ICSV13 - Vienna
The Thirteenth International Congress on Sound and Vibration
Vienna, Austria, July 2-6, 2006

CLASSIFICATION STRATEGIES FOR AUDIO SIGNALS USING WAVELET ANALYSIS AND ARTIFICIAL NEURAL NETWORKS

Neil McLachlan*[1] and Dinesh Kant Kumar[2]

[1]Department of Psychology. University of Melbourne
Victoria 3001, Australia
[2] School of Electrical and Computer Engineering
RMIT University, GPO Box 2476V Melbourne, Victoria 3001, Australia
mcln@unimelb.edu.au

Abstract

This paper describes an audio event identification system named Cyber Ear. Cyber Ear is loosely based on human hearing and uses statistical descriptors of the detail wavelet coefficients. Neural network has been used for classification of feature vectors. The system only requires the weight matrix descriptors to classify an unknown sound event. Two experiments that are closely related to real world problems have been conducted. The results demonstrate that Cyber Ear is suitable for identifying vehicular sounds at large distances under varying conditions, and for monitoring many different train station sounds with good accuracy.

INTRODUCTION

Computational auditory scene analysis (CASA) attracted some interest after the publication of Bregman's text on human auditory scene analysis [2] and was expected to find many applications in data retrieval, autonomous robots, security and environmental analysis. Ellis describes approaches to CASA as being either data driven (or "bottom-up" processing in psychological terms) where raw data is sequentially transformed into abstract representations, or prediction driven ("top-down") where predictions generated by higher level abstractions that have evolved in relation to previous data are used to segregate new data on an ongoing basis [7]. Classification of environmental noise has been attempted using statistical classification methods on 1/3 octave filtered spectra of transport sounds [4]. Higher order statistical features have been used to statistically classify musical instrument

sounds [6,3], and wavelet transforms have been used for feature extraction prior to statistical or neural net classification of marine mammal and musical instrument sounds [5,8]. Gygi [9] found that octave band spectral energy distributions and amplitude fluctuations within these distributions over time produce timbral cues that are important in human recognition of environmental sounds. The rate of successful recognition of environmental sounds decreased significantly when participants were presented with full bandwith amplitude modulated representations of sounds similar to the models used by Ellis.

The authors have developed a system named Cyber Ear© to identify a small dictionary of sounds. This paper describes tests on this system to classify complex real world sounds. To be computationally inexpensive and have real time capabilities such a system requires expedient choices of data reduction, feature extraction and classifiers that utilize the principles of human sound recognition processing without trying to accurately model them. Supervised training of classifiers with example sets of sounds for each class will be undertaken prior to classification trials. A detailed description of the development and performance of Cyber Ear© compared to other classification strategies is detailed elsewhere [12] and a short description follows.

THEORY

The Cyber Ear

Wavelet filter banks have many features in common with the multi-resolution frequency-time filtering of the human middle ear [1]. They also have the advantage over Fast Fourier Transforms of high frequency resolution at low frequencies and high temporal resolution at higher frequencies [14]. In keeping with the findings of Gygi, the feature vectors used for classification were the RMS amplitude and variance of short frames of wavelet coefficients over 12 decomposition levels. A sensitivity analysis was undertaken on the type of wavelet that would best separate these features for a range of sounds and the Db3 wavelet was found to perform best [12]. It has been used throughout the subsequent research described here.

There are a number of classification tools available that may be suitable for classifying signals based on the small set of features extracted. Artificial Neural Networks (NN) is an option that can be used when the separation may be non-linear. The other advantages of such a system are that it can learn from examples and it is easy for a user with limited technical skills to change the configuration for changed data set. This paper reports the use of supervised back-propagation neural networks with the following specifications:

- A feed-forward network with a back propagation learning algorithm and ‘logsig’ threshold function to produces binary outputs for classification.
- The network was trained using ‘momentum’. The initial learning rate was chosen to be 0.01 selected after an iterative process.

- Three layers were used. The output layer had 1 node for every 2 signal classes, the input layer had 24 nodes, one each for the mean and variance of all 12 wavelet decomposition levels, and the hidden layer had 12 nodes.
- The limit for sum-squared error was chosen to be 0.001.

The user cannot predict every sound that may occur in a natural environment and incorporate them into an individual training class, so some provision should be made for an 'unknown' output. Thresholds were applied to the NN outputs such that outputs within the range of the threshold from a binary output were classified as belonging to a known sound class, and all other outputs were classified as unknown. These thresholds could be varied depending on the conditions of use and their values may be based on prior knowledge of the signals, and the number of classes (generally the more classes, the greater the threshold). The value of the threshold also determines the rate of false negatives and false positives. Selection of the threshold value may also be done iteratively, which is not a very difficult task as there are only few options between 0.5 and 1.

Example recordings are concatenated and used to develop a training file for each sound class. The closeness of different sound classes is user defined, and the success of Cyber Ear© to separate the classes is dependent on the number of classes and difference between examples of the same class. The recordings of training and test files are time framed for further analysis. The length of frame is user defined and preliminary experiments suggested a use of 0.1 to 1 second windows. After training, a computer file with the weights and biases of the NN is saved for later use in classifying a test file.

The mean and variation of the detailed wavelet coefficients are measures of the statistical variation of the signal and not a direct measure of the spectral content which is more prone to variation in changed acoustic environment. It is important to characterise a background sound that comprises the baseline signal and include it in the set of training files so that the NN does not make spurious outputs in the absence of target sounds. Transient sources should be segmented from the recordings used for the background training file as far as possible. These sounds may comprise a new target sound class if they occur often, or they may be simply ignored. The NN is likely to give an 'unknown' response when transient sounds that it has not been trained for occur in a test recording.

A simple model was created in Matlab using well-established empirical data from the literature [13] to better understand the acoustic properties of background sounds when they arise from many uncorrelated sources. Outdoor sound intensity is be expected to decay at 3 dB with every doubling of distance to a sound source. The attenuation also depends on the frequency of the signal, atmospheric humidity and the density of intervening vegetation, assuming no reflections (apart from the ground) and ignoring lens effects due to atmospheric temperature gradients or wind effects. Assuming an even distribution of noise sources, their number increases as the square of distance from the recording location.

Depending on the sound attenuation and the density and amplitude envelopes of sources, at a certain distance from the recording location they should overlap to create a reasonably constant sound that we define as the background. Figure 1a

approximates how constant, white noise sources evenly distributed about a recording location sum over distance to produce background sound levels under different vegetation densities at 60% humidity. Sources at distances greater than 500 metres make no contribution to the background with heavy vegetation, whereas with light vegetation lower frequencies continue to contribute for more than 2 km.

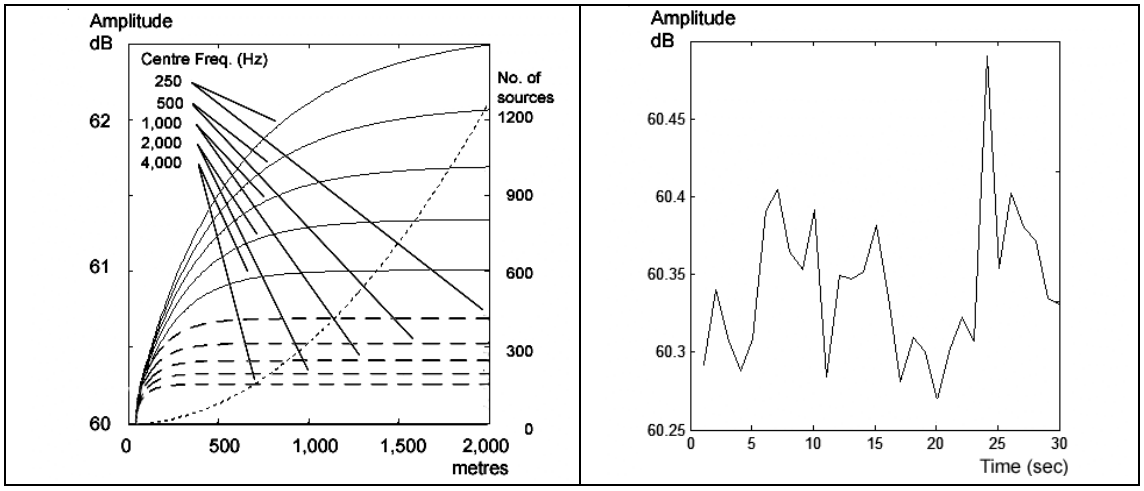

Figure 1. a) An approximation over 5 octave bands of the summation of white noise sources of 65 dB at 1 m, evenly distributed 100 m apart, over a radius of 2 km. Solid lines =light vegetation and broken lines =heavy vegetation. Dotted line = number of sources. b) The time varying behavior of the 1000 Hz octave band in Figure 1a (light vegetation) assuming random onsets of 1 second sources occurring for 10% of the time summed over 1 km.

Figure 1b plots the summation over time of intermittent 1-second sources the same as those used in Figure 1a for light vegetation. Figure 1b shows that even for widely dispersed sources the background sound tends towards constant amplitudes (a range of only 0.3 dB) within a 1 km range. An instance of the same source located as close as 32 meters to the recording location would produce amplitudes of only 50 dB, 10 dB below the average background level, and therefore would not be discernable from the background variation by amplitude alone. In real situations the constancy of the background sound will depend on how many of these sources are correlated (as may be the case in wind or traffic noise) and their peak amplitude variation.

An amplitude threshold value to segment recordings into examples of background sounds and louder transient sounds can be evaluated manually by application of a Receiver Operating Condition (ROC) [11]. When the target sound levels are within a few dB of the background, target training files need to be recorded in the presence of similar background conditions, or have the background sound artificially added. This ensures that the signals used for training have similar properties as the test data.

In some applications sounds occurring over a range of relatively close distances to the recording location will need to be classified. These range differences may cause signal amplitudes to vary by up to 20 dB with very little change in the signal spectrum due to atmospheric absorption. This amplitude variation will cause large

discrepancies in the RMS amplitude feature vectors for the training and test files unless they are first normalized.

METHOD

Two different examples of the application of The Cyber Ear© are presented in the following. The first example is for a public security system for an unmanned railway station with sounds of men and women screaming as target sounds amongst the usual sounds of the station. The second example is a military or boarder protection acoustic surveillance system in which vehicle sounds are identified at long ranges.

Recordings were made of the background sound, and of trains arriving and departing from a suburban station using an omni-directional microphone. Six examples each of men and women screaming were taken from a Hollywood sound effects library. The train sounds were manually segmented into 5 classes for training the NN along with 3 other classes as described in Table 1. Example recordings were concatenated into individual training files for each class and normalized before the calculation of feature vectors for 0.1-second frames and NN training. Training required 10,000 epochs to arrive at an error of 0.005. A new, 16-second recording of a train arriving at the station was tested with Cyber Ear©.

Table 1. Sound classes used in railway station monitoring

Class No.	Description	Class No.	Description
1	Compressed air released from the train brakes	5	Engine noise
2	Background sound	6	Female screams
3	Brake squeal	7	Male screams
4	Clanking of the train wheels over joins in the track	8	Train whistles

For the second example, two sets of recordings were made of vehicle sounds using a directional microphone (Sennheiser M61). The variations in the recordings conditions included the environmental conditions and presence of intervening obstacles. The experiments were also conducted with one or two target classes to test the system's capability to detect multiple targets under these conditions. The vehicles were traveling on sealed roads and the distances were of the order of 1-2.5 Km.

RESULTS

Figure 2 shows the Cyber Ear© output for the arrival of a train. The sequence of acoustic events heard in the recording is well represented by Cyber Ear. Sounds rarely occur in isolation, so the Cyber Ear© output at any time represents the dominant feature vectors arising from the recording. For example, the clanking sound

dominates the NN output for the time it is present, possibly due to its highly impulsive nature creating distinctive feature vectors with high variance. However because the clanking is a short loud sound, other sounds can be classified in the frames between the frames dominated by the clanking. The overlapping sound classes cause numerous scattered 'unknown' classifications but only 5 segments (out of total 160 frames) were classified incorrectly (2 background, 1 female and 2 male screams). This is an error rate of only 3% of false positives. Such spurious results for the screams can be readily ignored, as the real event must last for more than 1 second.

New examples of female and male screams have been classified with 100% percent accuracy when superimposed over the background sound with a signal to noise ratio greater than 10 dB. This is sufficient for the screams to be successfully classified for any location on the station platform at a centrally located microphone. However the screams cannot be successfully classified while a train is arriving or leaving.

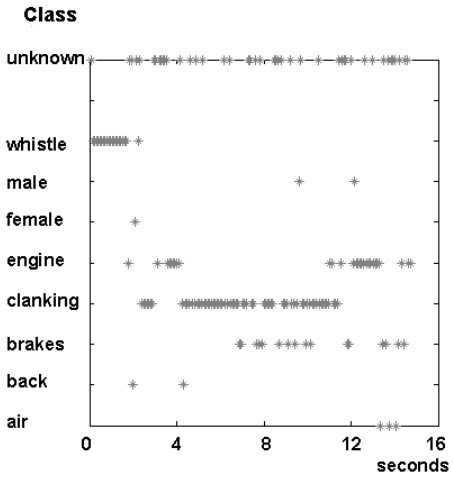

Figure 2. Classification results for the arrival of a train calculated in 0.1 second frames. Classification threshold = 0.2.

Figure 3 shows the output of Cyber Ear© for over 360 seconds (in 1 second frames) when trained to detect the presence of cars and trucks travelling at 60 km/h on a sealed road at ranges greater than 2.5 km using a directional microphone with line of sight in low noise conditions. The outputs of the 2 NN output nodes compared to the classification thresholds, and the NN binary outputs corresponding to each class are also shown Figure 3. Node 1 is the solid line and node 2 is the broken line. The regions of extended elevation of either line are indicative of a positive classification of a car or truck while other short peaks may be noise.

Figure 4 shows the output of Cyber Ear over 2 ½ minutes in adverse recording conditions when trained to detect the presence of cars travelling at 80 km/h on a sealed road at ranges greater than 500 meters using a directional microphone. Wind speeds were variable up to 15 km/h across the range, and there was medium density of trees and shrubs and low landforms between the vehicle and the recording location.

Under these conditions it was found to be necessary to initially segment background recordings and create 2 new sound classes. Therefore the system was

trained with 4 classes of sound; 1) background sound of wind in vegetation, 2) cars travelling at 80 km/h on a sealed road at a range of 500-1000m in the same background sound conditions, 3) cockatoo bird calls, and 4) wind noise on the microphone. The training files were were broken into 0.1 second frames before feature vectors were calculated. The car was detected at a range of 1 km for about 10 seconds at around 100 seconds into the recording as it moved between landform obstacles. Other detections are possibly spurious. The node 1 outputs between 60 and 100 seconds reflect the transient nature of individual birdcalls compared to the smoother node 2 responses for the car.

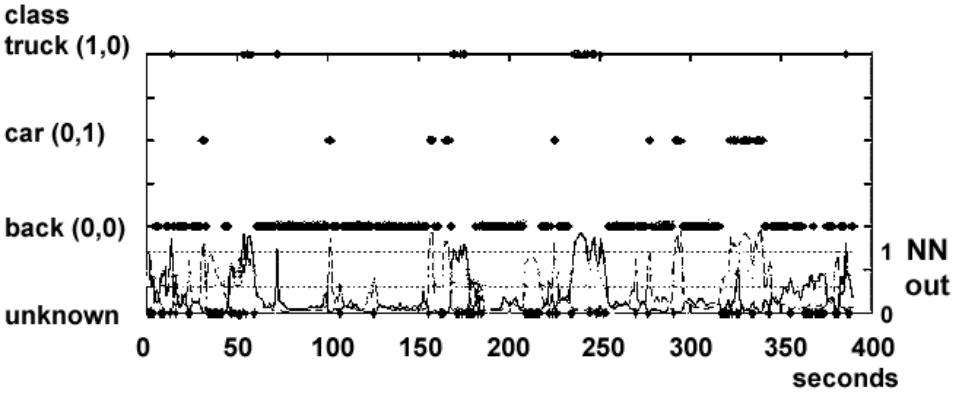

Figure 3. Classification results for the detection of a cars and trucks traveling at 60 km/h on a sealed road at 2.5 km (line of sight). Solid line = NN output 1, broken line = NN output 2, dotted line = classification threshold of 0.2.

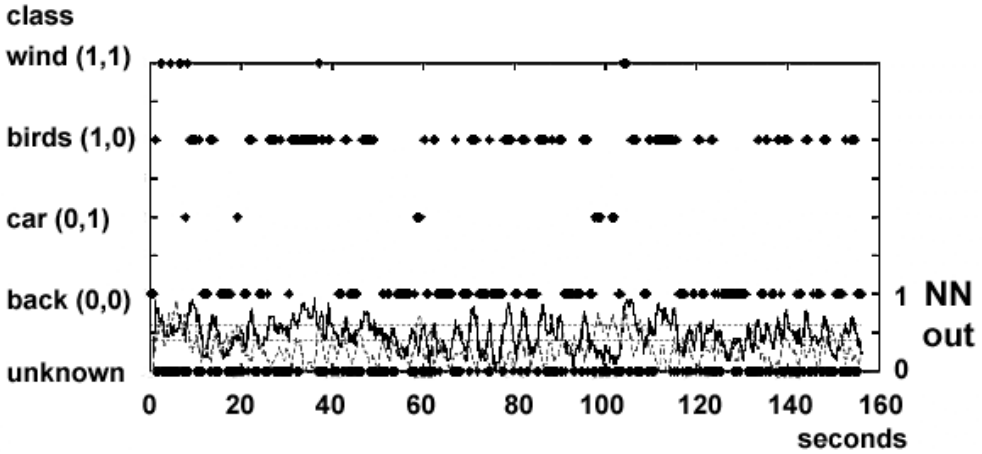

Figure 4. Classification results for the detection of a car traveling at 80 km/h on a sealed road at 0.5 – 1 km in the presence of substantial background sounds and physical obstacles.

CONCLUSIONS

The Cyber Ear©, a patented audio event identification tool, based on wavelet statistical description and neural net classification, has been described in this paper. The application of the system requires that acoustic signals recorded at the site be

segmented into transient and constant sounds. Commonly occurring transient sounds may become a target class used in training the NN, and less common transient sounds can be ignored. Constant sounds that don't contain embedded target sounds become another training class called the background. For mobile sources at close range all sound files should be normalized before feature vectors are calculated to account for large amplitude changes.

The background needs to be carefully defined when target sounds are within a few decibels. In this case all training files may need to include the background sound, and care needs to be taken to ensure that the training and test files are presented to the system at consistent amplitudes. In a simple mathematical approximation, intermittent, widely dispersed uncorrelated sources were shown to sum over time to produce a constant background sound in an idealized outdoor environment. Two real-world applications of Cyber Ear© in which target sounds are at close or distant range have been shown to produce promising results.

REFERENCES

1. Agerkvist, F. T., "A Time-Frequency Auditory Model Using Wavelet Packets" J. Audio Eng. Soc. 44 (1-2), pp 73-50, 1996.
2. Bregman A. S., *Auditory Scene Analysis: The Perceptual Organisation of Sound,* (MIT press, Cambridge USA 1990).
3. Brown J. C., Houix O. and McAdams S. "Feature Dependence in the Automatic Identification of Musical Woodwind Instruments" J. Acoust. Soc. Am., **109** (3), pp. 1064-1072, 2001.
4. Couvreur C. and Bresler, Y. "Automatic Classification of Environmental Noise Sources by Statistical Methods", Noise Control Eng. J**. 46** (4) pp. 167-182, 1998.
5. Delfs C. and Jondral F., "Classification of Piano Sounds Using Time-Frequency Signal Analysis", IEEE International Conference on Acoustics, Speech, and Signal Processing, **3,** pp. 2093 –2096, 1997.
6. Dubnov S. and Tishby N., "Analysis of Sound Textures in Musical and Machine Sounds by Means of Higher Order Statistical Features". IEEE International Conference on Acoustics, Speech, and Signal Processing, **5**, pp. 3845 –3848, 1997.
7. Ellis D. P. W., *Prediction Driven Computational Auditory Scene Analysis*, (Ph. D Thesis, MIT, 1996).
8. Huynh Q. Q., Cooper L. N., Intrator N. and Shouval H., "Classification of Underwater Mammals Using Feature Extraction Based on Time-Frequency Analysis and BCM Theory", IEEE Trans. Sig. Proc. **46** (5), pp. 1202-1207, 1998.
9. Gygi B., Kidd G. R. and Watson C. S., "Spectral-Temporal Factors in the Identification of Environmental Sounds", J. Acoust. Soc. Am., **115** (3), pp. 1252-1265, 2004.
10. Kumar S., Kumar D., Sharma A. and McLachlan N., "Visual Hand Gesture Classification" *International Journal of Wavelets, Multiresolution and Information Processing,*
11. Marques de Sa, J. P., *Applied Statistics using SPSS, Statistica and Matlab*, (Springer, Heidelberg, New York, 2003).
12. McLachlan, N. M., Kumar, D. K. and Becker J., "Wavelet Classification of Indoor Environmental Sound Sources", *International Journal of Wavelets, Multiresolution and Information Processing*
13. Smith, B. J., Peters, R. J., and Owen S., *Acoustics and Noise Control,* pp. 64 67, (Longman, Essex, 1985).
14. Szu, Harold H., Kadambe, Shubha, 'Neural Network Adaptive Wavelets for Signal Representation and Classification', Optical Engineering, **31** no. 9 pp 1907-1916, 1992.