Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 2895–2907

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Comparison of techniques for environmental sound recognition

Michael Cowling *, Renate Sitte [1]

School of Information Technology, Griffith University, Gold Coast Campus, PMB 50, Gold Coast Mail Centre, Queensland 9726, Australia

Received 14 January 2003; received in revised form 21 May 2003

Abstract

This paper presents a comprehensive comparative study of artificial neural networks, learning vector quantization and dynamic time warping classification techniques combined with stationary/non-stationary feature extraction for environmental sound recognition. Results show 70% recognition using mel frequency cepstral coefficients or continuous wavelet transform with dynamic time warping.

Keywords: Non-speech sound recognition; Environmental sound recognition; Audio signal processing; Acoustic signal processing; Joint time-frequency feature extraction

1. Introduction

In this paper we present the results of a comprehensive comparative study on different techniques that are typically used in speech/speaker recognition and musical instrument recognition, with the purpose to study their suitability for environmental sound identification. We found that the techniques traditionally known as best performers in the speech/speaker or in the musical instrument recognition scenario, are either not suitable, or not as good performers for the neglected field of environmental sound recognition, which has very few publications (e.g. Goldhor, 1993; Liu, 1999; Sampan, 1997). In particular, non-stationary Continuous Wavelet Transform together with Dynamic Time Warping were the clear winners for our purpose.

In previous work, the authors (Cowling and Sitte, 2000, 2001, 2002a,b) investigated the use of stationary feature extraction techniques for environmental sound recognition. Although the results from this comparative study were promising, the application of stationary feature extraction techniques to non-speech sounds is not ideal, since most environmental sounds are by their nature non-stationary. However, emerging research by Orr et al. (2001), among others, in the field of speech and speaker recognition has demonstrated that non-stationary (time-frequency) techniques

* Corresponding author. Tel.: +61-7-5552-8503.
E-mail addresses: m.cowling@mailbox.gu.edu.au (M. Cowling), r.sitte@mailbox.gu.edu.au (R. Sitte).
[1] Tel.: +61-7-5552-8203.

0167-8655/$ - see front matter
doi:10.1016/S0167-8655(03)00147-8

can be applied to sound and can produce good results. Therefore, we also tested these techniques (Cowling and Sitte, 2002a,b) for their suitability to environmental sound recognition.

This paper discusses the use of both stationary frequency-based techniques and non-stationary time-frequency-based feature extraction techniques as a means of classifying environmental sounds. The techniques are analysed and results are presented for testing them in combination with several common classification techniques (traditionally used in speech or in musical instrument recognition).

This paper contributes to research in the field of audio signal processing, specifically in the area of computational auditory scene analysis (CASA). The development of an environmental sound recognition system, in turn, contributes to the development of intelligent machines that are capable of understanding sound. Our immediate application is as a core component in a security system. In surveillance systems, sound systems have the advantage over video surveillance camera only systems, because line-of-sight is not an issue with these systems. Additionally, these systems can work hand in hand with video-based systems, when the direction of the camera is triggered by the type of sound heard.

The remainder of this paper is split into four sections. Section 2 discusses how the feature extraction techniques and classification techniques were selected. Specifically, it highlights those techniques that can (or cannot) be used for non-speech sound recognition. Section 3 discusses the specific implementation of these techniques. Section 4 presents the results of the comparative study on these techniques, and assesses their suitability for surveillance purposes. Finally, Section 5 concludes the paper and suggests areas of future research in the field of environmental sound recognition.

2. Sound analysis technique selection

This section analyses a number of techniques for their suitability to environmental sound recognition. Sound recognition (be it speech or environmental) is generally done in two phases: first feature extraction, followed by classification (using artificial intelligence techniques). This section discusses techniques in both these phases. Feature extraction is where a sound is manipulated in order to produce a set of characteristic features for that sound. For instance, a sound could be considered a high-pitched sound, or a low-pitched sound. Classification is then used to recognize the sound by cataloging the features of existing sounds in some way (training) and then comparing the test sound to this database of features (testing).

Feature extraction can be split into two broad types: stationary (frequency-based) feature extraction and non-stationary (time-frequency based) feature extraction. Stationary feature extraction produces an overall result detailing the frequencies contained in the entire signal. With stationary feature extraction, no distinction is made on where these frequencies occurred in the signal. In contrast, non-stationary feature extraction splits the signal up into discrete time units. This allows frequency to be identified as occurring in a particular area of the signal, aiding understanding of the signal.

2.1. Feature extraction (stationary)

For stationary feature extraction, speech, and musical instrument recognition relies on only a few different types of feature extraction technique (each with several different variations). Initially, we considered eight popular techniques (two of which are commonly used in musical instrument recognition and all of which are commonly used in speech recognition) as possible candidates for feature extraction of non-speech sounds. These were:

- Frequency extraction (music and speech)
- Homomorphic cepstral coefficients
- Mel frequency cepstral coefficients (music and speech)
- Linear prediction cepstral (LPC) coefficients
- Mel frequency LPC coefficients
- Bark frequency cepstral coefficients
- Bark frequency LPC coefficients
- Perceptual linear prediction (PLP) features

It should be noted that while frequency extraction is a stationary technique, other techniques using cepstral coefficients could be considered "pseudo-stationary" techniques because they split the signal into time-slices. These are not true time-frequency techniques, because each time-slice has to be taken in context with other time-slices in order to produce relevant information.

Techniques based on LPC coefficients were based on the idea of a *vocoder*, which is a simulation of the human vocal tract. Since the human vocal tract does not produce environmental sounds, these techniques typically seem to highlight non-unique features in the sound and are therefore not appropriate for recognition of non-speech sounds.

According to Lilly (2000), the results of the mel frequency-based filter and the bark frequency filter are similar, mainly due to the similar nature of these filters. Gold et al. (2000) also mentions that PLP and mel frequency are similar techniques. Based on these previous findings, we selected only the more popular mel frequency technique for testing.

2.2. Feature extraction (non-stationary)

The main time-frequency techniques that are commonly mentioned in the general literature (e.g. Cohen, 1995; Hubbard, 1995) are:

- Short-time Fourier transform (STFT)
- Fast (discrete) wavelet transform (FWT)
- Continuous wavelet transform (CWT)
- Wigner-Ville distribution (WVD)

All of these techniques use different algorithms to produce a time-frequency representation of a signal. While STFT uses a standard Fourier transform over several windows, Wavelet-based techniques apply a mother wavelet to a waveform to surmount the resolution issues inherent in STFT. WVD is a bilinear time-frequency distribution that also uses advanced techniques to try and combat these resolution difficulties. It has higher resolution than the STFT, but suffers from crossterm interference and produces results with coarser granularity than wavelet techniques (Hubbard, 1995). Of the two wavelet techniques, FWT is usually used for encoding and decoding of signals, while CWT is used for recognition tasks.

Despite its common usage for speech/sound coding, the FWT could be used successfully for recognition tasks, so it should be included in our comparative study. However, early tests on the Wigner-Ville distribution showed a transformation duration in excess of five minutes for signals of the length typical of environmental sounds. Given our intention to develop our system into a real-time surveillance system, this excessive duration was deemed unacceptable.

Based on these findings, three techniques (STFT, FWT, CWT) should be tested for their ability to classify non-speech sounds.

2.3. Classification

After feature extraction, a classification technique is used to catalogue the features. Test features can then be compared to this database.

The following classification techniques are commonly used for speech/speaker recognition or have, in the past, been used for this application domain. They are:

- Dynamic time warping (DTW)
- Hidden Markov models (HMM)
- Learning vector quantization (LVQ)
- Self-organising maps (SOM)
- Ergodic-HMM's
- Artificial neural networks (ANN)
- Long-term statistics

In addition to these techniques, we also highlighted three techniques commonly used on realistic recordings in musical instrument recognition (not just isolated tones):

- Maximum likelihood estimation (MLE)
- Gaussian mixture models (GMM)
- Support vector machines (SVM)

To aid in selection of techniques, comparison tables were built (using Gold et al. (2000), Lee et al. (1996a,b), Rodman (1999) as a base) to compare the different feature extraction and classification methods used by each of these techniques.

The comparison tables showed that some of these techniques, by their very nature, cannot be used for non-speech sound recognition. Any of the techniques that use subword features are not suitable for non-speech sound identification. This is because environmental sounds lack the phonetic structure that speech does. There is no set "alphabet" that certain slices of non-speech sound can be split into, and therefore subword features (and the related techniques) cannot be used.

Due to the lack of an environmental sound alphabet, the HMM-based techniques mentioned will be difficult to implement. However, this technique may be revisited in the future if necessary, and if a meaningful way of developing sound subunits can be devised. However, this is beyond the purpose of this research.

The SOM and LVQ techniques are complementary to each other. Kohonen developed both techniques, with specific applications intended for each technique. For classification, Kohonen (1997) suggests the use of the LVQ technique over the SOM technique. Therefore, LVQ will be the technique tested.

Long-term statistics cannot be applied in combination with non-stationary feature extraction techniques. Therefore, this classification technique will only be tested on its own feature extraction techniques.

Finally, all of the techniques used for musical instrument recognition work on a similar paradigm, that of unsupervised classification. For efficiency, we selected the most widely used of these techniques for testing, GMM's.

3. Comparison experiment

This section discusses the methodology used in our comparison of techniques. It includes the description of the experiment setup, the comparative study method and the implementation details. All calculations were done using Matlab 6 on a Pentium 4 1.6 GHz Desktop machine with 528MB of RAM.

3.1. Experiment setup

For the experiment setup, sound recording was conducted under quiet conditions. Dual Condenser Microphones were used to record to Sony Minidisc using the maximum sampling rate of 44,100 Hz, with 16 bits per sample. It should be noted that Sony Minidisc uses the lossy AATRAC3 compression format, but we do not expect the application of the lossy compression used in AATRAC3 to unduly effect our recognition process.

The experiment consists of tests on eight sounds, each with six different samples. The sounds used for this test are listed in Table 1 and are some typical sounds that would be classified in a sound surveillance system.

The techniques are tested using a jackknife method, identical to the method used by Goldhor (1993). A jackknife testing procedure involves training a classification system with all data except the sound sample that will be tested. This sound is then tested against the classification system and the classification is recorded. In cases where the setting of initial weights may affect the classification result (as is the case with LVQ and ANN techniques), training is repeated five times, with different weight initializations each time. A correct classification is only recorded if more than three of the training runs are correct. This jackknife procedure is repeated with all six of the samples from each of the eight sounds.

3.2. Comparative study method

The feature extraction and classification techniques shown in the comparison are tested for their ability to classify non-speech sounds in two ways. First, testing is performed, using these techniques, on non-speech sounds and data on the parameters, the resulting time taken and the final correct classification rate will be recorded. Then, these results are compared with statistics and previous results for the performance of the classi-

Table 1
Set of sounds used in the experiments

Sound type			
Jangling keys	Footsteps (close)	Footsteps (distant)	Wood snapping
Coins dropping	Footsteps on leaves	Footsteps on glass	Glass breaking

fication techniques with speech recognition and with musical instrument recognition. This demonstrates how these techniques perform compared against each other and provide an evaluation to the results for non-speech.

Moreover, since feature extraction and classification are both required to recognize a sound, each classification technique must also be tested against each feature extraction technique to determine the best combination of these two techniques. The exception to this is the long-term statistics technique, which generates its own features and therefore requires no feature extraction techniques.

Based on the above and on the selections made in Section 2, this produces a set of experiments summarized in Table 2.

3.3. Feature extraction techniques—stationary

In this comparison, we tested three stationary feature extraction techniques, whose implementation is discussed in this section.

3.3.1. Frequency extraction

Frequency extraction was performed using the fast Fourier transform (FFT) routine, which uses the following equation for a DFT:

$$X(k) = \sum_{j=0}^{N-1} x(j)\omega_N^{jk} \quad k = 0, \ldots, N-1 \tag{1}$$

where $\omega_N = e^{-i^{2\pi}/N}$ and is the frequency we wish to check for, j counts all the samples in the signal and N is the length of the signal being tested. The results of the FFT were then windowed into a set number of bands, each with a constant length. The mean signal power of each band was then taken to produce a reduced FFT feature, with a single value for each band. This FFT feature was then used as input to train the classification system. Empirical testing revealed that splitting the frequency signal into 256 bands produced good results. Since non-speech sound covers a wider frequency range than speech (anywhere from 0 to 20,050 Hz, the approximate limit of human hearing), a 44,100 point FFT ($N = 44,100$) was performed, to allow a greater frequency resolution across all the frequencies required.

3.3.2. Mel frequency cepstral coefficients

We used the MFCC algorithm from the Auditory Toolbox by Malcolm Slaney of Interval Research Corporation (1998). This toolbox is in wide use in the research community. The toolbox applies three steps to produce the MFCC. First, it splits the signal into sections (determined by the number of coefficients, which in this implementation is 13) and applies a Hamming window using the standard Hamming window equation:

$$h(k) = 0.54 - 0.46\cos\left(\frac{2\pi k}{N-1}\right) \quad k = 1, \ldots, N \tag{2}$$

where N represents the length of the subset of the signal which is being windowed. A mel frequency filterbank is then applied to each windowed segment. The mel frequency filterbank m is built using a logarithmic frequency mapping expressed by the following relation:

$$m = \frac{1000\ln\left(1+\frac{f}{700}\right)}{\ln\left(1+\frac{1000}{700}\right)} \approx 1127\ln\left(1+\frac{f}{700}\right) \tag{3}$$

where f represents the range of frequencies in the signal. The application of this filterbank produces a series of magnitude values (one for each filter). A cepstral coefficient formula (shown in the next

Table 2
Combination of feature extraction/classification techniques

	LTS	FE	MFCC	HCC	STFT	FWT	CWT
Learning vector quantization		✓	✓	✓	✓	✓	✓
Artificial neural networks		✓	✓	✓	✓	✓	✓
Dynamic time warping		✓	✓	✓	✓	✓	✓
Gaussian mixture models		✓	✓	✓	✓	✓	✓
Long-term statistics	✓						

section) is then used to perform a frequency warping using these magnitude values to produce MFCC and these features are then collected into a single feature vector, which is more appropriate for training a network. Special attention was paid to removing the first scalar within the vector, which represents the total signal power and is therefore too sensitive to the amplitude of the signal (as suggested by Lilly (2000) and Gold et al. (2000)).

3.3.3. Homomorphic cepstral coefficients

Our implementation of the homomorphic cepstral coefficient (HCC) algorithm was based on the MFCC algorithm from the Auditory Toolbox by Malcolm Slaney of Interval Research Corporation (1998). This algorithm was modified to produce HCC as opposed to MFCC by removing convolution with the mel frequency filterbank.

To apply this method, we first split the signal using hamming windows. We then calculate the cepstrum ($X(n)$) for each of the windowed segments. The cepstrum is the Fourier transform of the log magnitude spectrum. Once we have done this, we can calculate cepstral coefficients using the following relation:

$$y(k) = w(k)\sum_{n=1}^{N} X(n)\cos\frac{\pi(2n-1)(k-1)}{2N} \quad k = 1,\ldots,N \tag{4}$$

where

$$w(k) = \begin{cases} \sqrt{1/N}, & k = 1 \\ \sqrt{2/N}, & 2 \leqslant k \leqslant N \end{cases}$$

and n is the length of the windowed segment being manipulated. We selected the first 13 coefficients produced by this relation. These features were then used in a vector notation, which is more appropriate for training a network. As with the MFCC, special attention was paid to removing the first scalar within the vector, which represents the total signal power and is therefore too sensitive to the amplitude of the signal (as suggested by Lilly (2000) and Gold et al. (2000)).

3.4. Feature extraction implementation—non-stationary

This section explains and discusses the implementation details of the three feature extraction techniques that we tested in our comparative study.

In the case of STFT and CWT, a principal component analysis (PCA) was used after feature extraction to reduce the dimensionality of the resulting signal. An adaptive algorithm was used to calculate the maximum number of principal components required for the training data used (based on the energy in each dimension and a variable threshold). In both cases, a threshold value of 1% was found to produce the most accurate results. This process reduced the size of the signal significantly. For the STFT, it reduced the size of the matrix from 128×67 (or 8643 features) to 18×67 (or 1206 features). For CWT, it reduced the size of the matrix from 8820×55 (or 485,100 features) to 12×55 (or 660 features).

3.4.1. Short-time Fourier transform

A STFT was implemented using Matlab's FFT routine and a rectangular windowing algorithm. This approach allowed finer control over the resultant resolution of the STFT by allowing us to systematically change the number of samples in both time and frequency. The signal was windowed and then a FFT was calculated for each windowed segment (Cohen, 1995). This produces the following relation for the calculation of a STFT:

$$S_{\mathrm{t}}(\omega) = \frac{1}{\sqrt{2\pi}}\int \mathrm{e}^{-\mathrm{j}\omega t}s(\tau)h(\tau - t)\,\mathrm{d}\tau \tag{5}$$

where ω is the frequency, τ is the signal length, $s(t)$ is the signal and $h(t)$ is the windowing function. This algorithm was implemented in Matlab with a variable window size parameter (allowing the resolution of the STFT to focus more closely on either time data or frequency data). Empirical testing showed that a window size of the sample frequency scaled by 100 produced the most accurate results when tested.

3.4.2. Fast wavelet transform

For the fast wavelet transform (FWT), we used the periodized, orthogonal *FWT_PO* algorithm from the Matlab Wavelab toolbox by Stanford University (Donoho et al., 2002). Like all FWT algorithms, this algorithm convolves the signal

with a filter and then applies a subsampling relation:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k) \cdot x(2n-k) \tag{6}$$

This subsampling equation is then repeated on the lower half of the signal (and optionally the high half of the signal), such that:

$$y_{\text{high}}(k) = \sum_{n} x(n) \cdot g(2k-n) \tag{7}$$

$$y_{\text{low}}(k) = \sum_{n} x(n) \cdot h(2k-n) \tag{8}$$

As a filter (i.e. $h(t)$ and $g(t)$), we applied the popular Daubechies filters (Daubechies, 1992) to the signal. Dabauchies filters allow for the perfect reconstruction of a signal from the FWT. A vanishing moment variable can be set for these filters upon generation; however, the value of this coefficient seemed to make little difference to the classification rate. Due to the nature of the FWT, the signal requires no PCA to reduce its dimensionality, meaning the result of the FWT can be used directly in the classification system.

3.4.3. Continuous wavelet transform

For the CWT, we used the discretized CWT algorithm from Stanford Unviersity's Matlab Wavelab toolbox (Donoho et al., 2002):

$$\begin{aligned}\text{CWT}_x^{\psi}(\tau,s) &= \Psi_x^{\psi}(\tau,s) \\ &= \frac{1}{\sqrt{|s|}} \sum_{t=1}^{N} x(t)\psi^*\left(\frac{t-\tau}{s}\right)\end{aligned} \tag{9}$$

where τ represents translation, s represents scale and $\psi(t)$ is the mother wavelet, which was chosen to be the Morlet mother wavelet (Daubechies, 1992), defined as

$$\psi(t) = \mathrm{e}^{jat}\mathrm{e}^{\frac{-t^2}{2s}} \tag{10}$$

where a is a modulation parameter and s again represents scale. This mother wavelet has been used for recognition tasks in the past and produced acceptable results (Orr et al., 2001).

3.5. Classification

Four classification techniques will be tested in this comparison. The implementation of each of these techniques will be discussed in this section.

3.5.1. Learning vector quantization

The LVQ was implemented using the inbuilt LVQ routines in Matlab's neural network toolbox. The network was initialized with 50 competitive neurons and a learning rate of 0.05. These settings allowed the network to converge in ~50 iterations and were found to give the most accurate classification rate.

3.5.2. Artificial neural networks

The ANN was implemented using the fast back propagation algorithm (BPA) in the Matlab neural network toolbox. We used the Levenberg–Marquardt back propagation algorithm and tansig activation functions. The network was initialized with 50 hidden neurons and a learning rate of 0.05. The limit of the sum-squared error was set to 0.001 and the momentum constant was set to 0.95. These settings allowed the network to converge in ~500 iterations.

3.5.3. Dynamic time warping

We implemented DTW using the DTW function in the Auditory Toolbox developed by Malcolm Slaney (1998). DTW uses a dynamic programming approach (as opposed to a linear approach) to align the length of the signal with the length of the reference signal. DTW minimizes a global error by using a sequential optimization strategy where the current estimate of the global optimization function is updated for each possible step. Enough information is retained on the set of plausible hypotheses to allow the set of choices for the minimal global error to be reconstructed at the end. This means that a signal warped using DTW more closely resembles the original signal than a signal warped using a linear time warp (Gold et al., 2000).

To use DTW, feature extraction was first applied to each signal and then the test signal was warped against each of the reference signals and the error between these two signals was recorded. The smallest error was taken to represent the closest class of sound.

3.5.4. Long-term statistics

The long-term statistics (LTS) were implemented using the mean and covariance functions

available in the standard Matlab distribution. Mean and covariance were calculated for each of the reference signals and stored in a matrix. The mean and covariance of the test signal was then compared to this matrix. The closest match was selected as the correct class. If the closest mean and covariance occurred in different classes, the test was deemed to be inconclusive.

3.5.5. Gaussian mixture models

We implemented GMM using the Netlab toolbox developed by Ian Nabney (2002). GMM's use an unsupervised learning technique to determine the centres and variance of clusters within a search space. The GMM's in Netlab are initialized using the k-means classification technique and then trained with an expectation–maximization (EM) algorithm. We trained a GMM for each of the classes of sound in our domain. Once this was done, we worked out the probability of each of these models on the training data. Because each of the classes has equal priority, testing the system then simply involves finding the class C_i that produces the highest $p(\bar{x}|C_i)$, where $\bar{x}$ is the data under test.

4. Results and discussion

In this section, we present the results of our comparison and discuss the validity of these techniques to the domain of environmental sound recognition. Results are presented for each classification system using all feature extraction techniques. The result shown is the total classification rate over all sounds and all samples using the jackknife technique (Tables 3–7; Figs. 1–5).

4.1. Comparison of classification results with speech and music

This section shows results of selected techniques from the results section (LVQ, ANN, GMM) in other related domains (speech recognition and musical instrument recognition). This allows a comparison of these techniques among the different domains.

Table 3
Learning vector quantization (LVQ)

Method	% Correct
FT	50
MFCC	37.5
HCC	12.5
STFT	0
FWT	12.5
CWT	54

Table 4
Artificial neural network (ANN)

Method	% Correct
FT	0
MFCC	4
HCC	0
STFT	0
FWT	0
CWT	41

Table 5
Dynamic time warping (DTW)

Method	% Correct
FT	66
MFCC	70
HCC	29
STFT	58
FWT	12
CWT	70

Table 6
Long-term statistics (LTS)

Method	% Correct
FT	29
Power FT	29

Table 7
Gaussian mixture models (GMM)

Method	% Correct
FT	21
MFCC	46
HCC	12
STFT	46
FWT	25
CWT	21

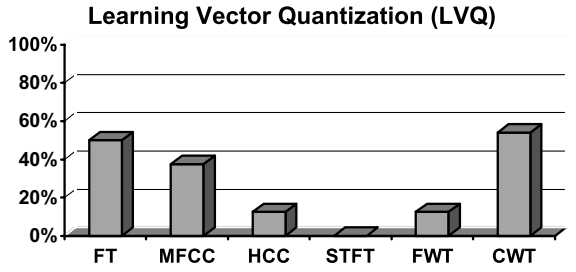

Fig. 1. Comparison of LVQ for environmental sound recognition.

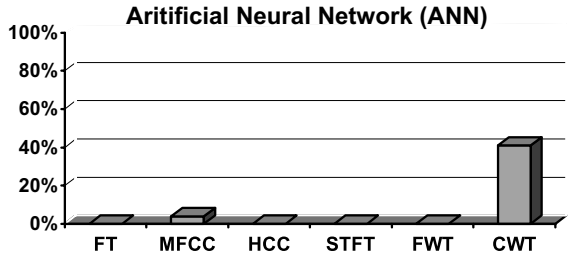

Fig. 2. Comparison of ANN for environmental sound recognition.

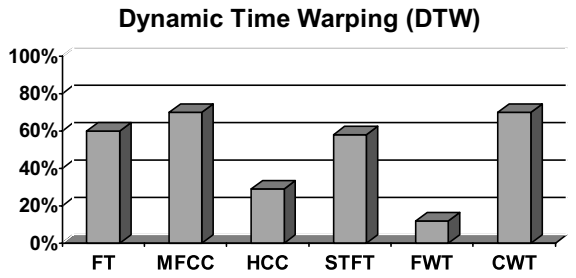

Fig. 3. Comparison of DTW for environmental sound recognition.

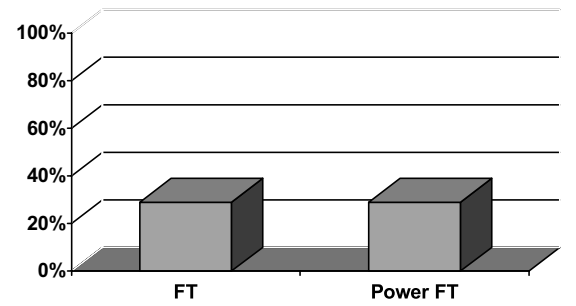

Fig. 4. Comparison of long-term statistics (LTS) for environmental sound recognition.

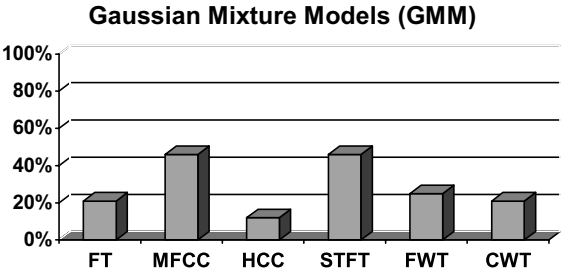

Fig. 5. Comparison of GMM for environmental sound recognition.

4.1.1. Speech recognition

For the sake of completeness, we compare our LVQ and ANN non-speech results with results reported for speech recognition systems. Due to the current popularity of HMM methods in speech recognition at the present time, results for DTW are difficult to find, therefore no DTW results are presented.

For ANN's, a selection of results from Castro and Perez (1993) are shown in Fig. 6. Their results were taken on an isolated word recognition set with typically high classification error, the Spanish EE-set. Castro and Perez's multi-layer perceptron (MLP) used the back propagation algorithm, contained 20 hidden neurons and was trained over 2000 iterations with various amounts of inputs. The figures given are the MLP's estimated error rate with a 95% confidence interval (Table 8).

For LVQ, results from Van de Wouver et al. (1996) are shown in Fig. 6 for both female and male voices. These results present statistics for both a standard LVQ implementation for speech recognition and an implementation of LVQ that then has fuzzy logic performed on it (FILVQ). As

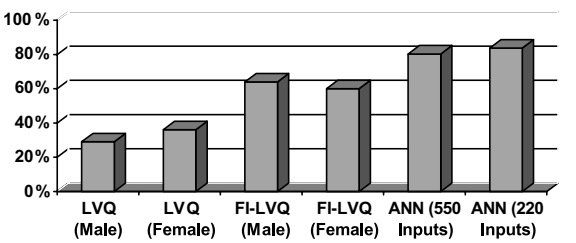

Fig. 6. Comparison of speech recognition results.

Table 8
ANN for speech recognition

Number of inputs	% Correct
550	80.3
220	83.7

Table 9
LVQ for speech recognition

Method	% Correct (female)	% Correct (male)
Standard LVQ	36	29
FILVQ	60	64

can be seen from the results, the use of LVQ for speech recognition produces rather low recognition results (Table 9).

Compared to the results from Figs. 1 and 2, these results are quite interesting. Fig. 1 shows the best result for LVQ in speech is 54%, when combined with a CWT feature extraction technique. The results for LVQ in speech are only between 6% and 10% above this, even with the application of fuzzy logic. Without the application of fuzzy logic, the results for speech using LVQ are 18% worse than the non-speech results.

For ANN's, the results from speech show a much higher percentage rate than the results from non-speech. For non-speech, the best result is 41% using the CWT feature extraction technique. This is much lower than the 83.7% achieved using ANN's for speech. We believe this is due to the non-speech data being non-linearly separable, and will elaborate more on this in the discussion section.

4.1.2. Musical instrument recognition

We looked at techniques in musical instrument recognition, considering that it might be closer to environmental sound recognition than to speech. In this field, two seminal works stand out as using GMM's. Marques and Moreno (1999) used several different techniques for the feature extraction and classification of musical instruments, with the best results coming from a combination of mel frequency cepstral coefficients with either GMM or support vector machines (SVM). Martin (1999) reports initial results from Marques showing a classification rate of 72% for professional recordings and 45% for non-professional recordings. Since then, a further technical report from Cambridge Research Laboratory shows a classification rate of 70% when using mixed data, increased to 98% when using data from a single source (Table 11). This suggests the applicability of robustness techniques (also tested in speech recognition) to this domain, in order to combat this problem with variable training/test data.

Brown also presents a system using cepstral coefficients (Brown, 1999). In this case, the system uses Q-cepstral coefficients with a GMM for classification (one model for each instrument). On independent, noisy samples of music, the system achieves a classification rate of 94%, between oboe and saxophone recordings. However, this system achieves only 84% when extended to include four samples of instrument (oboe, saxophone, flute and clarinet) (Table 10).

If we compare these results (Fig. 7) to the results for non-speech sounds (Fig. 5), we see that GMM's can be much better applied in the musical instrument domain. The best result for GMM's in non-speech is 46%, while musical instrument recognition can achieve a 94% recognition rate. However, it must be considered that the results for non-speech are taken on eight classes of sounds. The results for musical instrument recognition are taken on only two and four classes of sounds. This could account for the higher recognition rate. Brown also shows with her results that classification rate decreases quickly as the number of classes increase (down from 94% to 84% for two and and classes respectively). It is possible that, if Brown ran her tests on eight classes of musical instru-

Table 10
Q-Cep/GMM for musical instrument recognition

2 Instrument types	4 Instrument types
94%	84%

Table 11
MFCC/GMM for musical instrument recognition

Mixed data	Single source data
70%	98%

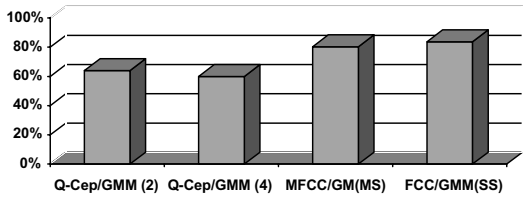

Fig. 7. Comparison of musical instrument recognition results.

ment, she would get similar results to those shown for non-speech sounds.

4.2. Discussion of results

The results obtained for this set of experiments are somewhat surprising (Fig. 8). Even though the results from speech recognition suggest that the ANN will outperform LVQ, the opposite occurs for non-speech recognition. We propose that this is due to the closeness of the various environmental sounds presented to the two networks.

It is widely accepted that one of the main advantages of LVQ over ANN's is their ability to correctly classify results even where classes are similar. In this case, sounds such as footsteps (close) and footsteps (distant) appear the same but contain slightly lower or higher amplitudes. LVQ is able to classify these sounds properly where the ANN cannot distinguish them. Furthermore, the detailed results of each test show that the ANN was classifying footsteps (close) as footsteps (distant) and vice versa. To support this hypothesis, further tests were performed on the ANN using several different MSE values (to allow more training time). The results these different experiments are presented in Table 12 and compared in Fig. 9.

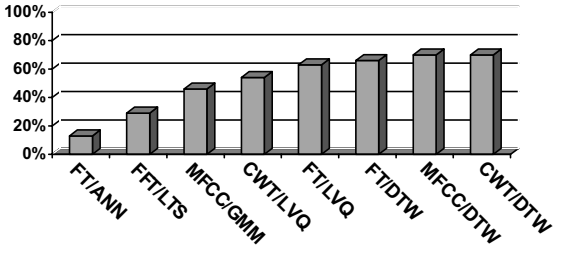

Fig. 8. Comparison of best results for non-speech sound recognition.

Table 12
Further ANN results for environmental sound recognition

Method	% Correct (MSE – 0.001)	% Correct (MSE – 0.0001)
FT	0	0
MFCC	4	4

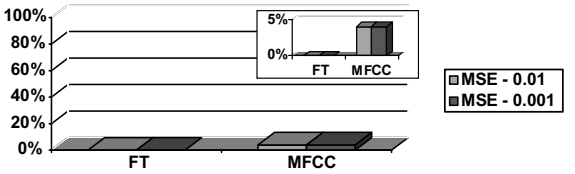

Fig. 9. Comparison of ANN results with alternative MSE values for environmental sound recognition.

From these results, it can be seen that the ANN results remain the same regardless of the MSE value. This suggests that the ANN has problems training the sample sounds, most likely because these sounds are non-linearly separable.

The performance of the MFCC feature extraction algorithm over the Fourier transform (FT)-based frequency extraction algorithm is also interesting. Surprisingly, in all cases except when DTW or GMM is used as a classifier, the FT algorithm outperforms the MFCC algorithm. However, to achieve the same results with the FT algorithm, it has to spend almost 10 times as long training as the MFCC algorithm does.

For the LVQ and ANN tests, it seems that the MFCC algorithm can achieve a maximum classification rate of approximately 37.5%. In contrast, the FT algorithm can achieve a slightly higher rate, reaching its maximum at around 50%.

The DTW algorithm also produces surprising results. This algorithm shows only a small difference (equivalent to one classification) in performance between the MFCC algorithm and the FT algorithm. This is in contrast to the large difference between these algorithms in the LVQ and ANN tests. DTW performs classification much quicker than the LVQ and ANN techniques. This is most likely due to the fact that DTW does not require any training and instead relies on a series of

reference models. The downside to this approach is the extra storage space required for these templates.

In contrast to results presented by Lilly (2000), our results show a substantial difference in classification between the HCC technique and the MFCC technique. Due to the fact that other researchers report similar classification rates using these two techniques (e.g. Lilly, 2000; Gold et al., 2000), implementation of these techniques could conceivably be improved. However, since MFCC seems to be the more popular technique and produces the better results of the two techniques, at this stage we will continue to use it in its current form.

For time-frequency techniques, these results show that the combination of CWT with DTW produces the best results, with the CWT producing a top comparative study percentage of 70% with DTW. Results are also promising for the use of the CWT with LVQ and ANN, producing top results of 54% and 41% respectively.

The results from this comparative study reveal some interesting findings. It is interesting to note the poor performance of the STFT algorithm with both ANN and LVQ. Despite providing average performance using DTW (29%), a STFT combined with either an LVQ or ANN network fails to provide any correct classifications. Although results for stationary feature extraction techniques support this low classification rate for ANNs, performance with LVQ is surprising. Further research may endeavour to manipulate the resolution of the STFT in order to improve LVQ classification rate (an issue that does not affect the wavelet family of time-frequency techniques). Nonetheless, the performance of STFT with GMM produces quite good results, suggesting that maybe the problem lies with the learning algorithms in LVQ/ANN.

Overall, it is clear from these results that the DTW classification technique could be considered the most suitable for environmental sound recognition, especially in a surveillance context. Not only does the DTW technique perform the best of all the techniques that we investigated and compared, but it also produces the results quickly (under 1 second for a testing classification as compared to an average of 5 seconds for artificial intelligence techniques). However, this technique still needs to be investigated, as it uses a template matching method, which could turn out to be a weakness when the amount of sounds in the database increases. Nevertheless, there is ample opportunity to improve the technique in these circumstances with the use of difference measures to produce general representations of each class of sound.

For feature extraction, the results are not so clear-cut. They show that the pseudo-frequency technique of MFCC's produces a classification rate of 70% when used with the DTW technique. However, they also show that the same classification rate can also be achieved using the time-frequency technique of CWT. The relative effectiveness and classification efficiency of these two techniques will become apparent when they are applied to a larger database of sounds. Once again, opportunity exists to improve upon these techniques by systematic testing and refinement of these techniques over several iterations. The results presented in this paper demonstrate the obvious superiority of these techniques over the other techniques that we investigated for environmental sound recognition.

It could be argued that these classification rates do not parallel with the accuracy that can be achieved in speech recognition using HMM's. However, as was explained earlier, they are not suitable for environmental sounds.

In general, due to the variability inherent in environmental sounds, accuracy with techniques such as DTW will probably always be lower than the classification rate that can be achieved in the more constrained area of speech recognition.

5. Conclusion

This paper presented the results of a comparative study of time-frequency and frequency-based (or pseudo-frequency) techniques for non-speech environmental sound recognition and showed the applicability of either of these representations to environmental sound recognition. However, classification rates do not parallel with the accuracy

that can be achieved in speech recognition using HMM's. Due to the variability inherent in environmental sounds, accuracy is probably lower than with the more constrained area of speech recognition.

The results revealed that a combination of continuous wavelet transform with dynamic time warping produces a classification rate of 70%. Combination of MFCCs with dynamic time warping also produced 70%. From this, it is clear that DTW is a superior technique for classification of environmental sounds. Now that this obvious superiority of techniques has been shown, further refinements can be performed on these techniques to possibly produce even better classification rates.

Work is underway for applying a new type of technique to the recognition of environmental sounds, in the direction of structured classification.

References

Brown, J., 1999. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. J. Acoust. Soc. Am. 105, 1933–1941.

Castro, M.J., Perez, J.C., 1993. Comparison of geometric, connectionist and structural techniques on a difficult isolated word recognition task. In: Proceedings of European Conference on Speech Communication and Technology, ESCA, Berlin, Germany. vol. 3, pp. 1599–1602.

Cohen, L., 1995. Time-Frequency Analysis. Prentice-Hall, New Jersey, USA.

Cowling, M., Sitte, R., 2000. Sound identification and direction detection in Matlab for surveillance applications. In: Proceedings of Matlab Users Conference, Melbourne, Australia.

Cowling, M., Sitte, R., 2001. Sound identification and direction detection for surveillance applications. In: Proceedings of ICICS 2001, Singapore.

Cowling, M., Sitte, R., 2002a. Recognition of environmental sounds using speech recognition techniques. In: Advanced Signal Processing for Communications Systems. Kluwer Academic Publishers.

Cowling, M., Sitte, R., 2002b. Analysis of speech recognition techniques for use in a non-speech sound recognition system. In: Proceedings of DSPCS 2002, Manly, Australia.

Daubechies, I., 1992. Ten lectures on wavelets. Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Vermont, USA.

Donoho, D., Duncan, M., et al., 2002. WaveLab Toolbox™, Stanford University, version 0.802.

Gold, B., Morgan, N., 2000. Speech and Audio Signal Processing. John Wiley & Sons, New York, NY.

Goldhor, R.S., 1993. Recognition of environmental sounds. In: Proceedings of ICASSP, New York, NY, USA. vol. 1, pp. 149–152.

Hubbard, B., 1995. The World According to Wavelets: the Story of a Mathematical Technique in the Making. Wellesley, Mass, USA.

Kohonen, T., 1997. Self-Organizing Maps. Springer-Verlag, Berlin, Germany. Printed in the USA.

Lee, C.H., Soong, F.K., Paliwal, K., 1996a. An overview of automatic speech recognition. In: Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers, Norwell, MA.

Lee, C.H., Soong, F.K., Paliwal, K., 1996b. An overview of speaker recognition technology. In: Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers, Norwell, MA.

Lilly, B., 2000. Robust speech recognition in adverse environments, Ph.D. Thesis, Faculty of Engineering, Griffith University, Nathan Campus.

Liu, L., 1999. Ground vehicle acoustic signal processing based on biological hearing models, Masters Thesis, University of Maryland, College Park.

Marques, J., Moreno, P.J., 1999. A study of musical instrument classification using Gaussian mixture models and support vector machines. Cambridge Research Laboratory Tech. Report.

Martin, K., 1999. Sound-source recognition: A theory and computational model, Ph.D. Thesis, MIT Media Lab, Massachusetts Institute of Technology, USA.

Nabney, I., 2002. Netlab: Algorithms for Pattern Recognition. Springer-Verlag, London, UK. Printed in Great Britain.

Orr, M., Pham, D., Lithgow, B., Mahony, R., 2001. Speech perception based algorithm for the separation of overlapping speech signal. In: Proceedings of The Seventh Australian and New Zealand Intelligent Information Systems Conference, Perth, Western Australia. pp. 341–344.

Rodman, R., 1999. Computer Speech Technology. Artech House Inc., Norwood, MA.

Sampan, S., 1997. Neural fuzzy techniques in vehicle acoustic signal classification, Masters Thesis, Virginia Polytechnic Institute and State University, USA.

Slaney, M., 1998. Auditory Toolbox™, Interval Research Corporation, version 2.

Van de Wouver, G., Scheunders, P., Van Dyck, D., 1996. Wavelet-FILVQ classifier for speech analysis. In: Proceedings of International Conference on Pattern Recognition, Vienna. pp. 214–218.