sound [5]. Sound classification is a pattern recognition problem and being a relatively new area of research, most of the techniques involved are inspired from other pattern recognition problems, speech recognition in particular. This applies to the choice of features and classifiers.

Cepstral features, mel-frequency cepstral coefficients (MFCCs), in particular, have been the traditional feature choice in sound recognition. MFCCs have been shown to be effective in structured environments but its classification performance is poor in the presence of noise [6]. However, features extracted from the spectrogram image of speech or sound signals have proved effective for classification in the presence of noise [6, 7]. The intensity values in the spectrogram image represent the dominant frequency components against time. Features which capture this information can improve the recognition rate in the presence of additive noise provided the noise spectrum does not contain strong spectral peaks.

In [7], spectral subband centroids (SSCs) are used as supplementary features to achieve improvement in classification accuracy in the presence of noise in speech recognition. For robust sound event classification in [6], the spectrogram image is divided into multiple blocks and second and third central moments are computed in each block which forms the feature vector, referred as the spectrogram image feature (SIF). In [3], we proposed a technique to reduce the dimension of the SIF without compromising the classification accuracy, which we referred as the reduced spectrogram image feature (RSIF).

In this work, we propose a number of improvements when compared to our earlier work in [3] in trying to achieve robust sound recognition in an audio surveillance application. We consider various features for this purpose which can be broadly categorized as cepstral features and time-frequency image features. For cepstral features, in [3] we considered MFCCs only but in this work we also present results using gammatone cepstral coefficients (GTCCs), which we found to be more robust than MFCCs. Similarly, we consider the SIF and RSIF as the spectrogram image derived features, as in [3]. However, we also consider a new feature based on the image texture analysis technique of gray-level co-occurrence matrix (GLCM), also known as gray-tone spatial dependence matrix [8], which gives the spatial relationship of pixels in an image. We refer this as the spectrogram image texture feature (SITF) [9]. Furthermore, for all the time-frequency image features, we propose feature extraction using a gammatone filter-based time-frequency image, referred as a cochleagram [10], instead of the conventional spectrogram image. In the case of cochleagram feature extraction, we refer the features SIF, RSIF, and SITF as CIF, RCIF, and CITF, respectively.

In addition, feature vector combination has been shown to

Preliminary results from this work were presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 19–24 April 2015, Brisbane, and at the IEEE International Conference on Digital Signal Processing (DSP), 21–24 July 2015, Singapore.

The authors are with the School of Engineering, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand (e-mail: roneel.sharan@aut.ac.nz; tom.moir@aut.ac.nz).

improve the classification performance in a number of literature. A combination of cepstral features and SSCs improved the robustness in speech recognition in [7]. Various feature combinations were experimented with in a similar work in [2]. In [3], we achieved significant improvement in classification accuracy under noisy conditions with a combination of linear MFCCs and RSIF. In this work, we propose a combination of linear GTCCs and cochleagram image derived features in trying to achieve further improvement in classification performance when compared to the individual features on their own.

The rest of this paper is organized as follows. Section II provides literature review and background on the methods that we are proposing for this work. In section III and IV, we discuss the current methods or feature extraction techniques and the proposed framework, respectively. The experimental results, discussions, and analysis are presented in section V and conclusions and recommendations are given in section VI.

II. Literature Review

Every sound signal produces a unique texture which can be visualized using a spectrogram image and analyzed for automatic sound recognition. In music genre recognition in [1], texture analysis is carried out using the GLCM texture analysis technique. The spectrogram image is firstly divided into zones for feature extraction. Due to the non-uniform nature of the sound signal spectrograms, this local feature extraction technique was shown to give higher results than global features. The following seven features are then extracted from the GLCM from the fourteen textural descriptors proposed in [8]: entropy, correlation, homogeneity, third order momentum, maximum likelihood, contrast, and energy.

The GLCM method of image texture analysis using the fourteen textural descriptors of [8], a subset of these features, or with other textural descriptors has been employed in various other applications. These include insect recognition [11], fabric surface roughness evaluation [12], and urban and agricultural land classification [13]. It has also been applied for diagnosis of abdominal tumors using texture classification of ultrasound images [14] and mammogram texture classification for breast cancer detection [15]. In [16], however, instead of extracting features from the GLCM, the matrix values itself are used to form the feature vector in a face recognition problem. This approach was generally shown to give significantly better results than using the combined fourteen textural descriptors as features. We adopted this approach in [9] but as a spectrogram image texture analysis tool for sound classification. Texture analysis was carried out in subbands and instead of extracting textural descriptors from the GLCM, we concatenated the columns of the matrix to form the feature vector, which we referred as the SITF.

While the spectrogram image is the most commonly used tool in time-frequency analysis of sound signals, it may not be the best choice depending on the application. Short-time Fourier transform (STFT) is a commonly used method for spectrogram image formation where the signal is divided into short duration frames and discrete Fourier transform (DFT) is applied to the windowed frames. The spectrum values from each frame are stacked side-by-side to form the spectrogram image. The spectrogram image shows dominant frequency information against time and the frequency components are equally spaced along the vertical with constant bandwidth. However, most sound signals hold greater frequency components in the lower frequency range and, therefore, the information in these frequency bands are not fully revealed in this time-frequency representation.

A cochleagram is a variation of the spectrogram which uses a gammatone filter. A gammatone filter is a linear filter modeling the frequency selectivity property of the human cochlea. A commonly used cochlea model is that proposed by Patterson et. al. [17] which is a series of bandpass filters where the bandwidth is given by equivalent rectangular bandwidth (ERB). An efficient implementation of the gammatone filter bank is provided in [18] which has been used for computing GTCCs [19] and extended to gammatone wavelet (GTW) features in a similar application [20]. Feature extraction using cochleagram images finds applications in a number of areas involving signal processing and pattern recognition such as speech recognition [21] and audio separation [22]. In this work, we also use the cochleagram image for feature extraction but for sound classification. We consider the same features as for the spectrogram images and compare the classification performance against the spectrogram image derived features.

We consider a total of 10 sound classes to evaluate the robustness of the proposed features. The performance of the features is evaluated under clean conditions and in the presence of different noise environments at different signal-to-noise ratios (SNRs) using support vector machines (SVMs) for classification. Being a binary classifier, a number of techniques have been proposed for multiclass classification. The most common technique is to reduce the multiclass classification problem into multiple binary classification problems. Four commonly 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). In [3], we compared the performance of these four methods and the kNN classifier and found the OAA multiclass [23] classification method to be generally more noise robust with a better overall performance. Therefore, in this work, we report results using OAA multiclass classification method only. Refer to [3] for an overview of SVMs, the multiclass classification methods, and the comparison of classification performance.

In addition, for the problem of audio surveillance considered in this work, including the choice of sound and noise databases, we take a similar approach to [2], which is one of the most comprehensive piece of work in this area. Their sound database has a total of 1015 sound files with 9 sound classes: *human screams*, *gunshots*, *glass breaking*, *explosions*, *door slams*, *dog barks*, *phone rings*, *children voices*, and *machines*. Each sound class has multiple subclasses with interclass similarity and intraclass diversity.

They considered various features which can be broadly classified as time-domain, frequency-domain, cepstral, and wavelet-based features. The highest classification accuracy values achieved with the best performing feature set are 96.89% under clean conditions and 93.33%, 89.22%, 82.80%, and 72.89% at 20dB, 10dB, 5dB, and 0dB SNRs, respectively, with 70% of clean data used for training and the remaining for testing. While we take a similar approach with the experimental setup, our method is based on time-frequency image features.

III. Current Methods

In this section, we review some of the features used in similar previous works which we have considered in this work. We first describe the cepstral features which includes MFCCs and GTCCs. Spectrogram image generation and feature extraction for SIF and RSIF are explained next.

A. Cepstral Features

1) MFCCs

In computation of MFCCs, firstly, the DFT is applied to the windowed signal as

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

where N is the length of the window, $x(n)$ 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 triangular filter banks used in computing MFCCs are equally spaced on the mel-scale [24] and the adjacent filters overlap such that the lower and upper end of a filter are located at the center frequency of the previous and next filter, respectively, while the peak of the filter is at its center frequency. The output of the m^{th} filter can then be determined as

$$E(m,t)=\sum_{k=0}^{\frac{N}{2}-1}V(m,k)\left|X(k,t)\right|^2,\qquad m=1,2,...,M_1 \tag{2}$$

where $E(m,t)$ represents the filter bank energies, $V(m,k)$ is the normalized filter response, and M_1 is the total number of mel-filters. Some literature do not take square of the DFT values in computing the filter bank energies but we achieved better results using this approach. The results without taking the square of the DFT values is given in our earlier work in [3].

The MFCCs are then obtained as the discrete cosine transform (DCT) of the log compressed filter bank energies given as

$$c(l,t)=\sqrt{\frac{2}{M_1}}\sum_{m=1}^{M_1}\log\left(E(m,t)\right)\cos\left(\frac{\pi l}{M_1}(m-0.5)\right) \tag{3}$$

which is evaluated from $l=1,2,\dots,L$, where L is the order of the cepstrum.

We also report results using linear MFCCs where no compression is applied to the filter bank energies before computing the cepstral coefficients which was seen to be more noise robust in [3].

2) GTCCs

Extraction of GTCCs follow the same procedure as MFCCs except that gammatone filters are used instead of mel-filters. Gammatone filter banks are a series of bandpass filters the impulse response for which can be given as [17]

$$g(r)=Ar^{j-1}e^{-2\pi Br}\cos\left(2\pi f_c r+\phi\right) \tag{4}$$

where A is the amplitude, j is the order of the filter, B is the bandwidth of the filter, f_c is the center frequency of the filter, ϕ is the phase, and r is the time.

The ERB is used to describe the bandwidth of each cochlea filter in [17]. ERB is a psychoacoustic measure of the auditory filter width at each point along the cochlea and can be given as

$$f_{c,ERB}=\left[\left(\frac{f_{c,Hz}}{Q_{ear}}\right)^p+\left(B_{min}\right)^p\right]^{1/p} \tag{5}$$

where Q_{ear} is the asymptotic filter quality at high frequencies and B_{min} is the minimum bandwidth for low frequency channels. The bandwidth of a filter can then be approximated as $B=1.019\times f_{c,ERB}$. The three commonly used ERB filter models are given by Glasberg and Moore [25] ($Q_{ear}=9.26$, $B_{min}=24.7$, and $p=1$), Lyon's cochlea model as given in [26] ($Q_{ear}=8$, $B_{min}=125$, and $p=2$), and Greenwood [27] ($Q_{ear}=7.23$, $B_{min}=22.85$, and $p=1$).

The human cochlea has thousands of hair cells which resonate at their characteristic frequency and at a certain bandwidth. In [18], the mapping between center frequency and cochlea position is determined by integrating the reciprocal of (5) with a step factor parameter to indicate the overlap between filters. This can then be inverted to find the mapping between filter index and center frequency which can be given as

$$f_{cm}=-Q_{ear}B_{min}+\left(f_h+Q_{ear}B_{min}\right)e^{-ms/Q_{ear}} \tag{6}$$

where $m=1,2,\dots,M_2$, M_2 is the number of gammatone filters, f_h is the maximum frequency in the filter bank, and s is the step factor given as

$$s=\frac{Q_{ear}}{M_2}\log\left(\frac{f_h+Q_{ear}B_{\min}}{f_l+Q_{ear}B_{\min}}\right) \tag{7}$$

where f_l is the minimum frequency in the filter bank.

We use a 4th order gammatone filter with four filter stages and each stage a 2nd order digital filter as given in [18]. The gammatone filter was implemented using the Auditory Toolbox for Matlab [28].

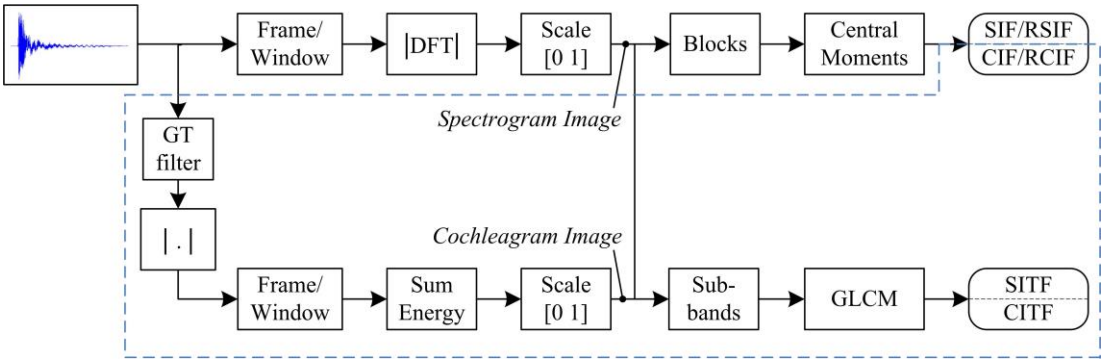


Fig. 1. Steps in time-frequency image generation and feature extraction (proposed time-frequency representation and feature extraction methods are enclosed in dashed lines).

B. Time-Frequency Image Features

The procedure for time-frequency image generation and feature extraction is explained with reference to Fig. 1.

For the SIF and RSIF, central moments are extracted as features from the spectrogram images. To obtain the spectrogram image, the linear values are firstly obtained from the DFT values as

$$S(k,t)=\left|X(k,t)\right|. \tag{8}$$

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

$$I(k,t)=\frac{S(k,t)-\min(S)}{\max(S)-\min(S)}. \tag{9}$$

Illustration of spectrogram images, color representations are shown for the grayscale values for better visualization, under clean conditions and with the addition of noise at 0dB SNR can be found in Fig. 2(a) and (b), respectively.

The time-frequency image is divided into blocks and the v^{th} central moment for any given block of image is then determined as

$$\mu_v=\frac{1}{K}\sum_{i=1}^{K}(I_i-\mu)^v \tag{10}$$

where K is the sample size or the number of pixels in the block, I_i is the intensity value of the i^{th} sample in the block, and μ is the mean intensity value of the block.

IV. Proposed Framework

In this section, we present the proposed feature, SITF, and the proposed time-frequency image representation, cochleagram. The steps in the proposed feature extraction and time-frequency image generation are given in Fig. 1.

A. SITF

The intensity values in a spectrogram image are determined by the spectral energy in the sound signal at any given time and frequency. The dominant frequency components in the sound signal are mostly unaffected by the noise as long as the noise signal does not contain strong spectral peaks, as shown in the spectrogram images in Fig. 2(a) – (b) with *factory* noise. As such, the SITF aims to capture the patterns of the subband spectral energy in trying to achieve noise robust classification performance.

The SITF uses the GLCM method of texture analysis which is a matrix of frequencies where each element (i,j) is the number of times intensity value j is located at a certain distance and angle, given by the displacement vector $[d_k\ d_t]$, where d_k is the offset in the y direction and d_t is the offset in the x direction, from intensity value i in an $N_t \times N_k$ image I. Mathematically, this can be given as

$$P(i,j)=\sum_{k=1}^{N_k}\sum_{t=1}^{N_t}\begin{cases}1, & \text{if } I(k,t)=i \ \&\ I(k+d_k,t+d_t)=j\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

where the size of the output matrix is $N_g \times N_g$, N_g is the number of quantized gray levels. The typical angles for computing the GLCM are $0°, 45°, 90°,$ and $135°$ corresponding to the displacement vector $[0\ d]$, $[-d\ d]$, $[-d\ 0]$, and $[-d\ -d]$, respectively. The feature vector for SITF is then formed by concatenating the GLCM values into a column vector.

B. Cochleagram

The cochleagram is another form of time-frequency representation and it mimics the components of the outer and middle ear [29]. In this representation, the signal is broken into different frequencies which are naturally selected by the cochlea and hair cells. This frequency selectivity can be modeled by a filter bank, such as a gammatone filter.

A representation similar to the conventional spectrogram image can be obtained by smoothing the time series associated with each frequency channel of the gammatone filter and then adding the energy in the windowed signal for each frequency component which can be given as

$$C(m,t)=\sum_{n=0}^{N-1}\left|\hat{x}(m,n)\right|w(n), \qquad m=1,2,...,M_2 \tag{12}$$

where $\hat{x}(n)$ is the gammatone filtered signal and $C(m,t)$ is the m^{th} harmonic corresponding to the center frequency f_{cm} for

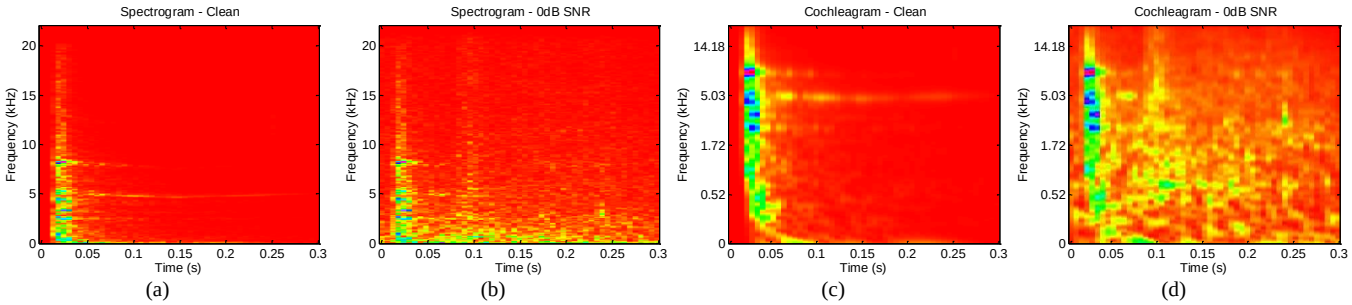


Fig. 2. Spectrogram and cochleagram images for a sample sound signal from *construction* sound class. (a) Spectrogram image under clean conditions, (b) spectrogram image at 0dB SNR with *factory* noise, (c) cochleagram image under clean conditions, and (d) cochleagram image at 0dB SNR with *factory* noise.

the t^{th} frame.

These values are then normalized using (9) to get the grayscale cochleagram image intensity values.

Illustration of cochleagram images under clean conditions and with the addition of noise at 0dB SNR can be found in Fig. 2(c) and (d), respectively, using the same sound signal as the spectrogram images of Fig. 2(a) and (b).

C. Motivation

The GLCM essentially captures the frequency of repeating patterns or intensity value combinations in the time-frequency image. We use only two intensity levels, $N_g = 2$, as determined to give the best results in [9], which means that the grayscale time-frequency image is treated as a binary image for feature extraction, therefore, revealing only the dominant frequency components. This also means that small linear transformations caused by the noise to the intensity values of the sound signal in the time-frequency image would not affect its transformation to binary format as long as the threshold for binary conversion is not crossed. In addition, as shown in Fig. 2(b) and (d), the noise affects only certain frequency bands in the time-frequency images and the use of subband feature extraction, with the optimal number of subbands determined as 64 in [9], ensures that feature data in subbands not affected by noise remain unchanged.

This is better illustrated in Fig. 3(a) and (b) where we have the normalized spectral energy distribution of a sound signal for the spectrogram image and cochleagram image, respectively. The spectral energy, in this context measured as the number of white pixels in the binary transformed image, is given in each of the 64 subbands without noise and with noise at 0dB SNR. The noise mostly affects subbands 13, 18, and 19 in the spectrogram image and subbands 40, 45, and 46 in the cochleagram image. Otherwise, there is generally a good degree of correlation between the energy distributions of the clean and noisy signals in both representations. As such, except in these bands, the repeating patterns captured by the GLCM will largely remain unchanged from clean to 0dB SNR conditions, explaining the usefulness of the proposed feature extraction technique.

In addition, while the spectrogram and cochleagram images of Fig. 2 use the same frequency range, $\left[0, {}^{F_s}/_2\right]$, the cochleagram offers a number of advantages [29]. Firstly, with

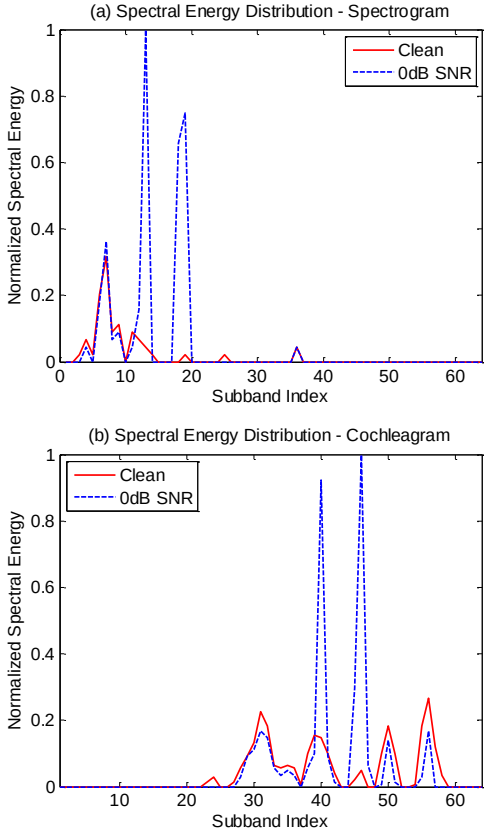


Fig. 3. Subband spectral energy distribution of a sound signal from *construction* sound class with and without noise for (a) spectrogram and (b) cochleagram.

the ERB spacing of the filter center frequencies, the cochleagram offers an expanded representation at low frequencies, where most of the spectral information lies. Secondly, depending on the type of sound signal, formants in the lower frequencies can be resolved into harmonics in the cochleagram since they have a narrower bandwidth. Therefore, a cochleagram offers more frequency components in the lower frequency range with narrower bandwidth and fewer frequency components in the higher frequency range with wider bandwidth, showing much more spectral information than a spectrogram, as a result. The cochleagram also emphasizes acoustic onsets which can be effective for audio separation [22].

The difference in the spread of spectral energy for the two representations is also illustrated in Fig. 3. For example, for the spectrogram image, the spectral energy is mainly distributed between subbands 2 to 20 and subbands 26 to 59 for the cochleagram image, that is, over 18 subbands for the spectrogram image and 33 subbands for the cochleagram image. As such, the cochleagram image clearly reveals more spectral information which makes it a more effective time-frequency representation for feature extraction.

V. Experimental Evaluation

A description of the database of sounds used in this work is given first followed by an overview of the noise conditions and the experimental setup. We then present results using the cepstral features, MFCCs and GTCCs. This is followed by the results for the spectrogram image features, SIF, RSIF, and SITF. Results for the three time-frequency image features but using the cochleagram image for feature extraction are presented next. We then present results using feature vector combinations and, finally, the performance of the proposed techniques is analyzed.

A. Sound Database

The sound database has a total of 1143 files belonging to 10 classes: *alarms*, *children voices*, *construction*, *dog barking*, *footsteps*, *glass breaking*, *gunshots*, *horn*, *machines*, and *phone rings*. Each sound class contains multiple subclasses with interclass similarity and intraclass diversity as demonstrated in [3]. The sound files are largely obtained from the Real World Computing Partnership (RWCP) Sound Scene database in Real Acoustic Environment [30] and the BBC Sound Effects library [31]. All signals in the database have 16-bit resolution and a sampling frequency of 44100 Hz. More details about the sound database and its comparison with that used in other similar work can be found in [3].

B. Noise Conditions

The performance of all features is evaluated under three different noise environments taken from the NOISEX-92 database [32]: *speech babble*, *factory floor 1*, and *destroyer control room*. The signals are resampled at 44100 Hz and the performance is evaluated in clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs.

C. Experimental Setup

For all experiments, signal processing is carried out using a Hamming window of 512 points (11.61 ms) with 50% overlap. The classification accuracy is given in percentage as *number of correctly classified test samples* divided by the *total number of test samples*. All results are reported using nonlinear SVM with a Gaussian radial basis function (RBF) kernel as it was found to give the best results. The classifier parameters, refer to [3], were tuned using cross validation. In tuning the parameters, one set of parameters which gave the best average classification accuracy were selected rather than determining the optimal parameters for each noise condition. For all experimentations, the classifier is trained with two-third of the clean samples with the remaining one-third of the samples used for testing under clean and noisy conditions.

D. Results using Cepstral Features

For the cepstral features, MFCCs and GTCCs, in determining the optimal minimum and maximum frequency limits for the mel and gammatone filter banks, we set the limits as multiples of the sampling frequency with the lower limit as $\left[0, {}^{F_s}/_{N}, {}^{2F_s}/_{N}, {}^{3F_s}/_{N}, {}^{4F_s}/_{N}\right]$ and the upper limit as $\left[{}^{F_s}/_{8}, {}^{F_s}/_{4}, {}^{3F_s}/_{8}, {}^{F_s}/_{2}\right]$. We also experimented with various number of filter channels and the fined-tuned parameter settings were determined as: MFCCs: $M_1 = 26$, $f_l = 258.4$ Hz, and $f_h = 16537.5$ Hz; GTCCs: $M_2 = 24$, $f_l = 172.27$ Hz, and $f_h = 16537.5$ Hz. Also, of the three ERB filter models considered for GTCCs, Lyon's filter model was shown to give the best results so we present results using this model only. More details on such parameter tuning can be found in [19].

For both the features, the feature vector for each frame is 36-dimensional: 12 cepstral coefficients plus the first and second derivatives. The overall feature vector dimension for a signal is $36 \times N_t$, where N_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 dimension. As such, the final feature vector is 72-dimensional.

We experimented with both log compressed cepstral coefficients and root compressed cepstral coefficients [33]. In root cepstrum, instead of applying log compression to the filter bank energies, it is raised to the root value, γ, normally in the range $0 < \gamma \leq 1$. We experimented with various root values and achieved the best results around the root value of 1. Therefore, for both MFCCs and GTCCs, we use $\gamma = 1$, which we refer as linear cepstrum. This means that no compression is applied to the filter bank energies.

The classification accuracy values for MFCCs and GTCCs using log and linear compression are given in Table I. Both the cepstral features give highest average classification accuracy with linear cepstrums. While there isn't a significant difference in the classification accuracy for log and linear cepstrums under clean conditions and at 20dB SNR, the classification accuracy is considerably better at 10dB, 5dB, and 0dB SNRs with linear cepstrums. GTCCs give the highest average classification accuracy for both compression methods.

E. Results using Spectrogram Image Features

We now present results using the spectrogram image derived features. For the SIF and RSIF, the spectrogram image is divided into 9×9 blocks and second and third central moments are computed in each block. We also experimented with $3 \times 3, 5 \times 5$, and 7×7 blocks but best results were obtained with 9×9 blocks which was the maximum that could be experimented with due to limitations in the length of the sound signal and the length of the spectrogram image as a result. For the SIF, the central moment values computed in

TABLE I
CLASSIFICATION ACCURACY VALUES USING MFCCS AND GTCCS

	Log						Linear					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
MFCC	97.11	92.21	73.32	60.54	47.77	74.19	96.85	93.44	84.95	74.28	58.88	81.68
GTCC	96.33	94.58	77.78	70.43	55.03	78.83	96.85	93.96	87.75	80.93	61.77	**84.25**

TABLE II
CLASSIFICATION ACCURACY VALUES USING SIF, RSIF, AND SITF

	Clean	20dB	10dB	5dB	0dB	Average
SIF	91.60	91.34	88.80	67.19	40.51	75.89
RSIF	92.13	92.04	89.33	78.57	53.37	81.08
SITF	89.76	89.41	89.33	87.66	71.92	**85.62**

each block are concatenated to form the final feature vector which is 162-dimensional. With the RSIF, the mean and standard deviation of the central moment values along the rows and columns of the blocks are concatenated to form the final feature vector which is 72-dimensional.

The classification accuracy values using SIF and RSIF are given in Table II. The best average classification accuracy is achieved using RSIF which is also the most noise robust. When compared to the conventional cepstral features, that is, log-compressed cepstrums, the average classification accuracy using RSIF is significantly better than MFCCs and marginally better than GTCCs. The RSIF is generally seen to be more effective at 10dB, 5dB, and 0dB SNRs. While the average classification accuracy using RSIF is only slightly lower than linear MFCCs, it is not able to match the average classification accuracy of linear GTCCs, which, at 84.25%, is the best performing baseline feature.

SITF was presented in our earlier work in [9] where we performed classification with feature vector extracted using GLCM analysis at the individual angles, $0°, 45°, 90°,$ and $135°$, and then with a combined feature vector. The results showed that the combined feature vector gave only marginally better average classification accuracy then the individual feature vectors and had the disadvantage of a feature vector which was four times more than the individual feature vectors. Of the four angles considered, feature vector from analysis at an angle of $45°$ generally gave the best average classification accuracy. Therefore, in this work, we present results with analysis at an angle of $45°$ only. In addition, each subband spans four frequency bins with $N_g = 2$ and $d = 1$, as determined to give the optimal results in [9]. As such, for the SITF, the final feature vector dimension is 256 ($N_g^2 \times 64$, where 64 refers to the number of subbands). The classification accuracy values using the SITF are given in Table II.

With an average classification accuracy of 85.62% for the SITF, it could be said that the proposed feature gives significantly better performance than the SIF and RSIF. The classification performance is also higher than the best performing cepstral feature, linear GTCCs, which produced an average classification accuracy of 84.25%. The SITF is unable to match the classification accuracy of linear GTCCs under clean conditions and at 20dB SNR, however, marginally higher classification accuracy is achieved at 10dB SNR and significantly better classification accuracy is achieved at 5dB and 0dB SNRs. This makes the SITF more noise robust than linear GTCCs.

F. Results using Cochleagram Image Features

Cochleagram feature extraction follows the same procedure as the spectrogram images but now using a cochleagram image. To get the same image resolution as the spectrogram images, the number of gammatone filters, M_2, is set to 256 with the same window size, $N = 512$. The classification accuracy values for CIF and RCIF are given in Table III and the classification accuracy values for CITF are given in Table IV. The results in each case are presented using the three ERB filter models.

The average classification accuracy values for CIF and RCIF with all the ERB filter models show significant improvement when compared to SIF and RSIF, respectively. The highest average classification accuracy for both CIF and RCIF is achieved using Greenwood [27] parameters. As such, the average classification accuracy value increases from 75.89% using SIF to 86.30% using CIF, an increase of 10.41%, and from 81.08% using RSIF to 89.03% using RCIF, an increase of 7.95%. In addition, the improvement in the classification accuracy increases as the SNR decreases. From SIF to CIF, the classification accuracy value increases 1.58%, 1.75%, 3.41%, 21.87%, and 23.44% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs, respectively. Similarly, from RSIF to RCIF, the classification accuracy value increases 2.62%, 2.71%, 5.25%, 13.12%, and 16.01% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs, respectively. This shows that while the classification accuracy value under all noise conditions has improved, the most improved results are at low SNRs, 5dB and 0dB SNRs, in particular. In addition, the overall results for CIF and RCIF are significantly higher than linear GTCCs, and, as with spectrogram image feature extraction, the reduced feature method once again gives the best average classification accuracy.

Furthermore, for the CITF, the highest average classification accuracy is achieved using Glasberg and Moore [25] parameters, as per the results in Table IV. There is also an improvement in the average classification accuracy when compared to the spectrogram based features, increasing from 85.62% with SITF to 89.24% with CITF, an increase of 3.62%. For the individual noise conditions, the improvement in classification accuracy is 2.89%, 3.24%, 2.88%, 2.72%, and 6.38% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs, respectively.

TABLE III
CLASSIFICATION ACCURACY VALUES USING CIF AND RCIF FOR THE THREE ERB FILTER MODELS

ERB Filter Model	CIF						RCIF					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
Glasberg and Moore [25]	92.13	91.78	90.73	85.74	63.08	84.69	94.75	94.58	94.14	89.68	65.44	87.72
Lyon [26]	91.60	91.25	90.46	83.38	58.88	83.11	95.01	94.40	93.35	89.59	65.44	87.56
Greenwood [27]	93.18	93.09	92.21	89.06	63.95	**86.30**	94.75	94.75	94.58	91.69	69.38	**89.03**

TABLE IV
CLASSIFICATION ACCURACY VALUES USING CITF FOR THE THREE ERB FILTER MODELS

ERB Filter Model	CITF					
	Clean	20dB	10dB	5dB	0dB	Average
Glasberg and Moore [25]	92.65	92.65	92.21	90.38	78.30	**89.24**
Lyon [26]	92.13	91.78	91.34	89.41	80.75	89.08
Greenwood [27]	91.86	91.78	91.78	89.85	78.04	88.66

TABLE V
CLASSIFICATION ACCURACY VALUES USING LINEAR GTCCS COMBINED WITH COCHLEAGRAM IMAGE FEATURES

Linear GTCCs +	Clean	20dB	10dB	5dB	0dB	Average
CIF	96.06	95.98	95.28	93.35	76.99	91.53
RCIF	97.64	97.38	96.59	91.51	79.79	92.58
CITF	96.59	96.59	95.36	94.23	83.73	**93.30**

Therefore, all the time-frequency image features show improvement in classification accuracy under all noise conditions when using a cochleagram image for feature extraction instead of the spectrogram image. Unlike CIF and RCIF, the improvement in classification accuracy value is generally much more even for CITF and the improvement in the average classification accuracy lower. Also, while the RSIF could not match the average classification accuracy of the SITF, RCIF gives comparable average classification accuracy to CITF. However, CITF can be considered the most noise robust feature with a classification accuracy of 78.30% at 0dB SNR for the best overall performing ERB filter model.

G. Results using Feature Combinations

Results using a combination of cepstral and time-frequency image features is presented in this subsection. For the cepstral features, we consider linear GTCCs only since it gives the best classification accuracy of all the cepstral features. Similarly, we use cochleagram image features only since they give much better classification performance than spectrogram image features. The classification accuracy values using a combination of linear GTCCs and the cochleagram image features, with the best overall performing ERB model used in each case, is given in Table V.

The average classification accuracy values for all cochleagram image features show some improvement when combined with linear GTCCs. The improvement is 5.23%, 3.55%, and 4.06% for CIF, RCIF, and CITF, respectively. As such, CIF combined with linear GTCCs gives the most improved results. However, CITF is once again the best performing feature with an average classification accuracy of 93.30% when combined with linear GTCCs. In addition, this combination also gives the most noise robustness performance with a classification accuracy of 94.23% and 83.73% at 5dB and 0dB SNRs, respectively.

H. Performance Analysis

The proposed method of feature extraction using the GLCM gives the most noise robust performance and also the best overall classification performance with spectrogram feature extraction, cochleagram feature extraction, and when combined with linear GTCCs. The peak of the filter bank energies play a key role in characterizing a sound signal which is demonstrated by the superior performance of both the cepstral features under clean conditions. However, the conventional log compression can produce high variations in the output for low energy components [34] which explains its poor performance as the SNR decreases. While the introduction of linear cepstrums improved the noise robustness, the cochleagram image derived features give a far superior performance at low SNRs.

While we have been presenting the overall classification accuracy values so far, to understand the classification performance between classes, we present the classification and misclassification values of classes. The confusion matrix for the CITF, the best performing individual feature, under clean conditions and in the presence of noise at 0dB SNR is given in Table VI and Table VII, respectively. The values in the confusion matrix are given in percentage as *number of correctly (or incorrectly) classified samples* divided by *number of test samples in the class*. The rows in the confusion matrix denote the classes that we intend to classify while the classified results are given in the columns.

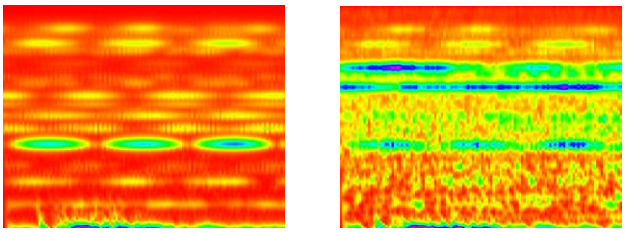
For example, for the confusion matrix under clean conditions given in Table VI, 96.67% of the test samples from *alarms* were correctly classified while the remaining 3.33% were misclassified into *children voices*, which includes children crying and screaming. *Dog barking*, *footsteps*, and *glass breaking* also have misclassification in one class only while *gunshots*, *horn*, *machines*, and *phone rings* are the best performing classes with no misclassifications. *Children voices* and *construction* are the worst performing classes with a classification accuracy of 70% and 83.33%, respectively, with both classes also having multiple misclassifications. In addition, there is only one-sided confusion between *footsteps* and *dog barking* whereby test samples from *footsteps* are misclassified into *dog barking* but not vice-versa. *Alarms*, *construction*, *dog barking*, and *glass breaking* have two-sided confusion with *children voices* whereby test samples from each of these classes is misclassified into *children voices* and

TABLE VI
CONFUSION MATRIX FOR TEST SAMPLES UNDER CLEAN CONDITIONS USING CITF

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	96.67	3.33	0	0	0	0	0	0	0	0
Children voices	3.33	70.00	5.00	11.67	6.67	1.67	0	1.67	0	0
Construction	0	6.67	83.33	0	0	6.67	0	0	0	3.33
Dog barking	0	3.57	0	96.43	0	0	0	0	0	0
Footsteps	0	0	0	1.75	98.25	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	0	0	0	0	0	0	0	0	100.00
Overall Classification Accuracy = 92.65%										

TABLE VII
CONFUSION MATRIX FOR TEST SAMPLES AT 0DB SNR USING CITF (MISCLASSIFICATIONS OF MORE THAN 10% HAVE BEEN HIGHLIGHTED)

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	86.11	7.22	0	6.67	0	0	0	0	0	0
Children voices	3.89	61.67	7.78	12.78	6.11	4.44	0	3.33	0	0
Construction	0	6.67	85.56	0	0	2.22	0	0	0	5.56
Dog barking	0	5.95	0	92.86	0	0	0	0	0	1.19
Footsteps	0	1.17	4.09	2.92	77.19	0	9.36	0.58	0	4.68
Glass breaking	0	5.00	0	0	0	91.67	0.00	0	0	3.33
Gunshots	0	2.38	11.90	0	3.57	0	82.14	0	0	0
Horn	0	0	0	1.52	0	0	0	98.48	0	0
Machines	0	11.11	3.33	0	0	22.22	0	0	47.78	15.56
Phone Rings	0	5.07	0.72	1.45	0	13.04	0	0	0	79.71
Overall Classification Accuracy = 78.30%										



(a) (b)

Fig. 4. Cochleagram images of sample sound signal from sub class 1 of sound class *machines*. (a) Cochleagram image of sound signal under clean conditions and (b) cochleagram image of sound signal at 0dB SNR with *destroyer control room* noise.

vice-versa.

Looking at the confusion matrix at 0dB SNR, Table VII, all classes now have misclassifications when compared to only six classes which had misclassification(s) under clean conditions. Once again, most classes have misclassification into children voices, all except horn, which, with a classification accuracy of 98.48%, is also the best performing class and the only one not to have multiple misclassifications. While there were no misclassifications for machines under clean conditions, it is the worst performing class at 0dB SNR with a classification accuracy of just 47.78%. It also has two of the highest misclassifications into any single class, 22.22% into *glass breaking* and 15.56% into *phone rings*.

To further understand the effect of the different environmental noises on the classification performance, we computed the average classification accuracy for each noise type at 0dB SNR which are as follows: *speech babble* – 69.29%, *destroyer control room* – 78.74%, and *factory floor 1* – 86.88%. This shows that most of the misclassifications are due to *speech babble* noise while *factory floor 1* has the least misclassifications. The *machines* sound class has three subclasses and, upon further analysis, it was observed that under *destroyer control room* noise, most of the test samples from subclasses 1 and 2 were misclassified into *children voices* and *phone rings*, respectively. The cochleagram image of a sample sound signal from subclass 1 under clean conditions and with the addition of *destroyer control room* noise at 0dB SNR is shown in Fig. 4(a) and (b), respectively. The dominant frequency components of the sound signal are clearly evident under clean conditions in Fig. 4(a). While they are also largely visible with the addition of noise, Fig. 4(b), the *destroyer control room* noise introduces strong spectral peaks which significantly alters the intensity distribution in the cochleagram image, hence, making the classification task much more difficult. While we still manage a decent overall classification accuracy of 78.74% using *destroyer control room* noise at 0dB SNR, it could be said that the proposed features are more suited to noise environments which do not contain strong spectral peaks, such as *factory floor 1*, as shown in the time-frequency images in Fig. 2.

Moreover, compared to CITF, GTCCs have significantly higher confusion at 0dB SNR as seen in Table VIII. To some extent, there is a reversal in the classification performance of individual classes. For example, with CITF, *children voices*,

TABLE VIII
CONFUSION MATRIX FOR TEST SAMPLES AT 0DB SNR USING GTCCS (MISCLASSIFICATIONS OF MORE THAN 10% HAVE BEEN HIGHLIGHTED)

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	48.89	15.00	0	0	0	23.89	0	0	0	12.22
Children voices	0	85.00	1.11	0.56	0	6.11	0	0	6.11	1.11
Construction	0	7.78	64.44	0	0	23.33	4.44	0	0	0
Dog barking	9.52	54.76	2.38	22.62	0	10.71	0	0	0	0
Footsteps	0	9.36	1.75	0	54.39	32.16	2.34	0	0	0
Glass breaking	0	10.00	0	0	0	81.67	0	0	3.33	5.00
Gunshots	0	7.14	15.48	0	25.00	30.95	21.43	0	0	0
Horn	1.52	25.76	1.52	0	0	1.52	0	69.70	0	0
Machines	0	3.33	0	0	1.11	18.89	0	0	61.11	15.56
Phone Rings	0	1.45	0	0	0	0.72	0	0	5.80	92.03
Overall Classification Accuracy = 61.77%										

machines, and *phone rings* are amongst the worst performing classes at 0dB SNR. However, the classification accuracy of these classes is higher with GTCCs. For all the other classes, however, CITF gives much better classification performance than GTCCs. There are also some similar trends as far as misclassifications are concerned. With CITF, all except one class has misclassifications into *children voices* and with GTCCs, all classes have misclassifications into *children voices*. Also, misclassifications of more than 10% are most into *glass breaking* for both features, two classes for CITF and six classes for GTCCs.

Furthermore, the number of classes in this work is one more than in [2] with 66.67% of data used for training when compared to 70% in [2]. As such the classification task in this work can be considered slightly more challenging. While we attain a more noise robust performance, it is difficult to conclusively say that our approach is better due to the variations in sound and noise databases. However, in our earlier work in [3], we used exactly the same sound and noise databases with the best average classification accuracy of 90.29% achieved using a combination of linear MFCCs, RSIF, and some time and frequency domain features. The classification accuracy values achieved were 98.16%, 96.41%, 94.23%, 90.81%, and 71.83% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs, respectively. In this work, we have proposed methods to improve the classification performance of cepstral and time-frequency image features. As far as the performance of the individual features is concerned, we have achieved significant improvement in average classification accuracy and noise robustness using the CITF when compared to [3] where RSIF was the best performing feature. In addition, with the best performing feature set of linear GTCCs combined with CITF, we achieve a better overall classification performance and significantly improved results at 5dB and 0dB SNRs when compared to the combined features of [3].

VI. CONCLUSION

We considered a number of cepstral and time-frequency image features in trying to achieve improvement in classification accuracy in the presence of noise in an audio surveillance application. The proposed SITF, based on the GLCM method of image texture analysis, showed greater noise robustness when compared to two cepstral features, MFCCs and GTCCs, and two time-frequency image features, SIF and RSIF. For the three time-frequency image features, significantly improved classification performance was achieved with feature extraction using a cochleagram image, which uses a gammatone filter based on the human cochlea model. With cochleagram image feature extraction, the SIF, RSIF, and SITF were referred as CIF, RCIF, and CITF, respectively. For both time-frequency image representations, feature vector formation using the GLCM texture analysis technique gave the best overall performance. In addition, the performance of the cochleagram image features was further improved when combined with linear GTCCs, which was the best performing cepstral feature. The combination of linear GTCCs and CITF gave the best overall classification accuracy and also the most noise robust.

While the proposed features show improvement in classification performance when compared to related work, there are still a few areas to improve on. The proposed features are not well suited to noise types which contain strong spectral peaks and more research is needed in this regards to attain an even better performance. In addition, this work does not consider out-of-class sound signals which is necessary for a practical implementation of an audio surveillance system.

REFERENCES

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

[2] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 763-775, 2008.

[3] R. V. Sharan and T. J. Moir, "Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM," *Neurocomputing*, vol. 158, pp. 90-99, 2015.

[4] J. Roth, X. Liu, A. Ross, and D. Metaxas, "Investigating the discriminative power of keystroke sound," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 333-345, 2015.

[5] F. Beritelli and S. Serrano, "Biometric identification based on frequency analysis of cardiac sounds," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 596-604, 2007.

[6] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130-133, 2011.

[7] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 617-620.
[8] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, 1973.
[9] R. V. Sharan and T. J. Moir, "Robust audio surveillance using spectrogram image texture feature," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 1956-1960.
[10] R. V. Sharan and T. J. Moir, "Cochleagram image feature for improved robustness in sound recognition," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, Singapore, 2015.
[11] L.-Q. Zhu and Z. Zhang, "Auto-classification of insect images based on color histogram and GLCM," in *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2010, pp. 2589-2593.
[12] X. Wang and N. D. Georganas, "GLCM texture based fractal method for evaluating fabric surface roughness," in *Canadian Conference on Electrical and Computer Engineering (CCECE '09)*, 2009, pp. 104-107.
[13] M. Umaselvi, S. S. Kumar, and M. Athithya, "Color based urban and agricultural land classification by GLCM texture features," in *IET Chennai 3rd International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2012)*, 2012, pp. 1-4.
[14] D. Mitrea, M. Socaciu, R. Badea, and A. Golea, "Texture based characterization and automatic diagnosis of the abdominal tumors from ultrasound images using third order GLCM features," in *4th International Congress on Image and Signal Processing (CISP)*, Shanghai, 2011, pp. 1558-1562.
[15] S. Beura, B. Majhi, and R. Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, pp. 1-14, 2015.
[16] A. Eleyan and H. Demirel, "Co-occurrence matrix and its statistical features as a new approach for face recognition," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 19, no. 1, pp. 97-107, 2011.
[17] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*. vol. 83, Y. Cazals, L. Demany, and K. Horner, Eds. Pergamon, Oxford, 1992, pp. 429-446.
[18] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer, Technical Report 35, 1993.
[19] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684-1689, 2012.
[20] X. Valero and F. Alias, "Gammatone wavelet features for sound classification in surveillance applications," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, 2012, pp. 1658-1662.
[21] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029-3038, 2013.
[22] B. Gao, W. L. Woo, and L. C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1171-85, Mar 2014.
[23] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.
[24] D. O'Shaughnessy, *Speech communication: human and machine*. Addison-Wesley Pub. Co., 1987.
[25] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103-138, 1990.
[26] M. Slaney, "Lyon's Cochlear Model," Apple Computer, Technical Report 13, 1988.
[27] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *Journal of the Acoustical Society of America* vol. 87, no. 6, pp. 2592-2605, Jun 1990.
[28] M. Slaney, "Auditory Toolbox for Matlab," Interval Research Corproation, Technical Report 1998-010, 1998.
[29] D. Wang and G. J. Brown, "Fundamentals of computational auditory scene analysis," in *Computational auditory scene analysis: Principles, algorithms and applications*, D. Wang and G. J. Brown, Eds. IEEE Press/Wiley-Interscience, 2006, pp. 1-44.
[30] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and 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.
[31] *BBC Sound Effects Library*. Available: http://www.leonardosoft.com
[32] A. Varga and 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*, vol. 12, no. 3, pp. 247-251, Jul. 1993.
[33] R. Sarikaya and 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.
[34] V. Tyagi and C. Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 2005, pp. 529-532.

Roneel V. Sharan (M'08) completed M.Sc. in engineering from the University of the South Pacific (USP), Fiji, in 2006. He is currently pursuing the Ph.D. degree at the School of Engineering, Auckland University of Technology (AUT), New Zealand.

From 2004 to 2005, he was a graduate assistant at the School of Engineering at USP and an assistant lecturer from 2006 to 2013. His research interests include pattern recognition, image processing, and signal processing.

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.