An Overview of Applications and Advancements in Automatic Sound Recognition

*Roneel V Sharan and Tom J Moir

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

*Corresponding Author: Email: roneel.sharan@aut.ac.nz

Abstract

Automatic sound recognition (ASR) has seen an increased and wide ranging interests in recent years. In this paper, we carry out a review of some important contributions in ASR techniques, mainly over the last one and a half decades. Similar to speech recognition systems, the robustness of an ASR system largely depends on the choice of feature(s) and classifier(s). We take a wider perspective in providing an overview of the features and classifiers used in ASR systems starting from early works in content-based audio classification to more recent developments in applications such as sound event recognition, audio surveillance, and environmental sound recognition. We also review techniques that have been utilized in noise robust sound recognition systems and feature optimization methods. Finally, some of the less commonly known applications of ASR are discussed.

Keywords: automatic sound recognition, cepstral coefficients, deep neural networks, sound event recognition, support vector machines, time-frequency image

1.0 Introduction

Any given environment generally contains a number of different sounds. In early literature, these sounds were often divided into *speech* and *non-speech*. The task of non-speech sound classification is now more commonly known as automatic sound recognition (ASR). It is also referred as sound event recognition (SER) and acoustic event detection in some contexts.

An ASR system aims to recognize sounds automatically using signal processing and machine learning techniques. In concept, it is very similar to an automatic speech recognition system except that the input signal is non-speech. While research in speech recognition has received significant attention over the past few decades, research in ASR only seems to have intensified over the past two decades or so.

There are many applications of an ASR system. Initial interests in ASR were mostly centered around content-based audio classification and retrieval [1-3] and speech and non-speech recognition [4, 5]. The specific application of most of these initial works were unclear but were eventually streamlined into applications such as music genre classification [6] and musical instrument sound classification [7].

However, applications have diversified since then with interests in areas such as audio surveillance [8], sound event recognition [9], and environmental sound recognition [10]. While audio surveillance can be seen as a subclass of sound event recognition, there are some differences as discussed in section 4.1. Applications of audio surveillance and sound event recognition systems include security monitoring in a room [11] and public transport [12], intruder detection in wildlife areas [13], and monitoring of elderly people, also referred as medical telemonitoring [14]. Environmental sound recognition poses a greater challenge when compared to audio surveillance and sound event recognition applications. This is because an environmental sound can comprise a number of different sound events within the environment which can be present in different combinations at any given time.

While the applications of ASR are many, the general approach to these pattern recognition problems are same and generally inspired from techniques employed in speech recognition

systems. An overview of a statistical pattern classifier adopted in most ASR systems is given in Figure 1. The three key steps in implementing an ASR system are *signal preprocessing*, *feature extraction*, and *classification*. Signal preprocessing seeks to prepare the sound signal for feature extraction. Typically, a signal is divided into smaller *frames*, often in the range of 10-30 ms, and a *window* function is applied to smooth the signal for further analysis. Hamming window seems to be the preferred choice in most ASR systems. While speech recognition systems typically use a sampling frequency of 8000 Hz, ASR systems employ a sampling frequency of 8000 Hz or higher, common values are 16000 Hz, 22050 Hz, and 44100 Hz, largely depending on the frequency bands of the sound signals considered in the application. Depending on the sampling frequency of the signal, a frame size of 256, 512, or 1024 samples are normally chosen with some degree of overlap between adjacent frames, such as 25% or 50%, to prevent loss of information around the edges of the window. Inherent features are then extracted from the signals and the input signal is represented by a *feature vector* in a much simpler and condensed form, which is referred as *feature extraction*. The time domain signal is often transformed to frequency domain or time-frequency domain for this purpose. Based on a set of training data containing observations whose classes are known, the task of the *classifier* is then to assign unknown observations to one of the classes.

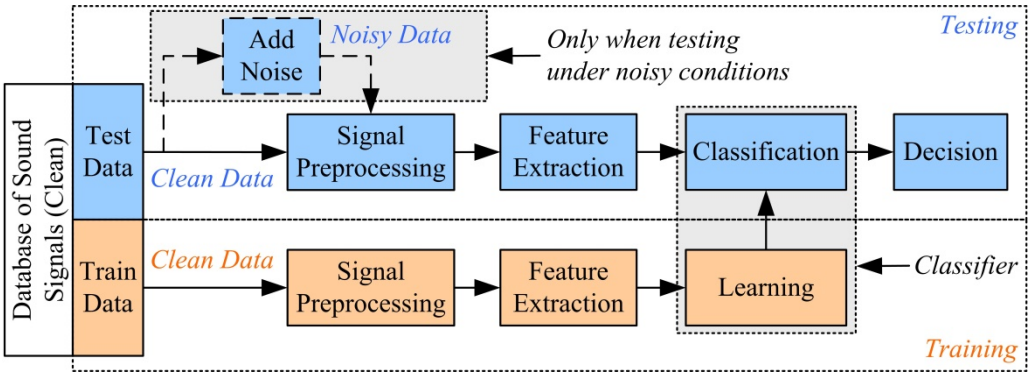


Figure 1: Model of a typical statistical pattern classifier employed in ASR systems

Features and classifiers from speech recognition systems are often employed in ASR systems. While most of the traditional features continue to be used today, they are often complemented with new features for improved performance. A thorough review of features for audio classification is provided in [15] where the features are distinguished based on its domain and can be summarized as:

- *Temporal domain* – based on the aspect of the signal the feature represents such as amplitude, power, and zero-crossing rate[1].
- *Frequency domain* – which can be further divided into *perceptual features*, which have a semantic meaning to the human listener, and *physical features*, which give description in terms of mathematical, statistical, and physical properties of the audio signal.
- *Cepstral features* – approximate the spectral envelope.
- *Modulation frequency features* – provide information on long-term amplitude or frequency variation of the signal.
- *Eigen domain features* – representing long-term information contained in sound segments with duration of several seconds.
- *Phase space features* – which capture information orthogonal to features originating from linear models.

Footnote: [1]Zero crossing rate is extracted from time domain but captures the frequency content of the signal.

Time domain, frequency domain, and cepstral features are by far the most commonly used features in ASR systems.

In addition, the commonly used classifiers are k-nearest neighbor (kNN), Gaussian mixture model (GMM), hidden Markov model (HMM), artificial neural networks (ANN), and support vector machines (SVMs), which have been well defined in many literature. While all of these continue to be used today, modifications and hybrid classification algorithms have been proposed over the years. Also, deep learning methods, such as deep neural networks (DNNs), have gained significant attention in various pattern recognition problems in recent years.

The classification performance of an ASR system is mostly reported using the *classification accuracy* which can be given in percentage as number of correctly classified test samples divided by the total number of test samples. The *error rate* (ER) can also be used for this purpose which can be stated as the number of misclassified test samples divided by the total number of test samples.

In this work, we focus on the features and classifiers used in ASR systems as seen through applications in content-based audio classification and retrieval, audio surveillance, sound event recognition, and environmental sound recognition. Due to its widespread usage in ASR systems and being a relatively new classifier, we put particular emphasis on SVMs. We also give an overview of DNNs which has seen an increased usage in speech recognition systems recently and employed in some ASR systems as well. In addition, we review literature in noise robust sound recognition and feature optimization methods, and discuss some of the lesser known applications of ASR.

2.0 Feature Extraction muscle

2.1 Time and Frequency Domain Features

One of the early works in content-based audio classification and retrieval is by Wold et al. [1] which also found commercial success and was called Muscle Fish (www.musclefish.com). It utilized some low-level acoustical features, such as loudness, pitch, brightness or spectral centroid (SC), and bandwidth (BW), with a nearest neighbor (NN) classifier based on normalized Euclidean distance. The sound database had 409 sounds files belonging to 16 classes: *alto trombone*, *animals*, *bells*, *cello bowed*, *crowds*, *female*, *laughter*, *machines*, *male*, *oboe*, *percussion*, *telephone*, *tubular bells*, *violin bowed*, *violin pizz*, and *water*. Content-based retrieval has been the main application of this work with Virage Inc., BBC, and Kodak amongst its licensees [16].

Some other commonly used time and frequency domain features include zero-crossing rate (ZCR), short-time energy (STE), subband energy (SBE), spectral flux (SF), and spectral roll-off (SR). While time and frequency domain features continue to be used in ASR systems, such as in audio surveillance applications [8, 17], they are often only used as supplementary features.

2.2 Cepstral Features

2.2.1 MFCCs

Li [2] used the Muscle Fish database and improved the ER when compared to [1] with the introduction of mel-frequency cepstral coefficients (MFCCs) as features and the nearest feature line (NFL) [18] method of classification. Humans are better at differentiating small changes in pitch at low frequencies than at high frequencies. MFCCs equally space the frequency bands on the mel-scale which more closely resembles how humans perceive sound when compared to linearly spaced cepstrums. In addition, a cepstrum gives information about

how the frequencies change in different spectrum bands. Therefore, a combination of mel-scale and cepstrum make MFCCs useful in audio classification.

In computation of MFCCs, firstly, the discrete Fourier transform (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 conversion from frequency in Hz, f_{Hz}, to frequency in mel, f_{Mel}, can be given as [19]

$$f_{Mel}=1127\log\left(1+\frac{f_{Hz}}{700}\right). \tag{2}$$

The center frequency for the m^{th} filter can be computed as

$$f_{cm}=f_l+\frac{m(f_h-f_l)}{M_1+1}, \qquad m=1,2,...,M_1 \tag{3}$$

where all the frequency values are given in mel, f_l and f_h are the minimum and maximum cut-off frequencies, respectively, and M_1 is the total number of mel-filters. 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.

For the t^{th} frame, the output of the m^{th} filter, referred as filter bank energies, can then be determined as

$$E(m,t)=\sum_{k=0}^{\frac{N}{2}-1} V(m,k)\left|X(k,t)\right|^2, \qquad m=1,2,...,M_1. \tag{4}$$

where $V(m,k)$ is the normalized frequency response.

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

$$c(i,t)=\sqrt{\frac{2}{M_1}}\sum_{m=1}^{M_1}\log\left(E(m,t)\right)\cos\left(\frac{\pi i}{M_1}(m-0.5)\right), \qquad i=1,2,...,L. \tag{5}$$

where L is the order of the cepstrum.

The number of filters in the bank is normally in the range of 20-24 with 19 filters used in [2], 20 filters in [8, 20], 23 filters in [10, 17] and 24 filters in [9]. In addition, the first and second derivatives of the coefficients, commonly known as delta and delta-delta coefficients, respectively, which provide trajectories of MFCCs over time, are often appended to improve the classification performance of ASR systems. The delta coefficients can be computed as [21]

$$c_{\Delta}(i,t) = \frac{\sum_{d=1}^{D} d\left[c(i+d,t) - c(i-d,t)\right]}{2\sum_{d=1}^{D} d^2} \tag{6}$$

where $c_{\Delta}(i,t)$ is the i^{th} delta coefficient in the t^{th} frame and the value of D is often set to 2. The same formula can be applied to the delta coefficients to compute the delta-delta coefficients, $c_{\Delta-\Delta}(i,t)$.

The feature vector for each sound signal is often represented by the mean and standard deviation values along each feature dimension. However, to reduce the effect of different environmental conditions, the coefficients are often normalized before feature vector formation. Cepstral mean and variance normalization (CMVN) is the most common data normalization method which can be given as

$$c_{CMVN}(i,t) = \frac{c(i,t) - \mu(i)}{\sigma(i)} \tag{7}$$

where $\mu(i)$ and $\sigma(i)$ are the mean and variance along the i^{th} dimension, respectively. Another data normalization method scales the data in the range [0 1], referred as cepstral scaling (CS), which can be given as

$$c_{CS}(i,t) = \frac{c(i,t) - \min\left(c(i)\right)}{\max\left(c(i)\right) - \min\left(c(i)\right)} \tag{8}$$

where $max\left(c(i)\right)$ and $min\left(c(i)\right)$ are the maximum and minimum data values along the i^{th} dimension, respectively. These formulas also apply to normalization of delta and delta-delta coefficients.

In [2], the audio is first classified as silent and non-silent where silent is defined as one which has the sum of the signal magnitude below a certain threshold. The mean and standard deviation of the features extracted from the non-silent features are then concatenated to form the feature vector with normalized values. The leave-one-out test is carried out first where each of the 409 sounds are used as query but the query sound is not used as a prototype. A combination of cepstral features and perceptual features, which includes total spectrum power, subband powers, brightness, bandwidth, and pitch, gives an improved performance than the individual features. The combined feature vector gives an error rate of 9.78% which is better than the error rate of 19.07% for the Muscle Fish system [1]. In the second test, evaluation is done using separate training and test sets, 211 files and 198 files, respectively. An error rate of 9.60% is achieved which is once again using a combination of cepstral and perceptual features.

While MFCCs are still probably the most common feature in both speech and sound recognition applications, it has been shown to perform poorly in noisy conditions [9, 22]. Multi-conditional training is one solution to this problem but it requires large datasets to capture the variations in environmental conditions.

A commonly used solution seen in speech recognition is root compression. Log spectrum compression used in conventional cepstral analysis is sensitive to noise [23]. The peaks of the mel-filter bank energies are important in characterizing the sound but the log compression can create high variations in the cepstral coefficients for low energy components [24]. Root cepstrum was proposed in [25] to improve the robustness of MFCCs. Root compressed

cepstral coefficients are computed similar to the conventional method but root compression is applied to the filter bank energies instead of log compression which can be given as

$$c(i,t) = \sqrt{\frac{2}{M_1}} \sum_{m=1}^{M_1} E(m,t)^{\gamma} \cos\left(\frac{\pi i}{M_1}(m-0.5)\right), \quad i = 1,2,...,L \tag{9}$$

where γ is the root value used to compress the filter bank energies, $0 < \gamma \leq 1$. In the event $\gamma = 1$, the filter bank energies are uncompressed which is often referred as linear cepstral coefficients. When evaluated on an audio surveillance database in [8], linear MFCCs were shown to give significant improvement in classification accuracy at low signal-to-noise ratios (SNRs) and also found to be more effective for feature vector combination.

In some other variations of MFCCs, in [26], independent component analysis (ICA) MFCCs are proposed for recognizing home environment sounds under air-conditioner noise for home automation. In [27], power normalized cepstral coefficients (PNCCs) [28] were shown to outperform MFCCs under various noise conditions and noise levels, including reverberant environments.

2.2.2 GTCCs

Gammatone cepstral coefficients (GTCCs) are a more recent addition to the family of cepstral features. GTCCs employ a gammatone filter, a linear filter which models the frequency selectivity property of the human cochlea. The most commonly used cochlea model is that proposed by Patterson et. al. [29] which is a series of bandpass filters with the bandwidth given by equivalent rectangular bandwidth (ERB). An efficient implementation of the gammatone filter bank has been provided in [30] which has been closely followed in speech [31] and non-speech [32] recognition applications.

Extraction of GTCCs follows the same procedure as MFCCs except that gammatone filters are used instead of mel-filters. The impulse response for the gammatone filter can be given as [29]

$$g(r) = Ar^{j-1}e^{-2\pi Br}\cos(2\pi f_c r + \phi) \tag{10}$$

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 [29]. ERB is a psychoacoustic measure of the auditory filter width at each point along the cochlea and can be given as

$$f_{c,ERB} = \left[\left(\frac{f_{c,Hz}}{Q_{ear}}\right)^p + (B_{min})^p\right]^{1/p} \tag{11}$$

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 [33] ($Q_{ear} = 9.26$, $B_{min} = 24.7$, and $p = 1$), Lyon's cochlea model as given in [34], ($Q_{ear} = 8$, $B_{min} = 125$, and $p = 2$), and Greenwood [35] ($Q_{ear} = 7.23$, $B_{min} = 22.85$, and $p = 1$).

The cochlea has thousands of hair cells which resonate at their characteristic frequency and at a certain bandwidth. In [30], the mapping between center frequency and cochlea position is

determined by integrating the reciprocal of (11) 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

$$f_{cm} = -Q_{ear}B_{min} + \left(f_h + Q_{ear}B_{min}\right)e^{-ms/Q_{ear}}, \quad m = 1,2,...,M_2. \tag{12}$$

where f_h is the maximum frequency in the filter bank, M_2 is the number of gammatone filters, and s is the step factor given as

$$s = \frac{Q_{ear}}{M_2}\log\left(\frac{f_h + Q_{ear}B_{\min}}{f_l + Q_{ear}B_{\min}}\right) \tag{13}$$

where f_l is the minimum frequency in the filter bank.

A 4th order gammatone filter with four filter stages and each stage a 2nd order digital filter is described in [30] and an implementation provided in the Auditory Toolbox for Matlab [36].

Valero and Alías [32] performed a detailed analysis on MFCCs and GTCCs and concluded that GTCCs are more effective than MFCCs in representing the spectral characteristics of non-speech audio signals, especially at low frequencies.

2.3 Time-Frequency Image Features

Every sound signal produces a unique texture which can be visualized using a spectrogram image. The intensity values of the spectrogram image represent the dominant frequency components against time. This can be utilized to improve the recognition rate of sounds in the presence of additive noise, provided the noise spectrum does not contain strong spectral peaks which corrupt the dominant sound signal components. This was demonstrated by Paliwal [37] where spectral subband centroids (SSCs) were used as supplementary features for improved robustness in speech recognition. Some similar works in ASR are discussed below.

2.3.1 Central Moments

For sound event recognition, Dennis et al. [9] extract central moments as features from the spectrogram image of sound signals which they refer as the spectrogram image feature (SIF). They consider both grayscale and quantized spectrogram image representations. To obtain the spectrogram image, the linear values are firstly obtained from the DFT values as

$$S_{Linear}(k,t) = \left|X(k,t)\right| \tag{14}$$

$$S_{Log}(k,t) = \log\left|X(k,t)\right|. \tag{15}$$

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

$$I(k,t) = \frac{S(k,t) - \min(S)}{\max(S) - \min(S)}. \tag{16}$$

These values are then quantized and mapped onto the red, green, and blue (RGB) monochrome components of a color space. The HSV color space was used in this case and the mapping of the grayscale image to the monochrome image can be given as

$$\eta_q(k,t) = h_c\left(I(k,t)\right) \qquad \forall q \in \left(q_1, q_2, ...q_N\right) \tag{17}$$

where η_q is a monochrome image (*R*, *G*, or *B*), h a nonlinear mapping function, and q the quantization regions.

Each image is then partitioned into 9×9 blocks and second and third central moments are computed in each block as features. The v^{th} central moment for any given block of image can be determined as

$$\mu_v = \frac{1}{Z} \sum_{z=1}^{Z} \left(I_z - \mu\right)^v \tag{18}$$

where Z is the sample size or the number of pixels in the block, I_z is the grayscale intensity value of the z^{th} sample in the block, and μ is the mean grayscale intensity value of the block.

For experimentation, 60 sound classes are taken from the Real World Computing Partnership (RWCP) Sound Scene database in Real Acoustic Environment [38] 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 are used for testing. Four noise types: *speech babble*, *destroyer control room*, *factory floor 1*, and *jet cockpit 1* from NOISEX-92 database [39] are added at 20dB, 10dB, and 0dB SNRs to test the robustness of the proposed approach. While MFCCs were seen to produce higher classification accuracy under clean conditions, the results using the SIF method were much better at low SNRs. The best results for training with clean signals were between 74-77% at 0dB SNR for the four noise types. This was achieved using the SIF method, quantized linear spectrogram, and HMM classification, implemented using the HTK toolkit [40]. For babble noise, with multi-conditional training, training with clean signals and at 20dB and 10dB SNRs, the accuracy at 0dB SNR increased from 74.4% to 79.4% which was using the quantized log spectrogram.

The SIF with reduced feature dimensions, referred as reduced SIF (RSIF), is proposed for an audio surveillance application in [8]. For the RSIF, the mean and standard deviation values of the central moment values are computed along the rows and columns of the blocks. These are concatenated to form the feature vector which is 2.25 times smaller than the SIF but without compromise in classification performance. Experimentations are carried out on 10 sound classes: *alarms*, *children voices*, *construction*, *dog barking*, *footsteps*, *glass breaking*, *gunshots*, *horn*, *machines*, and *phone rings*. Each class contains multiple subclasses and the sound files are largely obtained from the RWCP Sound Scene database in Real Acoustic Environment [38] and the BBC Sound Effects library [41]. The performance is evaluated under three different noise environments taken from the NOISEX-92 database [39]: *speech babble*, *factory floor 1*, and *destroyer control room*. The feature vector combination of RSIF and linear MFCCs produced significantly better results, especially at low SNRs, with a classification accuracy of 97.11%, 96.06%, 93.61%, 89.15%, and 70.95% under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs, respectively.

2.3.2 GLCM

Gray-level co-occurrence matrix (GLCM), also known as gray-tone spatial dependence matrix [42], is an image processing based texture analysis technique which has been extended to texture analysis of sound signal time-frequency images. GLCM gives the spatial relationship of pixels in an image. It 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

$$p(i,j)=\sum_{k=1}^{N_k}\sum_{t=1}^{N_t}\begin{cases}1, \text{ if } I(k,t)=i \ \& \ I(k+d_k,t+d_t)=j\\ 0, \text{ otherwise}\end{cases} \tag{19}$$

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.

After computing the GLCM, the fourteen textural descriptors proposed in [42], a subset of these descriptors, or new descriptors are often extracted as features. This approach has been utilized in a number of applications involving image texture analysis and some recent examples include diagnosis of abdominal tumors using texture classification of ultrasound images [43] and mammogram texture classification for breast cancer detection [44].

Costa et al. [45] applied this technique to texture classification of spectrogram images for music genre recognition. Their audio database consists of 900 music pieces from 10 music genres taken from the Latin music database [46]. The audio signal is first converted to a spectrogram using time decomposition [46] and the GLCM texture descriptors are extracted as features using a zoning technique, that is, the spectrogram image is divided into horizontal sections, with a total of 10 zones, and analysis is carried out in each zone. The following seven textural descriptors are utilized: entropy, correlation, homogeneity, third order momentum, maximum likelihood, contrast, and energy. The results are compared against those in [47] which takes an instance-based approach with feature vectors represented by short-term, low-level characteristics of the music audio signal. Only a marginal increase is seen in the average classification accuracy, increasing from 59.6% to 60.1%, but results showed an improvement of about 7% with a combination of the two methods.

In a face recognition problem [48], however, instead of extracting the textural descriptors as features, the matrix values itself form the feature vector. This was generally shown to give significantly better results than using the combined fourteen textural descriptors as features. This technique was adopted in an audio surveillance application in [49] which was referred as the spectrogram image texture feature (SITF). Analysis was performed independently in frequency subbands and the final feature vector is a concatenation of the feature vectors from each subband. When tested with the same databases and experimental setup as in [8], the SITF gave significantly better classification performance at low SNRs and a better overall performance when compared to MFCCs, SIF, and RSIF.

2.3.3 Other Time-Frequency Representations

Short-time Fourier transform (STFT) is probably the most commonly used time-frequency image representation. In this representation, the frequency components are equally spaced with constant bandwidth. However, most sound signals hold greater frequency components in the lower frequency range and less frequency components in the upper frequency range. As such, some information is lost in the lower frequency components while the higher frequency components hold little information in this time-frequency representation.

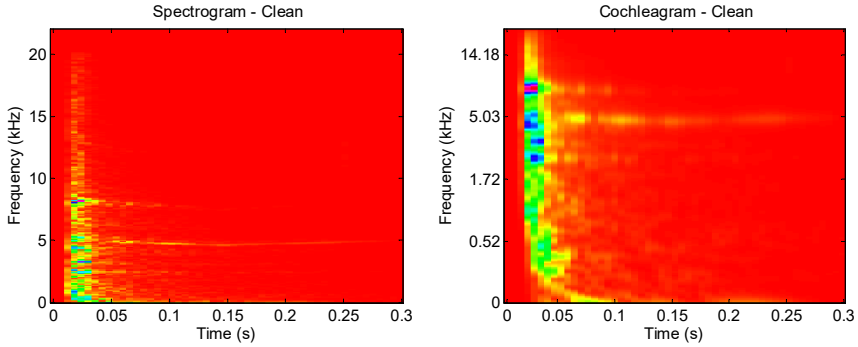
Mel-spectrogram and gammatone-spectrogram, also known as cochleagram, are variations of the STFT spectrogram utilizing the mel-filter and gammatone filter, respectively. Both these filters have more frequency components in the lower frequency range with smaller bandwidths and fewer frequency components in the higher frequency range with longer bandwidths. This makes their corresponding time-frequency representation more suitable for feature extraction. Cochleagram image-based feature extraction, in particular, has found usage in speech recognition [50] and audio separation [51] applications. This time-frequency representation was also used in an audio surveillance application in [52]. After filtering the

signal with the gammatone filter, the energy in the windowed signal for each frequency component is added which can be given as

$$C(m,t) = \sum_{n=0}^{N-1} \left|\hat{x}(m,n)\right| w(n), \qquad m = 1,2,...,M_2 \tag{20}$$

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 the t^{th} frame.

These values are then normalized using (16) to get the grayscale cochleagram image intensity values. Cochleagram image feature extraction improved the average classification performance of SIF, RSIF, and SITF, now referred as CIF, RCIF, and CITF, respectively, from 75.89%, 81.08%, and 85.62% to 86.30%, 89.03%, and 89.24% with significantly improved results at low SNRs. Further improvement in classification performance was also achieved when combined with linear GTCCs. Illustration of spectrogram and cochleagram image for a *construction* sound signal can be found in Figure 2 [52].



(a) (b)

Figure 2: Spectrogram and cochleagram images for a sample sound signal from *construction* sound class. (a) Spectrogram image and (b) cochleagram image.

Wavelet transform [53] also provides time-frequency representation of a signal and has an advantage over Fourier transform in that it provides better time and frequency localization. Nilufar et al. [54] use wavelet packet decomposition [55], an extension of wavelet transform that includes more signal filters, for robust *speech* and *music* discrimination. This technique is applied to the spectrogram to transform it into different subbands containing texture information. Multiple kernel learning (MKL) [56] is used to select the optimal subbands for discriminating the two classes.

2.4 Sparse Decomposition

Sparse decomposition aims to decompose a given input signal as a linear combination of a defined number of elementary signals from a large linearly dependent collection. While there are a few algorithms for this, such as basic pursuit (BP) [57], matching pursuit (MP) seems to be the most often used in ASR applications.

Chu et al. [10] consider MP in environmental sound recognition. Their sound database consists of fourteen environment types, taken from BBC sound effects library [41] and the Freesound project [58], which are as follows: *inside restaurants*, *playground*, *street with traffic and pedestrians*, *train passing*, *inside moving vehicles*, *inside casinos*, *street with police car siren*, *street with ambulance sire*n, *nature-daytime*, *nature-nighttime*, *ocean waves*, *running water/stream/river*, *raining/shower*, and *thundering*.

In simple terms, MP, originally proposed by Mallat and Zhang [59], allows extraction of time-frequency features through the sparse linear expansion of a waveform. This is done by

decomposing signals using an overcomplete dictionary of functions, such as Gabor dictionary [59] as used in their work. Other available dictionaries are wavelets [60], wavelet packets [61], multiscale Gabor dictionaries [62], and chirplets [63]. An overcomplete dictionary ensures that the signal converges to a solution with zero residual energy and, therefore, results in the best set of functions to approximate the original representation.

As explained in [10], given a dictionary D containing parameterized waveforms ϕ_γ as

$$D = \left\{ \phi_\gamma : \gamma \in \Gamma \right\} \tag{21}$$

where Γ is the parameter set and ϕ_γ is called an atom, the approximate decomposition of a signal x can then be given as

$$x = \sum_{i=1}^{a} \alpha_{\gamma_i} \phi_{\gamma_i} + R^{(a)}, \qquad i = 1,2,...,a \tag{22}$$

where $R^{(a)}$ is the residual. Given x, a, and D, the goal is to determine indices γ_i and compute ϕ_{γ_i} while minimizing $R^{(a)}$ with the initial approximation $x^{(0)} = 0$ and $R^{(0)} = a$. The sequence of sparse approximation is done stepwise using MP.

In [10], frequency and scale parameters are extracted from each atom as features together with the mean and standard deviation for each parameter, with five determined as the optimum number of atoms. A combination of MFCCs and MP features produced the highest classification accuracy at 83.9% using GMM classification.

Interestingly, they also gave a listening test to 18 individuals which produced an overall accuracy of 77%, 82%, and 85% for an audio clip of 2, 4, and 6 seconds, respectively. The confidence level of the individuals were also measured with each answer which showed direct correlation with the accuracy. Potential short falls in the listening test, such as short duration of clips, were discussed against the results in [64] where listening tests produced better results than the ASR system.

In addition, Scholler and Purwins [65] consider MP for signal approximation with sparse optimization method [66] for drum sound classification. Data samples from ENST database [67] and RWC Music Database: Musical Instrument database [68] are used and the following features are considered: MP features using a sparse coding dictionary (SC-MP), MP feature using a gammatone dictionary (GT-MP), and timbre descriptors (TD). Apart from the three individual feature sets, the combination of TD with SC-MP and GT-MP is also considered. Results are compared under clean conditions and at 20dB, 10dB, 0dB, and -10dB SNRs with the addition of white Gaussian noise. When trained with clean samples only, the overall performance of the MP features was much better than TD features. While all the features gave comparable results under clean conditions, MP features performed much better under noisy conditions, except at -10dB and 0dB SNRs, where all features gave poor results. For TD features, the addition of MP features and multi-conditional training improved the classification accuracy but the overall performance of the individual MP features was still better.

2.5 Feature Optimization

The addition of new features to a baseline feature set does not necessarily improve the classification accuracy as some features or feature dimensions are ineffective, depending of the application. Optimization techniques have been used in some literature to determine the optimal feature set. For musical instrument classification, Essid et al. [69] consider various features which can be broadly classified as: temporal, cepstral, spectral, amplitude

modulation (AM) [70], and octave band signal intensities (OBSI) [69]. Two feature selection methods, inertia ratio maximization using feature space projection (IRMFSP) [71] and genetic algorithms (GAs) [72, 73] are experimented with. Class pairwise feature selection with a fusion of the two optimization techniques was found to be most effective in discriminating between possible pair of instrument classes, giving a better overall performance as a result.

Alexandre et al. [5] argue the computational limitations of digital signal processing hardware in a hearing aid application. GA with restricted search [74] is proposed to select the optimal features so that the feature vector dimension could be reduced and the computation speed increased as a result. Three main classes are considered: *speech in quiet*, *speech in noise*, and *noise*. A two-layer structure is adopted for classification. The first layer distinguishes between *speech* and *non-speech* (*noise*) and the second layer classifies speech files into either a *quiet* environment or a *noisy* environment. They consider 38 features, mostly cepstral and time and frequency domain features, with the final feature vector 76-dimensional. The results show that while the unconstrained GA required 43 and 46 features to get the best probability of correct classification for the two classification problems, respectively, only 11 features are shown to give comparable performance using restricted GA, which is also always slightly better than the sequential methods [75], sequential forward search (SFS) and sequential backward search (SBS).

Chmulik and Jarina [76] experiment with particle swarm optimization (PSO) [77] and GA to select the optimum features from a collection of 20 audio feature descriptors, extracted using the YAAFE toolbox [78], for classification of six sound classes: *applause*, *crying*, *laughing*, *speech*, *music*, and *noise*. While comparable classification accuracy is achieved using both the optimization techniques, PSO gives the highest classification accuracy at 82.48% with a feature dimension of 86. This is much better than the classification accuracy with all the features included which is 72.94% with a feature dimension of 137. This shows that the inclusion of ineffective features not only increases the computation time but can also reduce the classification performance.

3.0 Classifiers

In this review, we focus on two relatively new classifiers, SVMs and DNNs. We provide a brief background on the two classifiers and discuss their classification performance.

3.1 Support Vector Machines

3.1.1 Binary SVM

SVM is a statistical learning classifier developed for binary classification. The initial SVM was a linear classifier proposed by Vapnik and Lerner in 1963 [79]. This was extended to nonlinear datasets by Boser, Guyon, and Vapnik in 1992 [80] and has gained widespread attention since the late '90s, around the same time research in ASR was generating interests. We provide a theoretical background of binary SVMs and then discuss the multiclass classification methods of SVMs. We then compare the performance of SVMs against other classification methods and also compare the performance of the multiclass classification methods.

3.1.1.1 Basic Theory

As described in [80-82], a support vector machine determines the optimal hyperplane to maximize the distance between any two given classes. Starting with a case of linearly separable dataset, consider a set of l training samples belonging to two classes 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 $\mathbf{x}_i$. The optimal hyperplane can be determined by minimizing $½\|\mathbf{w}\|^2$ subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$, where $\mathbf{w} \in R^d$ is a normal vector to the hyperplane and b is a constant. Solving this using classical Lagrangian duality gives the solution

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \tag{23}$$

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

$$b = y_i - \mathbf{w} \cdot \mathbf{x}_i \tag{24}$$

using any support vector or averaged over all support vectors.

For linearly nonseparable problems, the optimization can be generalized by introducing the concept of *soft margin* [81]. 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 solution is similar to the separable case except for a modification to the Lagrange multipliers: $0 \leq \alpha_i \leq C, i = 1, 2, \ldots l$, 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* [80] 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 $\phi(\mathbf{x}_j)$. A commonly used kernel function is Gaussian radial basis function (RBF) given as $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)$, where $\sigma > 0$ is the width of the Gaussian function.

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

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{25}$$

3.1.1.2 Multiclass 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 technique are one-against-all (OAA), one-against-one (OAO), decision directed acyclic graph (DDAG), and adaptive directed acyclic graph (ADAG).

OAA is probably the earliest of the multiclass SVM classification techniques [82]. For a P-class problem, P binary SVM classifiers are constructed and evaluated where 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 which can be given as

$$f(\mathbf{x}) = \arg\max_{i=1,2,\ldots,P} \left(\mathbf{w}^i \cdot \phi(\mathbf{x}) + b^i\right). \tag{26}$$

The OAO approach distinguishes between every pair of classes and classification is done using the max-wins voting strategy [83]. For a P-class problem, OAO-SVM constructs and evaluates $P(P-1)/2$ 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} classes. During classification, the class label of a test sample is predicted as

$$f(\mathbf{x}) = \arg\max_{i=1,2,\ldots,P} \sum_{j=1, j\neq i}^{P} \operatorname{sgn}\left(\mathbf{w}^{ij} \cdot \phi(\mathbf{x}) + b^{ij}\right). \tag{27}$$

DDAG [84] and ADAG [85] are also based on classification between pair of classes but utilize a decision tree structure in the testing phase. Similar to OAO-SVM, $P(P-1)/2$ nodes are created during training phase but only $P-1$ nodes are used during evaluation.

3.1.1.3 Comparison of SVM with Other Classifiers

Guo and Li [3] use the Muscle Fish database and similar features as in [2], that is, cepstral and perceptual features. However, a new metric called distance-from-boundary (DFB) is proposed for audio retrieval using SVMs to learn the boundaries. Using the same feature vector formation technique as [2], SVM performed better than NN, kNN, and NC (nearest center) classifiers. The lowest error rate is 11.00% for the leave-one-out test but 8.08% with separate training and test sets.

In another similar work, Lu et al. [86] consider five audio classes: *silence*, *music*, *background sound*, *pure speech*, and *non-pure speech*. SVM, with a Gaussian RBF kernel, is used for classification. For experimentation, a database with 2600 audio clips is created with a total duration of about 4 hours obtained from TV programs, internet, audio, and music CDs. When tested under different testing units (durations), in general, kNN classifier gave higher results than GMM while the SVM classifier always outperformed kNN and GMM classifiers.

In [87], multi-layer perceptron (MLP) neural network, trained using the Levenberg-Marquardt (LM) [88] back-propagation algorithm, and SVM, with a polynomial kernel, are experimented for classification for automatic ontology generation for musical instruments. The average classification accuracies for the MLP classifier were 76.0% and 46.7% for *solo music* and *isolated notes*, respectively, which increased to 83.0% and 86.3% with SVM classification.

In general, SVM has been seen to give a better classification performance than most traditional classifiers. In [9], HMM gives a better classification performance than SVM, however, they used linear SVM. Linear SVM is generally considered not suitable for complex classification tasks and most other similar works use nonlinear SVM. Results in [17] show that there isn't a significant difference between HMM and binary SVMs. In addition, it is important to tune the parameters C and σ correctly, avoiding local minima, for best results.

3.1.1.4 Comparison of Multiclass Classification Methods

There are various pattern recognition problems where multiclass SVM classification methods have been compared. Hsu and Lin [89] compare the performance of OAA, OAO, DDAG and two altogether methods, an approach for multiclass problems by solving a single optimization problem, on large classification problems. They conclude OAO and DDAG as being more suitable for practical use. A similar comparison is done by Seo [90] using OAA, OAO, DDAG together with the approach given by Weston and Watkins [91] and Crammer and Singer [92] for a face recognition application. While OAO was found to give marginally better results than DDAG, DDAG is suggested due to its low computational cost.

Various modifications to the multiclass SVM classification methods have also been proposed in literature. Liu and Zheng [93] introduce static reliability measures (SRM) and dynamic reliability measures (DRM) to improve the accuracy of OAA classification method. Kumar and Gopal [94] present reduced one-against-all (R-OAA), based on sample subset selection, which has a reduced computational time. Yang et al. [95] propose a partition based binary tree for the OAA method which has a faster training and evaluation time when compared to OAA and R-OAA. An optimized DAG method is proposed by Takahashi and Abe [96]. Another improvement to DAG, based on binary decision tree and Huffman code [97], is given by Chen and Liu [98]. Fei and Liu [99] propose a binary tree of support vector machine (BTS) which uses one-to-one training scheme but reduces the total number of classifiers by employing a probabilistic model. It gives a higher classification efficiency, that is, gives comparable classification accuracy to the traditional methods but is generally faster.

In general, the difference in the classification accuracy of the various multiclass SVM classification methods which have been proposed is marginal and the preference of one method over the others is largely based on faster training and/or evaluation times. However, most of these comparisons are limited to clean conditions only. In [8], the performance of OAA, OAO, DDAG, and ADAG multiclass SVM classification methods are compared under noisy conditions. The OAA classification method is seen to be the most noise robust and also gave a significantly better performance with feature vector combination. In addition, it required a significantly lower evaluation time than OAO but was slower than DDAG and ADAG methods.

3.1.1.5 Hybrid Classifier

Use of a hybrid SVM/kNN classifier using MPEG-7 audio descriptors can be found in [100]. MPEG-7, formally called “Multimedia Content Description Interface”, is a multimedia content description standard [101] which provides a comprehensive range of descriptors and descriptor schemes ranging from low level audio and video features to high level semantic features. Three MPEG-7 audio low-level descriptors, spectrum centroid, spectrum spread, and spectrum flatness, are used as features for classifying 12 sound classes: *male speech*, *female speech*, *cough*, *laughing*, *screaming*, *dog barking*, *cat meowing*, *frog wailing*, *piano*, *glass breaking*, *gun shooting*, and *knocking*. A frame-based classification strategy is used for the hybrid classifier where the output of the SVM classifier, using a RBF kernel and DDAG multiclass classification method, and kNN classifier are turned into probabilistic scores which are used to get a combined frame score. An unknown sound signal is then assigned to the class which most of the frames are assigned to. With 50% of the samples used for training and 50% for testing, a maximum classification accuracy of 85.1% is achieved using the hybrid classifier when compared to a maximum of 83.2% using HMM classification. However, the results are not compared to the standard SVM and kNN classification methods.

3.1.2 One-Class SVM

Unlike the binary SVMs considered so far, one-class SVM (1-SVM) is used in an audio surveillance application by Rabaoui et al. [17]. 1-SVM, proposed by Schölkopf et al. [102], is a modification of binary SVM to solve one-class classification problem. The feature is transformed by the kernel and the origin is treated as the second class. 1-SVM essentially separates the feature data points from the origin and maximizes the distance from this hyperplane to the origin.

1-SVM is more suited with high dimensional feature vectors. As such, unlike most other work where mean and standard deviation values of the extracted features across all frames are concatenated to form the feature vector, a slightly different approach to feature data

representation is taken in [17]. The overall feature data for the sound signal is divided into three portions: 30%, 40%, and 30% of the total number of frames. Mean value of the data across each dimension from each portion are concatenated to form the feature vector which results in a feature dimension which is 1.5 times longer than the conventional technique. The classification accuracy of 1-SVM was generally higher than HMM, OAA-SVM, and OAO-SVM classification methods when tested at various SNRs with a number of individual and combined features.

3.2 Deep Neural Networks

While SVMs have seen an increased usage in ASR systems, a new machine learning algorithm called deep learning is generating a lot of interest in speech recognition. Deep learning aims to learn high-level representations of data through a hierarchy of intermediate representations, such as deep neural networks (DNN) [103]. It has been used for acoustic modeling by research groups at University of Toronto, Microsoft Research, Google, and IBM Research, amongst others, and shown to outperform most other classification methods [104]. Furthermore, in two recent works in acoustics event classification [105, 106], DNN was shown to perform better than other classifiers. However, these works compare the classification performance using mel-frequency cepstral coefficients (MFCCs) only.

In another recent work in SER [107], the classification performance of SVM and DNN classifiers are compared with a number of features. These include MFCCs, features extracted from a stabilized auditory image (SAI) [108], and the SIF [9]. The DNN classifier utilized has *L*-layers with the feature vectors on the input layer and output layer in a one-of-*P* configuration (*P*-classes). The DNN is constructed using individual pre-trained restricted Boltzmann machine (RBM) pairs. Each pair comprises *V* visible and *H* hidden stochastic nodes, $\mathbf{v} = [v_1, v_2, \ldots, v_V]^T$ and $\mathbf{h} = [h_1, h_2, \ldots, h_H]^T$. Bernoulli-Bernoulli RBM structures are used on the intermediate and final layers while the input layer is Gaussian-Bernoulli RBM. Given the energy functions of the two RBM structures, $E(\mathbf{v}, \mathbf{h})$, the joint probability associated with configuration $(\mathbf{v}, \mathbf{h})$ is given as

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Y} e^{\{-E(\mathbf{v}, \mathbf{h}; \theta)\}} \tag{28}$$

where *Y* is a partition function given as $Y = \sum_v \sum_h e^{\{-E(\mathbf{v}, \mathbf{h}; \theta)\}}$.

The training data is used to estimate the RBM model parameters θ by maximum likelihood learning using the contrastive divergence (CD) algorithm [103]. After pre-training, a size *P* softmax output labeling layer is added to the pre-trained stack of RBMs [109] to convert a number of Bernoulli distributed units in the final layer into multinomial distribution. The stacked network is then trained using back propagation.

DNNs generally gave significantly higher overall classification accuracy than SVM with the best overall feature performance achieved using the SIF. With multi-conditional training, for example, the average classification accuracy using SVM is 88.55% but 92.58% with DNN. Similar conclusions were also drawn with MFCCs and SAI features.

4.0 Discussion and Conclusion

In this section, we first provide a summary of some of the key advancements in ASR. We then discuss some relatively lesser known applications of ASR.

4.1 Summary of Advancements

MFCCs have evolved as a baseline feature in many ASR systems. However, it is often supplemented with other features such as perceptual features [2, 3] and MP-based features

[110] for improved classification performance. A combination of linear cepstral coefficients and time-frequency image features has also shown to give robust performance in [8, 52]. While SVMs have been the preferred classifier in most ASR applications, DNNs have gained popularity in recent years with its superior classification performance, as demonstrated in a number of pattern recognition problems. In Table I, we summarize what we believe are some of the key works in ASR and highlight the advancements in features and classifiers using the work by Wold et al. [1] as basis.

While there has been a decent amount of development in ASR systems over these years, the inconsistency in the choice of sound databases in most literature makes it difficult to make direct comparison of the performance of the proposed techniques. While sound libraries, such as the Latin music database, RWCP databases, and the BBC sound effects library have been employed for research in certain ASR applications, the creation of the sound database for use from these available libraries is at the discretion of the researchers. Also, the number and complexity of sound classes and the amount of training data, amongst others, have a direct influence on the classification performance of an ASR system. Therefore, there is a need to standardize sound databases and experimental setups to make it easier for direct comparison of proposed techniques, similar to what was seen in [1-3], refer to Table 1.

In addition, different approaches have been noticed in structuring of classes in some similar applications, such as audio surveillance and sound event recognition. For an audio surveillance application in [8, 17], a sound class has a number of sound events. For example, shots fired from a rifle, shotgun, and machine gun are examples of different sound events but would be treated as a single sound class such as gunshots. In some cases, the signal properties of subclasses in a particular class are similar to the subclasses in other classes but different to subclasses in its own class. This creates interclass similarity and intraclass diversity, increasing the complexity of the problem as a result.

Moreover, there has been a lack of attention in recognition of overlapping sound events. An example of such a system is presented in [111] and, while there are some other similar works, more research is needed in this area.

Table I: An overview of some key works in ASR

Reference	Year	Application	Sound Database(s)	Noise Database(s)	No. of Classes	Total Files	% Training data	Best feature(s)	Classifier	Classification Accuracy (or Error Rate)
Wold et al. [1]	1996	Content-based audio classification	Muscle Fish	–	16	409	–	Perceptual features	NN	19.07% (ER)[1]
Li [2]	2000	Content-based audio classification	Muscle Fish	–	16	409	Leave-one-out test	**MFCC** + perceptual features	NFL	9.78% (ER)
							51.59% (211/409)			9.60% (ER)
Guo and Li [3]	2003	Content-based audio classification	Muscle Fish	–	16	409	Leave-one-out test	MFCC + perceptual features	**SVM**	11.00% (ER)
							51.59% (211/409)			8.08% (ER)
Rabaoui et al. [17]	2008	Audio surveillance	RWCP Sound Scene, Leonardo Software, hand recorded	NOISEX-92, hand recorded	9	1015	70%	(Multiple features[2])	**1-SVM**	Clean – 96.89% 20dB – 93.33% 10dB – 89.22% 5dB – 82.80% 0dB – 72.89%
Chu et al. [10]	2009	Environmental sound recognition	BBC Sound Effects, Freesound	–	14	–	75%	**MFCC + MP**	GMM	83.9%
Dennis et al. [9]	2011	Sound event recognition	RWCP Sound Scene	NOISEX-92	60	4800	62.5%	**SIF**	HMM	[3]Clean – 87.9% [3]20dB – 88.0% [3]10dB – 87.5% [3]0dB – 75.5%
Sharan and Moir [52]	2015	Audio surveillance	RWCP Sound Scene, BBC	NOISEX-92	10	1143	66.67%	**Linear GTCC + CITF**	SVM (OAA)	Clean – 96.59% 20dB – 96.59%

			Sound Effects							10dB – 95.36% 5dB – 94.23% 0dB – 83.73%
McLoughlin et al. [107]	2015	Sound event recognition	RWCP Sound Scene	NOISEX-92	50	4000	62.5%	SIF (De-Noised)	**DNN**	Clean – 96.20% 20dB – 95.80% 10dB – 94.13% 0dB – 85.47%

[1]ER as reported in [2, 3].
[2]Different combination of features are experimented with under clean and noisy conditions.
[3]Average classification accuracy value over four noise types.

4.2 Other Applications of ASR

The applications of ASR are not limited to content-based audio classification, audio surveillance, sound event recognition, and environmental sound recognition, which have been the focus of our review so far. A summary of some less conventional applications of ASR is given in Table II.

Similar to finger print recognition, face recognition, and, more recently, vein pattern recognition systems, heart sound recognition has the potential for human identification. Heart and lung sound recognition can also be used for diagnosis of disorders associated with the heart and lung, respectively. This extends to animals as well such as for identification of respiratory infections in pigs and dairy calves. Such technology can act as an early warning system which could help contain contagious viruses with some viruses from animals, such as swine flu, known to affect humans as well. Diagnosis of disorders using ASR technology extends beyond heart and lung sound recognition. An example of gastrointestinal motility monitoring system using bowel sounds, captured through abdominal surface vibrations, can be found in [112].

Table II: Some lesser known applications of ASR

Reference	Application	Description	Sound Database(s)[1]	Feature(s)	Classifier(s)
Beritelli and Spadaccini [113]	Biometrics	Heart sound recognition for human identification	–	MFCC + FSR[2] (first-to-second ratio)	Euclidean distance measure
Kwak and Kwon [114]	Biomedical	Heart sound classification for diagnosis of cardiac disorder	Heart Sounds and Murmurs [115]	MFCC	HMM, SVM
Lei et al. [116]		Breath sound classification for diagnosis of disorders associated with breathing	–	MFCC + perceptual features	SVM, ANN
Exadaktylos et al. [117]		Cough sound recognition in pigs	–	Power spectral density (PSD)	Euclidean distance measure
Dimoulas et al. [112]		Gastrointestinal motility monitoring using bowel sounds, captured through abdominal surface vibrations	–	Time and frequency domain features, wavelet analysis	ANN
Cai et al. [118]	Animal species recognition; sound classification; monitoring	Bird species recognition using bird calls	Backyards [119], Australian bird calls: subtropical east [120] and voices of subtropical rainforests [121], and recorded data	MFCC	ANN
Jaafar and Ramli [122]		Frog species recognition	–	MFCC	kNN
Brown and Smaragdis [123]		Northern resident killer whale sound classification	–	MFCC	HMM
Le-Qing [124]		Insect sound classification	United States department of agriculture [125]	MFCC	PNN [126]
Milone et al. [127]		Monitoring grazing behavior of cattle using ingestive sound classification	–	Spectral features	HMM

Aydin et al. [128]		Automatic measurement of feed intake of broiler chickens by detecting pecking sounds	–	PSD	(Adaptive threshold)
Yao et al. [129]	Context awareness	Context awareness for social activity recognition and recommendation using audio data gathered from mobile phone	–	MFCC, ZCR, SF, SC, BW	DTW [130]
Tong et al. [131]	Tile Inspection	Inspection of tile wall exfoliation through analysis of impact sound	–	PSD	ANN
Márquez-Molina et al. [132]	Aircraft classification	Aircraft classification using aircraft takeoff noise	–	MFCC, Octave analysis [133, 134]	ANN
Montazer et al. [135]		Helicopter type identification using rotor sound	–	Energy	RBFNN
Redel-Macías et al. [136]	Vehicle pass-by noise test	Identification of sound for pass-by noise test in vehicles	–	Spectral features	ANN
Tabacchi et al. [137]	Classification of cooking stages	Classification of cooking stages of boiling water using audio and vibration signals	–	MFCC	Parzen [138]

[1]Sound database provided only where known. Hand recorded signals were mostly used otherwise.
[2]Power ratio of the first and second heart sounds.

In addition, ASR can be used for animal species recognition through analysis of their call sound. Such a system can be used to carry out automatic animal species monitoring replacing the laborious manual recognition process. It could also be used for environmental monitoring since the abundance of wildlife would generally indicate a healthy environment and vice-versa. For example, researchers in Brisbane established a sensor network in the city's suburbs and forest park to study the impact of urbanization of neighboring suburbs on the ecological system, with the focus on recognition of bird species using acoustic signals [118].

Furthermore, while we have generally been looking at examples of standalone ASR systems so far, audio and video recognition systems could also be integrated for a more holistic approach to this problem, such as in the development of surveillance systems. Video surveillance systems have been around for many years but have limitations such as limited field of view and relatively expensive computation and data storage. An ASR system could be used to complement a video-based recognition system such as in public transports [139], detecting traffic events [140], fall detection [141], machine awareness [142], and in banks [143]. Audio and video recognition systems could also be combined for recognition of complex events in real movies [144]. Another such example is robotics. Robots are often aimed at mimicking human behavior and, similar to humans, acoustical information can be utilized to make a more complete description of the scene. Robotics based search and rescue operation is one such scenario where in the aftermath of a natural disaster, such as an earthquake, the injured could be behind collapsed structures and audio information such as screaming or crying could be used to reach them [145].

Author Biography

Roneel V Sharan received the B.E.Tech. and M.Sc. degrees in electrical and electronic engineering from the University of the South Pacific (USP), Fiji, in 2004 and 2007,

respectively. From 2004 to 2005, he was a graduate assistant at the School of Engineering at USP and an assistant lecturer from 2006 to 2013. He is currently pursuing Ph.D. degree at the Auckland University of Technology (AUT), New Zealand, under an AUT doctoral scholarship. His research interests include pattern recognition, image processing, and signal 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, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia,* vol. 3, no. 3, pp. 27-36, 1996.

[2] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Transactions on Speech and Audio Processing,* vol. 8, no. 5, pp. 619-625, 2000.

[3] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks,* vol. 14, no. 1, pp. 209-215, 2003.

[4] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing,* vol. 10, no. 7, pp. 504-516, 2002.

[5] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2249-2256, 2007.

[6] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing,* vol. 10, no. 5, pp. 293-302, 2002.

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

[8] 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.

[9] 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.
[10] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, no. 6, pp. 1142-1158, 2009.
[11] A. R. Abu-El-Quran, R. A. Goubran, and A. D. C. Chan, "Security monitoring using microphone arrays and audio classification," *IEEE Transactions on Instrumentation and Measurement,* vol. 55, no. 4, pp. 1025-1032, 2006.
[12] J. L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *IEEE Intelligent Transportation Systems Conference (ITSC '06)*, 2006, pp. 733-738.
[13] M. V. Ghiurcau, C. Rusu, R. C. Bilcu, and J. Astola, "Audio based solutions for detecting intruders in wild areas," *Signal Processing,* vol. 92, no. 3, pp. 829-840, 2012.
[14] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J. F. Serignat, "Information extraction from sound for medical telemonitoring," *IEEE Transactions on Information Technology in Biomedicine,* vol. 10, no. 2, pp. 264-274, 2006.
[15] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Features for Content-Based Audio Retrieval," in *Advances in Computers*. vol. 78, V. Z. Marvin, Ed. Elsevier, 2010, pp. 71-150.
[16] *Muscle Fish*. Available: http://www.musclefish.com
[17] 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.
[18] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Transactions on Neural Networks,* vol. 10, no. 2, pp. 439-443, 1999.
[19] D. O'Shaughnessy, *Speech communication: human and machine.* Addison-Wesley Pub. Co., 1987.
[20] C. Woo-Hyun, K. Seung-Il, K. Min-Seok, D. K. Han, and K. Hanseok, "Acoustic and visual signal based context awareness system for mobile application," *IEEE Transactions on Consumer Electronics,* vol. 57, no. 2, pp. 738-746, 2011.
[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu*, et al.*, *The HTK book (for HTK version 3.4).* Cambridge University: Engineering Department, 2009.
[22] X. Zhang and Y. Li, "Environmental sound recognition using double-level energy detection," *Journal of Signal and Information Processing,* vol. 4, no. 3B, pp. 19-24, 2013.
[23] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Communication,* vol. 12, no. 3, pp. 277-288, 1993.
[24] 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.
[25] 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.
[26] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," *IEEE Transactions on Automation Science and Engineering,* vol. 5, no. 1, pp. 25-31, 2008.

[27] B. Gao and W. L. Woo, "Wearable audio monitoring: Content-based processing methodology and implementation," *IEEE Transactions on Human-Machine Systems,* vol. 44, no. 2, pp. 222-233, 2014.
[28] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4101-4104.
[29] 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.
[30] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer, Technical Report 35, 1993.
[31] O. Cheng, W. Abdulla, and Z. Salcic, "Performance evaluation of front-end processing for speech recognition systems," The University of Auckland, New Zealand, Report 621, 2005.
[32] 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.
[33] 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.
[34] M. Slaney, "Lyon's Cochlear Model," Apple Computer, Technical Report 13, 1988.
[35] 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.
[36] M. Slaney, "Auditory Toolbox for Matlab," Interval Research Corproation, Technical Report 1998-010, 1998.
[37] 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.
[38] 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.
[39] 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.
[40] *HTK Toolkit*. Available: http://htk.eng.cam.ac.uk
[41] *BBC Sound Effects Library*. Available: http://www.leonardosoft.com
[42] 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.
[43] 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.
[44] 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.

[45] 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.
[46] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "The Latin music database," in *Proceedings of the 9th International Conference on Music Information Retrieval*, Philadelphia, PA, USA, 2008, pp. 451-456.
[47] M. Lopes, F. Gouyon, A. L. Koerich, and L. E. S. Oliveira, "Selection of training instances for music genre classification," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 4569-4572.
[48] 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.
[49] R. V. Sharan and T. J. Moir, "Robust audio surveillance using spectrogram image texture feature," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 1956-1960.
[50] 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.
[51] 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.
[52] R. V. Sharan and T. J. Moir, "Subband time-frequency image texture features for robust audio surveillance," *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 12, pp. 2605-2615, 2015.
[53] S. Mallat, *A wavelet tour of signal processing.* New York: Academic Press, 1999.
[54] S. Nilufar, N. Ray, M. K. I. Molla, and K. Hirose, "Spectrogram based features selection using multiple kernel learning for speech/music discrimination," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 501-504.
[55] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern Recognition Letters,* vol. 24, no. 9–10, pp. 1513-1521, June 2003.
[56] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research,* vol. 5, pp. 27-72, 2004.
[57] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing,* vol. 20, no. 1, pp. 33-61, 1998.
[58] *The Freesound Project*. Available: http://freesound.iua.upf.edu/index.php
[59] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing,* vol. 41, no. 12, pp. 3397-3415, 1993.
[60] P. Vera-Candeas, N. Ruiz-Reyes, M. Rosa-Zurera, D. Martinez-Munoz, and F. Lopez-Ferreras, "Transient modeling by matching pursuits with a wavelet dictionary for parametric audio coding," *IEEE Signal Processing Letters,* vol. 11, no. 3, pp. 349-352, 2004.
[61] G. Yang, Q. Zhang, and P.-W. Que, "Matching-pursuit-based adaptive wavelet-packet atomic decomposition applied in ultrasonic inspection," *Russian Journal of Nondestructive Testing,* vol. 43, no. 1, pp. 62-68, 2007.
[62] R. Gribonval, "Fast matching pursuit with a multiscale dictionary of Gaussian chirps," *IEEE Transactions on Signal Processing,* vol. 49, no. 5, pp. 994-1001, 2001.

[63] S. Ghofrani, D. C. McLernon, and A. Ayatollahi, "Comparing Gaussian and chirplet dictionaries for time-frequency analysis using matching pursuit decomposition," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology*, 2003, pp. 713-716.

[64] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa*, et al.*, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 1, pp. 321-329, 2006.

[65] S. Scholler and H. Purwins, "Sparse approximations for drum sound classification," *IEEE Journal of Selected Topics in Signal Processing,* vol. 5, no. 5, pp. 933-940, 2011.

[66] E. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation,* vol. 17, no. 1, pp. 19-45, 2005.

[67] O. Gillet and G. Richard, "ENST-Drums: An extensive audio-visual database for drum signals processing," in *Proceedings of 7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006, pp. 156-159.

[68] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proceedings of the 4th International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, 2003, pp. 229–230.

[69] S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1401-1412, 2006.

[70] A. Eronen, "Automatic musical instrument recognition," MSc Thesis, Tampere University of Technology, Tampere, Finland, 2001.

[71] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," in *Audio Engineering Society Convention 115*, New York, 2003.

[72] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* Ann Arbor, MI: University of Michigan Press, 1975.

[73] I. Fujinaga, "Machine recognition of timbre using steady-state tone of acoustic musical instruments," in *Proceedings of the International Computer Music Conference*, 1998, pp. 207-210.

[74] S. Salcedo-Sanz, G. Camps-Valls, F. Perez-Cruz, J. Sepulveda-Sanchis, and C. Bousono-Calzon, "Enhancing genetic feature selection through restricted search and Walsh analysis," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 34, no. 4, pp. 398-406, 2004.

[75] C. M. Bishop, *Neural networks for pattern recognition.* Oxford: Clarendon Press, 1995.

[76] M. Chmulik and R. Jarina, "Bio-inspired optimization of acoustic features for generic sound recognition," in *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2012, pp. 629-632.

[77] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, 1995, pp. 1942-1948.

[78] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an easy to use and efficient audio feature extraction software," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, 2010, pp. 441-446.

[79] V. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," *Automation and Remote Control,* vol. 24, no. 6, pp. 774-780, 1963.

[80] B. E. Boser, I. M. Guyon, and 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.
[81] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, Sep. 1995.
[82] V. N. Vapnik, *Statistical learning theory.* New York: Wiley, 1998.
[83] U. H. G. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 255-268.
[84] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems 12 (NIPS-99)*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge MA: MIT Press, 2000, pp. 547-553.
[85] B. Kijsirikul, N. Ussivakul, and S. Meknavin, "Adaptive directed acyclic graphs for multiclass classification," in *PRICAI 2002: Trends in Artificial Intelligence*. vol. 2417, M. Ishizuka and A. Sattar, Eds. Berlin Heidelberg: Springer, 2002, pp. 158-168.
[86] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems,* vol. 8, no. 6, pp. 482-492, 2003.
[87] S. Kolozali, M. Barthet, G. Fazekas, and M. Sandler, "Automatic ontology generation for musical instruments based on audio analysis," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, pp. 2207-2220, 2013.
[88] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks,* vol. 5, no. 6, pp. 989-993, 1994.
[89] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks,* vol. 13, no. 2, pp. 415-425, 2002.
[90] N. Seo, "A comparison of multi-class support vector machine methods for face recognition," The University of Maryland, Research Report, 6 Dec 2007.
[91] J. Weston and C. Watkins, "Multi-class support vector machines," Department of Computer Science, Royal Holloway, University of London, Egham, UK, Technical Report CSD-TR-98-04, 1998.
[92] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research,* vol. 2, pp. 265-292, 2001.
[93] Y. Liu and Y. F. Zheng, "One-against-all multi-class SVM classification using reliability measures," in *Proceedings of IEEE International Joint Conference on Neural Networks*, 2005, pp. 849-854.
[94] M. A. Kumar and M. Gopal, "Reduced one-against-all method for multiclass SVM classification," *Expert Systems with Applications,* vol. 38, no. 11, pp. 14238-14248, 2011.
[95] X. Yang, Q. Yu, L. He, and T. Guo, "The one-against-all partition based binary tree support vector machine algorithms for multi-class classification," *Neurocomputing,* vol. 113, pp. 1-7, 2013.
[96] F. Takahashi and S. Abe, "Optimizing directed acyclic graph support vector machines," in *Proceedings of Artificial Neural Networks in Pattern Recognition*, Florence, Italy, 2003, pp. 166-170.
[97] M. A. Weiss, *Data structures and algorithm analysis in C++.* 3rd ed. New York: Addison-Wesley, 2006.
[98] P. Chen and S. Liu, "An improved DAG-SVM for multi-class classification," in *Fifth International Conference on Natural Computation*, 2009, pp. 460-462.

[99] B. Fei and J. Liu, "Binary tree of SVM: a new fast multiclass training and classification algorithm," *IEEE Transactions on Neural Networks,* vol. 17, no. 3, pp. 696-704, 2006.
[100] J.-C. Wang, J.-F. Wang, K. W. He, and C.-S. Hsu, "Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor," in *International Joint Conference on Neural Networks (IJCNN '06)*, 2006, pp. 1731-1735.
[101] ISO/IEC, "Information technology - Multimedia content description interface - Part 4: Audio," ISO/IEC 15938-4, 2002.
[102] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Microsoft Research, Technical Report MSR-TR-99-87, 1999.
[103] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation,* vol. 18, no. 7, pp. 1527-1554, 2006.
[104] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly*, et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.
[105] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 506-510.
[106] Z. Kons and O. Toledo-Ronen, "Audio event classification using deep neural networks," in *INTERSPEECH*, 2013, pp. 1482-1486.
[107] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 3, pp. 540-552, 2015.
[108] T. C. Walters, "Auditory-based processing of communication sounds," Ph.D. dissertation, University of Cambridge, Cambridge, U.K., 2011.
[109] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," M.Sc. Thesis, Technical University of Denmark, Lyngby, Denmark, 2012.
[110] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition using MP-based features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, 2008, pp. 1-4.
[111] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley*, et al.*, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustical Society of America,* vol. 131, no. 6, pp. 4640-4650, 2012.
[112] C. Dimoulas, G. Kalliris, G. Papanikolaou, V. Petridis, and A. Kalampakas, "Bowel-sound pattern analysis using wavelets and neural networks with application to long-term, unsupervised, gastrointestinal motility monitoring," *Expert Systems with Applications,* vol. 34, no. 1, pp. 26-41, 2008.
[113] F. Beritelli and A. Spadaccini, "Human identity verification based on Mel frequency analysis of digital heart sounds," in *16th International Conference on Digital Signal Processing*, 2009, pp. 1-5.
[114] C. Kwak and O. W. Kwon, "Cardiac disorder classification by heart sound signals using murmur likelihood and hidden markov model state likelihood," *IET Signal Processing,* vol. 6, no. 4, pp. 326-334, 2012.
[115] D. Mason, *Listening to the heart: A comprehensive collection of heart sounds and murmurs.* 2nd ed. Philadelphia: F. A. Davis Company, 2000.
[116] B. Lei, S. A. Rahman, and I. Song, "Content-based classification of breath sound with enhanced features," *Neurocomputing,* vol. 141, pp. 139-147, 2014.

[117] V. Exadaktylos, M. Silva, J. M. Aerts, C. J. Taylor, and D. Berckmans, "Real-time recognition of sick pig cough sounds," *Computers and Electronics in Agriculture,* vol. 63, no. 2, pp. 207-214, 2008.
[118] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, 2007, pp. 293-298.
[119] F. V. Gessel. Top 40 Bird Songs [Online]. Available: http://www.birdsinbackyards.net
[120] D. Stewart, "Australian bird calls: Subtropical east," *CD, Nature Sound,* 2002.
[121] D. Stewart, "Voices of subtropical rainforests," *CD, Nature Sound,* 2002.
[122] H. Jaafar and D. A. Ramli, "Automatic syllables segmentation for frog identification system," in *2013 IEEE 9th International Colloquium on Signal Processing and its Applications (CSPA)*, 2013, pp. 224-228.
[123] J. C. Brown and P. Smaragdis, "Hidden Markov and Gaussian mixture models for automatic call classification," *Journal of the Acoustical Society of America,* vol. 125, no. 6, pp. EL221-EL224, 2009.
[124] Z. Le-Qing, "Insect sound recognition based on MFCC and PNN," in *2011 International Conference on Multimedia and Signal Processing (CMSP)*, 2011, pp. 42-46.
[125] R. Mankin. *Sound Library*. Available: http://www.ars.usda.gov
[126] D. F. Specht, "Probabilistic neural networks," *Neural Networks,* vol. 3, pp. 109-118, 1990.
[127] D. H. Milone, J. R. Galli, C. A. Cangiano, H. L. Rufiner, and E. A. Laca, "Automatic recognition of ingestive sounds of cattle based on hidden Markov models," *Computers and Electronics in Agriculture,* vol. 87, pp. 51-55, 2012.
[128] A. Aydin, C. Bahr, S. Viazzi, V. Exadaktylos, J. Buyse, and D. Berckmans, "A novel method to automatically measure the feed intake of broiler chickens by sound technology," *Computers and Electronics in Agriculture,* vol. 101, pp. 17-23, 2014.
[129] Y. Yao, G. Bin, Y. Zhiwen, and H. Huilei, "Social activity recognition and recommendation based on mobile sound sensing," in *2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing (UIC/ATC)*, 2013, pp. 103-110.
[130] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, no. 1, pp. 43-49, 1978.
[131] F. Tong, X.-M. Xu, S. K. Tso, and K. P. Liu, "Application of evolutionary neural network in impact acoustics based nondestructive inspection of tile-wall," in *Proceedings of International Conference on Communications, Circuits and Systems*, 2005, pp. 974-978.
[132] M. Márquez-Molina, L. P. Sánchez-Fernández, S. Suárez-Guerra, and L. A. Sánchez-Pérez, "Aircraft take-off noises classification based on human auditory's matched features extraction," *Applied Acoustics,* vol. 84, pp. 83-90, Oct. 2014.
[133] "IEC 1260: Electroacoustics - Octave-band and fractional-octave-band filters," *International Electrotech Commission,* 1995.
[134] "ANSI Standard S1.11-2004: Specification for octave-band and fractional-octave-band analog and digital filters," *American National Standards Institute,* 2004.
[135] G. A. Montazer, R. Sabzevari, and H. G. Khatir, "Improvement of learning algorithms for RBF neural networks in a helicopter sound identification system," *Neurocomputing,* vol. 71, no. 1–3, pp. 167-173, 2007.

[136] M. D. Redel-Macías, F. Fernández-Navarro, P. A. Gutiérrez, A. J. Cubero-Atienza, and C. Hervás-Martínez, "Ensembles of evolutionary product unit or RBF neural networks for the identification of sound for pass-by noise test in vehicles," *Neurocomputing,* vol. 109, pp. 56-65, 2013.
[137] M. Tabacchi, C. Asensio, I. Pavón, M. Recuero, J. Mir, and M. C. Artal, "A statistical pattern recognition approach for the classification of cooking stages. The boiling water case," *Applied Acoustics,* vol. 74, no. 8, pp. 1022-1032, 2013.
[138] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics,* vol. 33, no. 3, pp. 1065-1076, 1962.
[139] V. T. Vu, F. Bremond, G. Davini, M. Thonnat, P. Quoc-Cuong, N. Allezard*, et al.*, "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.
[140] K. Lopatka, J. Kotus, M. Szczodrak, P. Marcinkowski, A. Korzeniewski, and A. Czyzewski, "Multimodal audio-visual recognition of traffic events," in *22nd International Workshop on Database and Expert Systems Applications (DEXA)*, 2011, pp. 376-380.
[141] S. K. Tasoulis, C. N. Doukas, V. P. Plagianakos, and I. Maglogiannis, "Statistical data mining of streaming motion data for activity and fall recognition in assistive environments," *Neurocomputing,* vol. 107, pp. 87-96, 2013.
[142] J. Wang, K. Zhang, K. Madani, and C. Sabourin, "Salient environmental sound detection framework for machine awareness," *Neurocomputing,* vol. 152, pp. 444-454, 2015.
[143] J. Kotus, K. Lopatka, A. Czyżewski, and G. Bogdanis, "Audio-visual surveillance system for application in bank operating room," in *Multimedia Communications, Services and Security*. vol. 368, A. Dziech and A. Czyżewski, Eds. Springer Berlin Heidelberg, 2013, pp. 107-120.
[144] J.-X. Du, C.-M. Zhai, Y.-L. Guo, Y.-Y. Tang, and C. L. P. Chen, "Recognizing complex events in real movies by combining audio and video features," *Neurocomputing,* vol. 137, pp. 89-95, 2014.
[145] Q. Zhang, F.-Q. Zhao, Z.-J. Liu, and P. Yang, "Audio sensors fusion based on vote for robot navigation," in *25th Chinese Control and Decision Conference (CCDC)*, 2013, pp. 3219-3222.