Acoustic Hazard Detection for Pedestrians with Obscured Hearing

Justin Lee, and Andry Rakotonirainy

***Abstract*—Pedestrians' use of mp3 players or mobile phones can pose the risk of being hit by motor vehicles. We present an approach for detecting a crash risk level using the computing power and the microphone of mobile devices that can be used to alert the user in advance of an approaching vehicle so as to avoid a crash. A single feature extractor classifier is not usually able to deal with the diversity of risky acoustic scenarios. In this paper, we address the problem of detection of vehicles approaching a pedestrian by a novel, simple, non resource intensive acoustic method. The method uses a set of existing statistical tools to mine signal features. Audio features are adaptively thresholded for relevance and classified with a three component heuristic. The resulting Acoustic Hazard Detection (AHD) system has a very low false positive detection rate. The results of this study could help mobile device manufacturers to embed the presented features into future potable devices and contribute to road safety.**

***Index Terms*—Environmental Sound Recognition, Pedestrian Safety**

I. Introduction

THIS paper presents a novel approach to using environmental sound processing as a means of detecting motor vehicle hazards for pedestrians who are inattentive or have their hearing obscured by music listening devices.

Pedestrians who do not pay attention to other road users and road signs are more exposed to crashes. Pedestrians are particularly vulnerable to being struck by vehicles when their hearing is obscured, reducing their awareness of oncoming traffic. This problem has become more prevalent in the past few years due to the ubiquity of personal music players. Not only does this cause pedestrians hearing to be obscured by the music played through their headphones, but it also distracts the pedestrian from being aware of potential dangers. For example the use of mobile phones impairs pedestrians ability to cross roads safely.

With the convergence of mobile phones and mp3 players, in devices such as the iPhone, many personal music players will be equipped with microphones. By processing environmental sounds from the embedded microphone, potential vehicle hazards may be identified and the pedestrian warned in time to avert being hit by an oncoming vehicle.

In this paper we present a new Acoustic Hazard Detection (AHD) system for detecting when a motor vehicle is approaching a pedestrian, based solely on the environmental sounds detected via the microphone embedded within a mobile phone handset. The proposed algorithm is simple, decomposable and based on adaptive thresholding. It has a very small footprint and has a very low false positive detection rate. This system can be used to warn a pedestrian, who may be listening to music or engaged in a phone conversation, so as to avoid being struck by a vehicle.

In regards to audio classification, while there has been a considerable amount of work done on voice and speaker recognition, and some work done on music and musical instrument recognition [10], [11], [4], [17]), there has been comparatively little work done on environmental sound recognition.

Some of the work done on environmental sound recognition includes: mobile phone localization [2]; sound classification and localization [3]; sound classification for autonomous surveillance [7], [24]; auditory scene classification [5], [6]; automatic recognition of acoustic environment [1]; classification of fatigued bills [22]; and classification and retrieval from audio database [15].

Some work has already been done in using acoustic processing to classify approaching vehicles by type, and to determine their speed [3]. Although this uses specialized equipment and is primarily aimed at military applications. Recently, work has been done utilising mobile phone's microphone and camera capabilities to identify location within a GPS locale [2]. This work shows that it is possible to utilise a mobile phone's microphone for detecting environmental features.

In general environmental sound recognition is approached as a two part process, composed of feature extraction and classification. Feature extraction involves taking various time and frequency domain measures of the audio signal that can be used to characterise the sounds contained in the signal. Classification takes the extracted features and then compares these to known sounds, to form a hypothesis regarding what this sound may be. Typically classification is done using machine learning techniques, such as neural networks, or by comparing with a database of sounds indexed by features.

In this paper we present a method of identifying approaching vehicles using a manually chosen subset of audio features that are adaptively thresholded for relevance and classified with a three component heuristic. The use of such a simple and decomposable system, as opposed to machine learning techniques, has obvious advantages for deployment as an actual application embedded in real devices in real environments. Hence, computational and memory requirements are kept to a minimum, the behaviour of the system is predictable and reproducible, and the system can be further refined to improve performance.

The rest of this paper is organised as follows. Background is given in Section II followed by the presentation of the

J. Lee and A. Rakotonirainy are with the Centre for Accident Research and Road Safety – Queensland, Brisbane, Australia e-mail: jm.lee, r.andry@qut.edu.au.

Manuscript received January 25, 2010; revised November 2, 2011.

design requirements in Section III. Section V presents the experimental analysis followed by discussion on the findings. Section VIII concludes this paper and highlights future work.

II. Background

A typical approach to environmental sound recognition involves feature extraction with a time-frequency algorithm, such as wavelets or a Short Time Fourier Transform (STFT), followed by classification of the extracted features using machine learning approaches, such as a neural network.

Feature extraction involves the manipulation of audio to extract a set of characteristic features of the sounds present in the audio. Time-frequency extraction algorithms are able to extract both time (when a sound occurred) and frequency (the sound's properties) information. The STFT is easy to implement, intuitive and fast. It is implemented by applying a Fast Fourier Transform (FFT) to successive overlapping windows of audio data, and produces spectral information (including phase and magnitude) for the frequency components in the signal.

The STFT's main problem is that to increase the frequency resolution, the time resolution must be decreased, and vice versa. Wavelets are used to combat the resolution problem of the STFT by applying good time resolution (and hence poor frequency resolution) at high frequencies, and good frequency resolution (with poor time resolution) at low frequencies. The fast (discrete) wavelet transform is usually used for encoding and decoding of signals, while the continuos wavelet transform is used in recognition applications [9].

Another time-frequency algorithm is the Wigner-Ville distribution, which has higher resolution than the STFT, but suffers from cross-term interference and produces results with coarser granularity than wavelet techniques, while also being extremely slow to compute [9].

Mel Frequency Cepstral Coefficients (MFCCs) are a psuedo-frequency technique that is popular in speech and music recognition. However, while the signal is split into time-slices, it is not a true time-frequency technique as each time-slice needs to be taken in context with other time-slices to generate useful information [9]. While MFCCs have been shown to work well for structured sounds, such as speech and music, their performance degrades in the presense of noise. Furthermore, MFCCs are not effective at analyzing noise-like signals that have a flat spectrum. Environmental audio contains a wide variety of sounds, including those that have a strong temporal component but are noise-like with a broad flat spectrum, and these cannot be modelled by MFCCs [6]. Some examples of sounds like this, include rain, bird chirping, and vehicle noise.

Other feature extraction techniques and sound classification have been published and can be found in [9], [7], [5], [6],[16], [18], [20], [14], and [13].

III. Design requirements

For vehicle detection to be usefully deployed to pedestrians, it needs to be embedded within a portable music player or phone, provide reliable detection (low rate of false positive), and work in real time. These devices can vary considerably in their computational capabilities, and to perform in real time, it is necessary that a streamlined approach is taken.

Wavelets are a popular method of extracting time-frequency information from an audio signal. However, wavelets are resource intensive and hence slow. The STFT is able to give comparable results at the time and frequency resolutions required in this application, while being less computationally expensive. The STFT is also more intuitive, which helps when analysing extracted features.

While the use of machine learning techniques is prevalent in categorising extracted audio features, we have taken a heuristic approach that doesn't require training, and allows the mechanics of categorisation to be understood and incrementally improved upon. This also eliminates artefacts of the training data and equipment that may create unforeseen results when transferred to other equipment and environments. In other words, the results presented here should accurately reflect the performance in a real device.

IV. Mining acoustic records

There are many ways to integrate features and classifiers. Different existing tools are used to extract relevant acoustic features identifying approaching vehicles. Relevant frequency ranges and a set of signal feature extraction methods are presented in the next sections.

A. Identifying Relevant Frequency Ranges

A spectrogram was used on recordings of passing vehicles to observe acoustic patterns. The audio recordings were viewed with Spectrogram 16, using a logarithmic frequency axis, and varying the threshold, specified in decibels (dB) to eliminate background noise.

Using the spectrogram, it was straightforward to identify the strongest relevant frequency components in the audio, that stand out from other background noise. This revealed that wind (from a passing car or otherwise) tends to create the strongest signals, showing up as a thin but strong (short time, but intense) band extending across a wide range of frequencies, but strongest at low frequencies (around 0-200Hz or so).

Vehicle sounds include:

- tires on the road, which may be particularly strong when there is gravel on the road);
- the running engine, of which changes in speed (acceleration, deceleration) and gear changes are noticeable acoustically;
- theoretically brakes should create discernible sounds, but this never occurred even when brakes were applied in tests.

A spectogram, graphed logarithmically on the frequency (Y) axis, with a cutoff at -70dB and is illustrated in Figure 1. It shows an approaching vehicle, a gap, then another approaching vehicle, which had a second vehicle close behind. Analysis of spectograms from many scenarios showed that:

- There are two (usually two to five in all our recordings) frequency bands (typically 5-10 Hz wide) of high energy

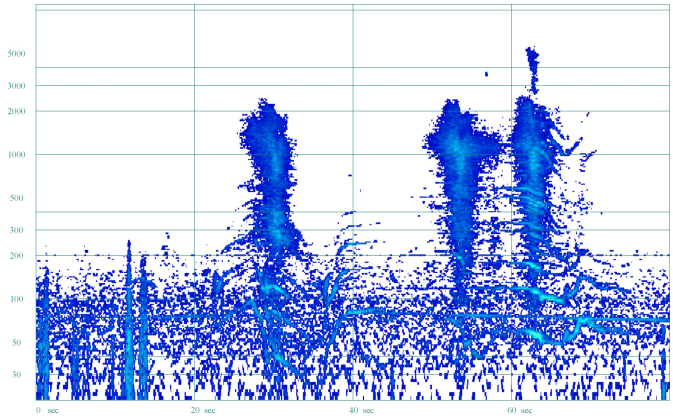


Fig. 1. Spectogram diagram: an approaching vehicle, a gap, then another approaching vehicle, which had a second vehicle close behind (scenario 39).

in the 50-200 Hz range (with shifting frequency over time) associated with the engine of an approaching or departing vehicle
- The sound of a vehicle's tyres as it passes is similar to wind, but covers a wider range of frequencies, from around 40 Hz - 3 kHz (and briefly up to around 5 kHz).
- Useful information for identifying vehicles appears to be within the frequency range extending up to around 5 kHz.
- Vehicle audio signatures are hard to generalise, and lower fidelity classification based on strength of frequency components doesn't appear sufficient for identifying a vehicle over other sound sources.

Based on these findings, subbands were chosen in a close-to-logarithmic fashion, inspired by MFCCs and human hearing models. According to [14], sounds are typically classified according to perceptual similarity, and perceptually relevant features often lead to robust classification. Therefore, perceptually logarmithmically equivalent frequency ranges, such as 100-200 Hz and 10-20 kHz, are approximately equally important.

The following frequency subbands (specified in Hz) were chosen, based on the spectographic analysis: 0-40, 41-70 71-110, 111-150, 151-200, 201-250, 251-300, 301-400, 401-500, 501-750, 751-1000, 1001-1500, 1501-2000, 2001-3000, 3001-5000, 5001-11025. Note that there is a 10 Hz granularity limit, and that for low fidelity recordings at 22050 Hz, 11025 Hz is the upper limit.

B. Adaptive Thresholding

Tests with spectrograms showed that to be able to discern useful spectral magnitude (loudness) information from an audio signal, an appropriate threshold needs to be set to filter out background noise. Unfortunately, background noise can vary considerably, from the low level noise in a park to the high levels of noise in a busy city location. This makes it impossible to specify a static threshold in dB that will work across a wide range of scenarios.

To address this, an approach was developed that adjusts over time to the loudness of the audio signal. Adaptive thresholds were developed based on the moving average and Standard Deviation (SD) of extracted signals over a window of time. Thresholds are based on the mean and SD of the frequency component magnitudes within subbands, or the entire spectrum. These can be used to eliminate background noise, while allowing sounds of interest to be identified. Thresholds are defined as:

$$\tau_i = SMA(m_i) \pm \kappa\rho\sigma(m_i) \quad (1)$$

Where i is the index in time, m_i is the mean magnitude (in dB) of the frequency components, for the spectrum or subband, at that given time. SMA is the Simple Moving Average, with addition or subtraction being used for testing above or below the moving average, and the distance from the moving average is set using the standard deviation, σ, scaled by a constant κ. Thresholding of the global spectrum was solely used for viewing the audio through a spectogram, for which κ was set to 2.5. Subband thresholding was used extensively (see Section V-F2), and in this case $kappa$ was set to 1.0. ρ is the roll off factor, defined in Section V-F1, and is set to 1 if roll off is not used to modulate the threshold.

The standard deviation can either be calculated across an entire recording, for post analysis, or for a window of audio samples in real time. In the latter case, a moving standard deviation is used. A moving standard deviation can be expressed in terms of the moving average of a set of samples, and the corresponding moving average of its squares as follows [21].

$$\sigma = \sqrt{\frac{n}{(n-1)(x2m - xm^2)}} \quad (2)$$

Where xm is the average of the last n (the window size) samples of the signal x, and $x2m$ is the average of its squares.

Prior to the first window of n samples being collected, the standard deviation of the existing samples can be used (i.e. n is set to the number of samples collected so far).

V. Experimental Analysis

This sections presents the methodology we used to build our Acoustic Hazard Detection (AHD) system. Recordings were done in many traffic (and non-traffic) locations and conditions, with vehicles approaching directly, passing in front and to the side (on both sides of the road). Different type of vehicles (2WD and 4WD), trucks and buses were present in these scenarios.

A. Methodology

The following steps has been used:

1) record audio with an iPhone from various environmental scenarios;
2) apply STFT to recorded audio signal;
3) identify useful frequency ranges, and create sub bands based on these;
4) extract and identify useful features;
5) combine features heuristically to determine vehicle presence.

The equipment used in the work presented here was:

- An Apple iPhone 3G.

- Recorder Pro (DAVA Consulting) software for recording audio in AIFF format, using 22050 Hz sampling rate.
- AIFF convert to WAV (for using with Matlab) with Winamp Nullsoft Disk Writer plugin.
- Spectogram 16 by Visualization Software LLC.
- Matlab to analyse data.

1) Recording procedure: We were not aware of any publicly available online database which we could use to test our system, therefore we have created our own set of recordings. A wide range of audio recordings were taken with an iPhone microphone so that the quality of the sound is similar to existing portable devices. The recordings covered staged vehicle drivebys, quiet suburban streets with infrequent vehicles, busy roads, a low density intersection, a high density Central Business District (CBD) intersection with buses and cars, a suburban park with no vehicle sounds, a noisy cafe, a supermarket, an underground bus station walkway with pedestrian and escalator noise, and an in vehicle recording. Staged vehicle driveby recordings lasted 15-60 seconds, while unstaged recordings lasted for several minutes. The speed of the passing vehicles and their distance to the iPhone were estimated subjectively by the person performing the recording.

A few representative recordings were chosen for examining how features correlate with audio events. In particular, recordings of a single vehicle passing were looked at first, then low level traffic on a quiet suburban road, an urban intersection, and finally a busy CBD intersection. Non-vehicle recordings were also looked at to determine how the features correlated with other environmental sounds. A noisy cafe, an underground bus station with pedestrian and escalator noise, and a quiet suburban park were chosen as representative scenarios.

To determine the utility of various features, they were calculated globally, where relevant for temporal features, and for subbands; normalised and graphed for comparison; scaled in terms of their standard deviation and graphed relative to the moving average.

B. Selection of Data Sets

The vehicle Acoustic Hazard detection (AHD) algorithm was written in Matlab. Ten representative scenarios were selected from the forty or so recordings. Nine of the scenarios ran for the first 80 seconds of audio from the selected recordings, while the other ran for the entire 66 seconds of the selected recording. The ten scenarios were:

- low traffic intersection [scenario 01];
- crossing busy road from bus stop [scenario 03];
- cafe morning [scenario 06];
- busy CBD intersection [scenario 08] ;
- underground bus station walkway [scenario 09] ;
- quiet park [scenario 11];
- suburban street in rain [scenario 13];
- supermarket [scenario 14] ;
- suburban street 1 [scenario 38] ;
- suburban street 2: representing an approaching vehicle, a gap, then another approaching vehicle, which had a second heavier vehicle close behind [scenario 39].

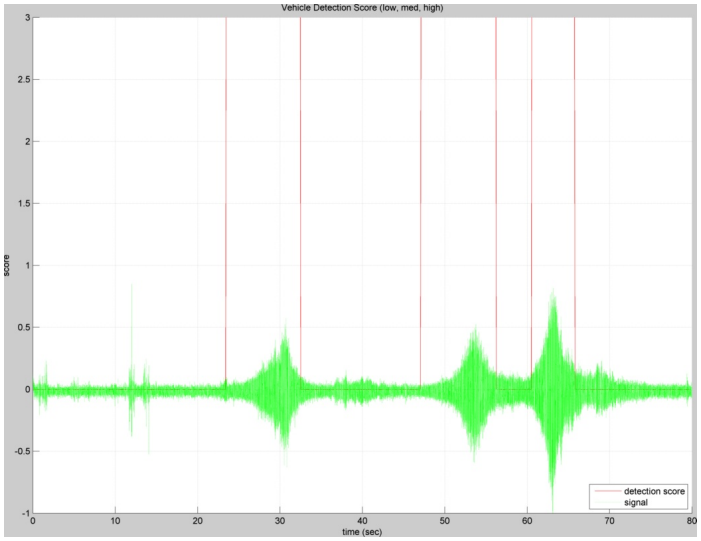


Fig. 2. Detection Score (scenario 39)

Vehicle detection scores are comprised equally of both sub band magnitude and sub band spectral features, i.e. 50% each (see Section V-F for more details). Thresholds for vehicle detection events were set as follows:

- none (score 0): below 10% activation – nothing or some partial features only, or a low magnitude event only;
- weak (score 1): $10-29\%$ activation – feature only, or magnitude event only gives scores 0-25 (both magnitude and spectral features should be present at the same time if they aren't, the score is halved);
- medium (score 2): $30-59\%$ activation – both feature and magnitude events present, but low magnitude and no vehicle immanence contribution;
- strong (score 3): $60-100\%$ activation – both feature and magnitude events present with good strength, and either immanence triggered and/or magnitude high.

In this classification, strong events correspond to a loud or close vehicle, medium corresponds to distinct vehicles, while weak correspond to background or distant vehicles. Figure 2 illustrates a scoring (red scare) for a given signal (green). It represents the signal of the three vehicles passing by in the spectrogram shown in Figure 1 (scenario 39).

Each scenario was examined by ear to identify vehicle events, and to manually categorise these as strong, medium, or weak events, so as to coincide with accurate classification of events by the AHD system. Scenarios were then tested with three variants of the AHD system: default (incorporating roll off modulated thresholds and using the standard deviation of signals over the entire recording), moving standard deviation (the same as default but with a moving standard deviation using a 30 second window), and no roll off (the same as default, but thresholds aren't adjusted according to roll off)

C. Correlating Events with Temporal and Global Spectral Measurements

We have mined the audio signal with temporal and spectral tools. Temporal features include: Root Mean Square (RMS),

Autocorrelation Coefficients, Zero Crossing Rate (ZCR), and Short Time Energy (STE). While spectral features derived from the STFT include: spectral centroid, bandwidth, skewness, kurtosis (flatness), roll off frequency, energy, band energy ratio, decrease, flux and entropy.

Amplitude, which reflects the loudness of the sound, was the first feature considered, as it can be used to indicate the strength and distance to the sound source. As expected, when graphed it shows the approach and passing of vehicles, as well as any other loud noises. Spectral energy (or equivalently RMS), which is derived from signal amplitude, shows the same effects, and when graphed together, they follow each other [15], [10] . Spectral energy, however, has the advantage of being able to be calculated for subbands, to show the loudness of sounds within the given frequency ranges. Short time energy [6] showed sharp peaks on the occurrence of wind, but otherwise followed signal amplitude (i.e. spectral energy). Spectral bandwidth [18] and spectral entropy [23] also followed the signal amplitude.

Roll off [18] has a tendency to oscillate a lot, and have many brief spikes, associated with short duration high frequency noises (e.g. bird calls, gusts of wind past the microphone). However, when roll off is smoothed with a median filter (with 1 second window), it shows changes in traffic conditions as shown in Figure 3. This graph represents the rolloff frequency (green), rolloff frequency smoothed over a 1 second window (blue) and the change in smoothed rolloff frequency over a 10 second window (red), for the traffic scenario presented in Figures 2 and 1. The figure is graphed logarithmically on the frequency (Y) axis. Figure 3 is the rolloff associated with the spectrogram shown in Figure 1 corresponding to an approaching vehicle, a gap, then another approaching vehicle, which had a 2nd vehicle close behind (scenario 39).

In a low traffic (or otherwise quiet) environment, roll off hovers around 100-300 Hz, but when a vehicle passes, roll off climbs to and plateau's around 1-2 Khz, before dropping back after the vehicle has passed.

However, in a noisy traffic environment (at an intersection for example), roll off may be continuously above 1 kHz, and actually drop when a loud, or closeby, vehicle passes, or starts from a stopped position. Roll off may also be high in noisy non-traffic environments, such as a busy cafe, but is also a lot flatter than in a traffic environment, where passing vehicles cause significant rises and falls. These findings indicate that significant and sustained changes in smoothed roll off frequency can be a good indicator of an approaching vehicle.

The spectral centroid [10], which correlates with the brightness of a sound, appeared to be a noisy signal that didn't correlate strongly with the sound of approaching vehicles as it was affected too much by other environmental sounds.

The zero-crossing rate [10] appeared to be a noisy signal, but seems to back up the roll off frequency to a large degree, with short high pitched components showing up as spikes, lower frequency sounds show up as a lower amplitude signal, and approaching and passing vehicles show up as peaks.

Spectral entropy [23] appears to largely follow signal amplitude, but with smoothing (i.e the signal doesn't drop off so fast) if higher frequency components are still present (i.e. after a vehicle has passed, or approaches).

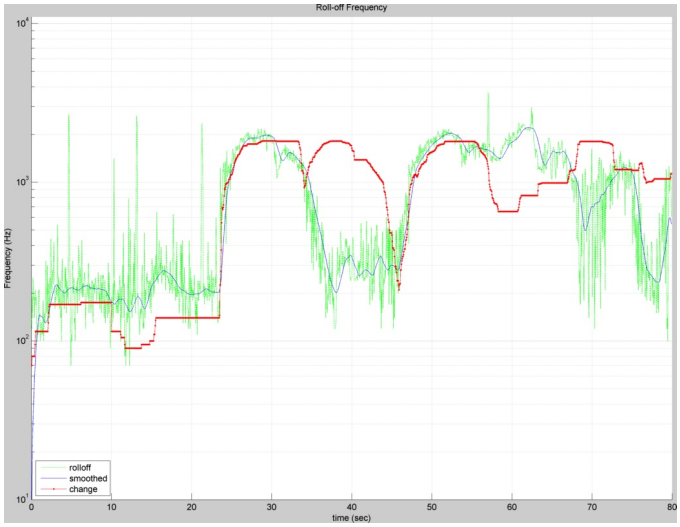


Fig. 3. Rolloff (scenario 39)

Spectral flux [18] largely appears to be a smoothed inverse of signal amplitude, dropping where amplitude increases rapidly, and reverting to baseline as amplitude returns to normal. It drops, in a smooth but rapid manner, to zero or less as a vehicle passes, creating a valley feature. While other street noises can cause flux to drop to less than zero, these often just spikes in the signal.

Spectral decrease [18] seems to correlate inversely with the recorded signal amplitude to some extent. It rises from negative to zero with loud noises, including vehicles and wind etc.

Spectral kurtosis (flatness) [12] appears to be the inverse of spectral decrease, but less noisy. It decreases and drops to around zero as a vehicle approaches, but is not as strongly affected by wind as spectral decrease, though it may also be affected by other strong noises.

Spectral skewness [12] follows flatness, to which it is closely related, but offset in amplitude slightly.

Autocorrelation coefficients [19] didn't appear to show a lot more than signal amplitude, although abrupt events show up strongly. We have used the inverse Fourier transform (IFFT) to calculate the correlation coefficients from each windowed spectral sample.

Overall, it appeared that the duration of a signal level (i.e. whether it is just a brief spike or not) needs to be considered along with the signal's level when determining what event the signal is characterising.

D. Correlating Events with Subband Measurements

Graphing of spectral energy, and equivalently mean band amplitudes, showed that wind can be seen clearly in subbands up to 500 Hz, and passing vehicles can be seen clearly in subbands from 251 Hz – 3 Khz, though there may be some noise in some bands above 1 kHz as illustrated in Figure 4 (which corresponds to the traffic scenario 39 presented in the previous graphs). It features E: energy (in black), BW: bandwidth (in green), entropy (in blue), flatness (dotted

magenta), autocorrelation coefficient zero (in red) and flux (in cyan).

The subband energy ratio wasn't as strongly correlated with the sound of a passing vehicle, however, but does show increases in low bands (0-40 Hz, 41-70 Hz) and decreases in subbands above 150 Hz, and notably low in the noisiest band (151-200 Hz) (see Figure 4 3rd graph on the right). This reflects the wide spectral spread of vehicle noise.

Bandwidth appeared to be a good indicator of vehicle activity. Subbands 251 Hz – 3 kHz increase as a vehicle approaches. Subbands 3 kHz and above also increase, but may pick up high frequency noises as well. Subbands below 251 Hz are quite noisy. Entropy shows they same behaviour as bandwidth.

Subband centroids appeared to be fairly flat and noisy in most bands, but with some movement in bands up to 200 Hz.

Flux is a weak indicator, being basically flat in most subbands, but with fairly small drops in a few bands as vehicles approach. However, flatness is a stronger indicator, dropping strongly as vehicles pass, except in the first few subbands up to 150 Hz. At other times, flatness is either high, or very noisy. Skewness also decreases as vehicles passes, mostly in bands from 251 Hz and up. However, it is a noisier signal than flatness.

Subband spectral decrease was simply too noisy to have any use in vehicle detection.

Auto correlation coefficient 0 increases in bands 251 Hz and up as a vehicle approaches. Lower bands corroborate this, but also pick up wind.

E. Choosing Features

Many of the temporal and global spectral features didn't appear useful in detecting vehicles (for example spectral centroid), and those that did (for example spectral energy) provide more detailed information in subbands. Roll off was the only global feature that appeared to provide useful information.

The most obvious feature that we have retained is signal amplitude, which indicates how close a vehicle is. There are three signal amplitude measures that are basically equivalent, these being spectral energy, RMS and STE. Of these spectral energy was chosen as the representative feature. In particular, subband energy provides information on the loudness of the sound, while allowing loud noises that sound nothing like a vehicle to be filtered out.

When a vehicle approaches energy, bandwidth, entropy, and auto correlation coefficient 0 should be high, while flatness, flux, and the absolute value of spectral decrease should be low. Many of these features are interrelated. From smoothed (one second window) graphs it appears that, ignoring offsets:

- flux follows flatness but is less detailed (flatter with just a few peaks and valleys);
- spectral decrease mirrors flatness;
- bandwidth follows spectral decrease (but is flatter);
- entropy follows bandwidth;
- entropy also follows energy to a large degree.

From these features, the strongest and least noisy are chosen as vehicle detection features: spectral energy, bandwidth and flatness. Autocorrelation coefficient 0 and flux have strong responses when an approaching vehicle is very close, so these are chosen as vehicle imminence detectors.

F. Heuristic For Approaching Vehicle Detection

From the previous section, it can be seen that many features provide the same information, so only the clearest of each of these sets of features has been utilised in the classification process.

Three sets of information are extracted from the audio:

- Subband audio signal amplitude (the mean of its constituent frequency component magnitudes).
- Subband spectral features: energy, bandwidth, flatness, flux and autocorrelation coefficient 0.
- Global spectral roll off.

The first two sets of information are thresholded, while the last is used to modulate the threshold for foreground vs background determination. Subband amplitude is used to determine if something that might be a vehicle is significantly loud to warrant attention, while subband spectral features assist in characterising the sound.

A heuristic based on these has been developed, and is outlined in the following sections.

1) Adjusting Sensitivity with Spectral Roll Off: Changes in spectral roll off frequency are used for adjusting the sensitivity of detecting activity in subband amplitude and spectral features. Roll off can be a very noisy signal with large oscillations, and so before it can be utilised, the signal needs to be smoothed to look at long term patterns. A moving average with a window of one second is used for smoothing.

After the signal has been smoothed, the logarithmic amount of change in the last ten second window is calculated as

$$roffD = \log_{10} \max(r) - \log_{10} \min(r) \tag{3}$$

where r is the smoothed rolloff values within the window. $roffD$ values typically range from less than 0.1 to above 1.0, with higher values generally associated with a vehicle passing. This measure of the logarithmic change in roll off is then used to increase or decrease sensitivity to subband amplitudes and spectral features.

A roll off factor (ρ) is then set based on $roffD$ as follows:

- roffD ≤ 0.65: set $\rho = 0.75$ to decrease the threshold, for increased likelihood that the sound is a vehicle;
- roffD between 0.35 – 0.65: set $\rho = 1.0$ to leave the threshold unaltered;
- roffD < 0.35: set $\rho = 1.5$ to increase the threshold, so reducing the likelihood that the sound is a vehicle.

Thresholds based around the standard deviation from the moving average are multiplied by $rofactor$ to scale the threshold of subband amplitude and features (see Eq. 1 in Section IV-B.

2) Subband Amplitude-based Classification: From examining the thresholded subbands, under various scenarios, it was shown that a good first pass, which allows through some false positives that can be eliminated using other measures, is to count the number (N) of "active" subbands, those having their mean amplitude above or equal to their subband amplitude

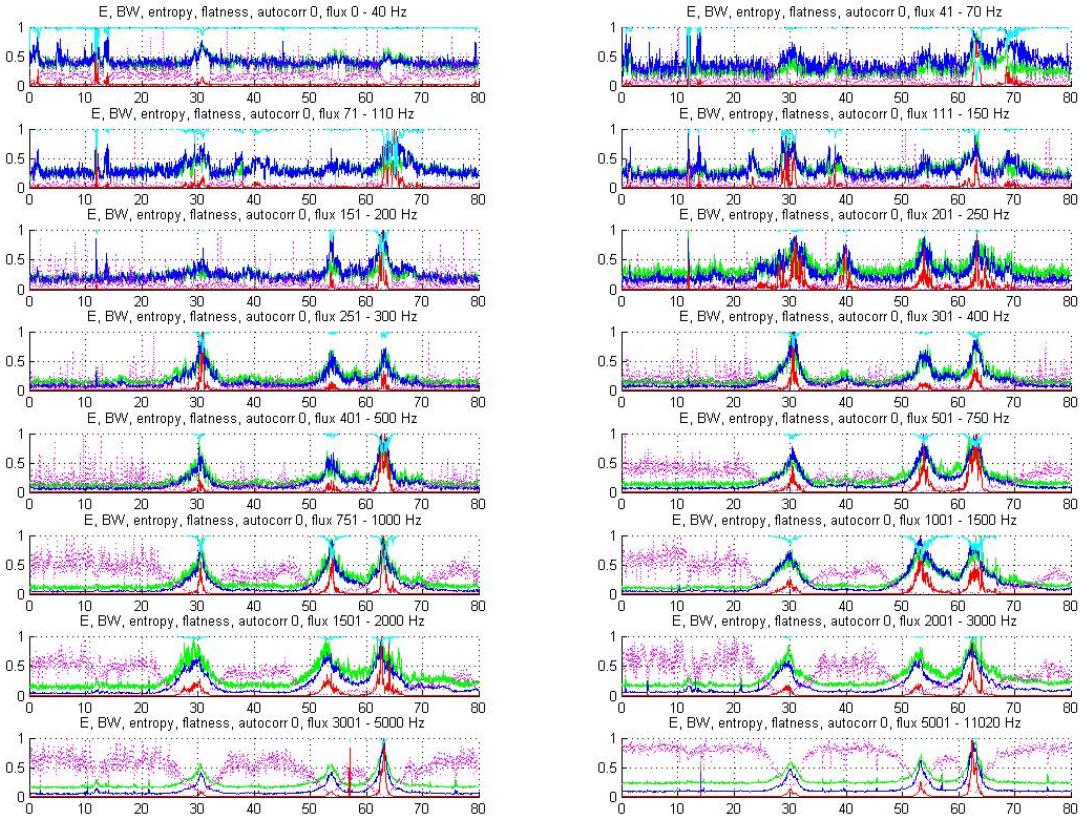


Fig. 4. Spectral energy and equivalent mean band amplitutes (scenario 39)

threshold, in bands 2-11 (41-1000Hz). Then, if band 11 is active, add the count of active subbands from 12-14 (1-3 kHz). Subbands above 3 kHz are not used here.

The top active subband (F), up to and including 14 (i.e. $\leq$ 3 kHz), is recorded, and if F $\geq$ 9 (i.e. if the top active subband is at least 401-500Hz), or if subband 1 (0-40 Hz) is inactive (indicating that the sound is unlikely to be wind), then set N to the number of active subbands with a ceiling at 10. Otherwise set N to 0. 1088c1091

This gives an instantaneous score, that taken in isolation can be a little noisy, containing brief spikes of activity. This can be improved by taking into account its history. At this point, a simple spike-removal approach has been taken, whereby spikes in the score, starting and returning to zero, are removed if their width of activity is less than 500ms. This successfully removes most false positives without adversely affecting the identification of significant vehicle activity.

3) Subband Feature-based Classification: The second part of the test for approaching vehicles uses the chosen extracted subband spectral features; the three primary features spectral energy, bandwidth and flatness; and the two imminence features Autocorrelation coefficient 0 and spectral flux.

As with subband amplitude, a thresholding approach is used to determine whether a feature is considered active or not. For any given subband of interest, a feature is considered significant or active, if it is at least $\kappa\rho$ times the standard deviation away from the moving average, in the direction of interest (i.e. above or below the average). See section IV-B for details.

For energy and bandwidth features, the main subbands of interest are from 401-2000 Hz (subbands 9-13), with optional subbands being 251-400Hz (subbands 7,8), 2 kHz - 11kHz (subbands 14-16). These two features are considered ”active” (i.e. a vehicle is detected) if either four out of the five main subbands, or three of the main subbands plus three out of the five optional subbands are at or above their subband feature threshold.

For the flatness feature, this feature is considered to be ”active” if the subband flatness is less than the moving average for most (four out of six) of the subbands $\geq$ 751 Hz (subbands 11-16).

These three features have their activity status counted to obtain a score from 0-3. Spikes (emanating from zero and then dropping back to zero) in activity with a width of activity less than 500 msec are then removed to eliminate noise

If, after spikes have been eliminated, the sum of these three features (energy, bandwidth and flatness) has a positive score, then the other two features, flux and autocorrelation coefficient 0, are tested as imminence (vehicle very close) indicators.

For the flux feature to be considered active, flux must be at or below the subband threshold for most (four out of six) subbands in the range 251 Hz - 2 kHz (subbands 7–14).

For the autocorrelation coefficient 0 feature to be considered active, it must at or above the subband threshold for most (six

out of ten) of the subbands from 251 Hz (subbands 7-16).

If both autocorrelation and flux are "active", then add these imminence detectors to the feature score to give a score from $0-5$, and then multiple this by two so as to have equal weighting to the active subband amplitude score.

4) Combining Amplitude and Feature Classifications: A total score indicating likelihood is made by combining the subband amplitude and feature scores. When a vehicle is approaching (or passing) both subband amplitude and spectral features should be present at the same time, however, thresholding based on roll off affects band amplitude more than features. Furthermore, while subband spectral features, in themselves are not sufficient for determining whether a vehicle is approaching, it will register an approaching vehicle before the vehicle's sound becomes prominent (that is, before subband amplitudes exceed the threshold).

For this reason, in an actual embedded application, if the subband amplitude score is 0, set the total score to be half the feature score, to reduce false positives, but to still have some effect. But if the subband spectral features score is 0 and the subband amplitude score is positive, then set the total score to be the amplitude score, as it has fewer false positives, and unexpected loud noises may be worth registering even if they are not from a vehicle. However, for the recognition testing experiments presented in this paper, rather than actual application, scores are done slightly differently. In this case, both band amplitude and spectral features should be present at the same, and if they aren't, then the score (from amplitude or features) is halved.

The behaviour of this heuristic is such that as a vehicle approaches, the total score will rise from zero to a peak determined by how close the vehicle comes, the vehicle size, and the refractory period associated with the roll off threshold scaling mechanism. With many passing vehicles, the first vehicle will trigger the greatest response, while following vehicles will have less impact. An isolated vehicle will trigger a stronger response than a constant stream of traffic.

Erroneous readings (false positives) are unlikely to rise to scores above 20, and are often narrow peaks that are easily filtered out. Background traffic may register with low (10-20) or medium (up to 40) scores, while out of the ordinary (relatively loud) passing vehicles or irregular traffic will register scores that peak above 50, and reach scores of 100 for isolated vehicles that pass close to the pedestrian. This total score can be used to notify the pedestrian of approaching vehicles.

G. Summary results

The results for each of these experiments is summarised by method in Table I. All type of vehicles (buses, trucks, 4-wheel drive, sedan) were all successfully detected. Each (Strong, Medium, Weak) column shows the number of events of this category matched, with trailing '+'s and '-'s indicate that an event was reported stronger or weaker than the actual event, while a '*' indicates that an event reported two categories higher than actual (i.e. a weak event reported as strong). The 'False' column gives the number false detection events, with 'w', 'm' and 's' to indicate the strength (weak, medium or strong). Parenthesis in the moving standard deviation method is used to indicate an event that occurred well before the sliding window for the moving standard deviation was filled – typically in the first 15 seconds of the recording. Our results shows that vehicle with strong and medium signal are mostly detected accurately.

TABLE I
SUMMARISED RESULTS BY METHOD

Method	Strong	Medium	Weak	False
actual	10	5	12	N/A
default	10	5–	9	8w
mov std	10	5+++(+)	7*	9w + (7w) + (1m) + [early trigger of strong vehicle alert]
no rolloff	10-	5+–	9+	9w + 1m + 1s

VI. DISCUSSION

There was a total of 786 seconds of audio tested in the experiments outlined in the previous section. All strong vehicle events were detected accurately, by all methods (though without roll off modulating the threshold, one strong event was detected as medium). Likewise all medium level events were detected, however the default method under classified two of these as weak, while the method not utilising roll off misclassified three (one as strong and two as weak), and using moving standard deviation, three were misclassified as strong, with an additional strong misclassification occurring well before the window was filled. All methods were unable to detect all of the 12 manually detected weak level events. The default method detected 9, as did the method without roll off (although one event was misclassified as medium). The method using moving standard deviation was only able to detect 7 weak events, with one of these misclassified as strong.

All methods produced false positives, although these were primarily weak level events. The default method reported 8 weak false positives, while using a moving standard deviation reported 9 weak (plus 7 weak and one medium well before the standard deviation's sliding window is filled) and an early triggering of a strong vehicle alert (where there was considerable wind before the vehicle approached), and the method that didn't modulate the thresholds based on roll off reported 9 weak, one medium and one strong false event.

Although the number of weak false positives appears significant, weak level events only postulate that the sound *may* be a distant or non-significant vehicle, and so in real application these would not be reported to the pedestrian. Furthermore, these primarily occur in situations where vehicle activity doesn't occur (for example, in a cafe), and could be removed through the incorporation of GPS information.

The misclassification of event levels by a single level is difficult to fault as level cut offs were chosen somewhat arbitrarily, and correlating medium and weak event strengths with what is heard can be difficult.

The experiment demonstrated that using roll off to modulate thresholds assists in removing false positives, especially medium and strong level false events caused by loud environmental noises, and increases the accuracy of categorisation of event levels.

However, using a moving standard deviation appears to lower the accuracy of classification level for medium and weak events, and has more difficulty in detecting weak events, which is not necessarily a bad thing in a real application, as it is detecting that they are simply background noise. However, it is possible that using a longer window would help to increase accuracy. In addition many false positives only occurred prior to the first window of samples being adequately filled. This could be countered in a real application by not activating alerts immediately (several false events occurred at the very start of recordings) and only reporting strong events in the first partial window.

VII. Potential safety impacts of AHD

The description of the human machine interface aspects of AHD is out of the scope of this paper. However we anticipated that when the portable device detects a car then it will block the current conversation or music and issue a strong auditory/vibration warning to the pedestrian. This section addresses the question about whether the detection provides enough reaction time to the pedestrian to act upon from the pedestrian safety viewpoint.

AHD is designed for an urban environment where the speed of vehicles are not too high (e.g less than 40 km/h). Brisbane city is an example of such a urban environment where the speed limit is 40 km/h. The stopping distance of a vehicle is a function of initial travel speed and can be calculated with basic laws of kinematics which is derived from Newton mechanics described as

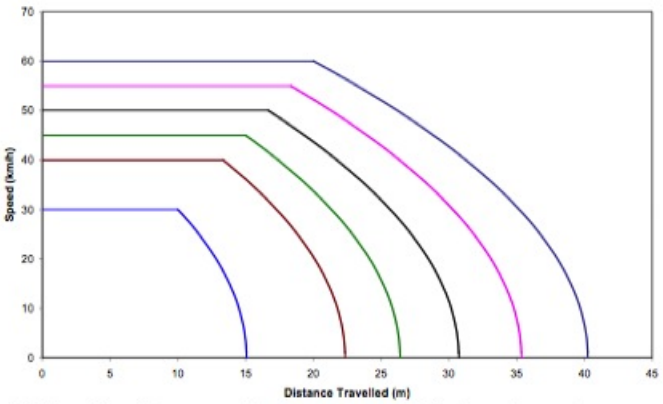
$$v^2 = u^2 + 2as \tag{4}$$

where v is final speed, u initial speed,s distance travelled, $a = \mu g$ with μ is friction coefficient and g is gravitational constant $9.8m/sec^2$. Corben et al. [8] use Equation 4 to calculate the stopping distance for different speed as illustrated in Figure 5 (a). It shows that it takes 15 m of vehicle traveling at 30 km/h to stop considering that a driver has on average 1.2 seconds of reaction time.

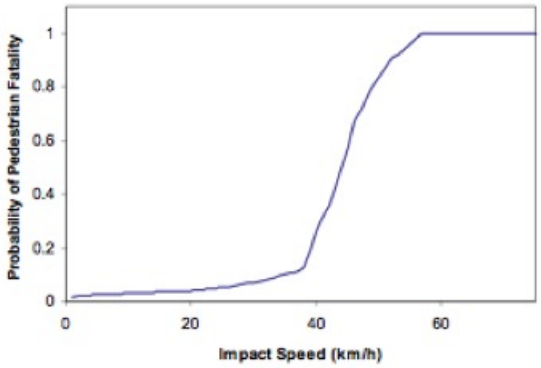
We have developed a prototype of AHD on the iPhone based on the algorithm developed in Matlab. The iPhone prototype is under development, however initial tests and estimations of the detection distance based on the Matlab algorithm show that vehicles are detected between 20 to 30 m from the pedestrian. The detection depends on the amount of background traffic noise (as obviously with more traffic it is harder to detect if a particular vehicle is of relevance to the pedestrian or not). Low level warnings could be available at around 3 seconds before the vehicle passes the pedestrian, while 2 seconds warning would be typical for a medium level warning, and a high intensity warning would typically occur around 1 second before the vehicle passes.

Assuming that a vehicle is detected at 25m and the reaction time of a pedestrian is 1.2 seconds, a car travelling at 30km/h would be 15m away from the pedestrian before she or he can reacts. The pedestrian has 1.8s left to take evasive actions. This is not much time but we could assume that (i) the pedestrian could potentially steps away or perform relevant action to avoid or reduce the crash severity (ii) the driver has started to break before the last 15m therefore reducing the likelihood of severe injury. The combination of the two actions provides sufficient time for the pedestrian to jolt back to an awareness of the traffic surrounds, and at least avoid stepping into a vehicles path, while at best, being able to take action to avoid being struck by an out of control vehicle or negligent or dangerous driver.

Figure 5 (b) shows that the probability of pedestrian fatality increases drastically other 40km/h. The increase is less pronounced under 40km/h which means that the severity of the crash decreases when the impact speed is reduced. AHD could reduce the likelihood or severity of impact by warning the pedestrian about the presence of an approaching vehicle.



(a) Stopping distance profiles for a range of Initial travel speeds



(b) The probability of a pedestrian fatality as a function of impact speed

Fig. 5. Stopping Distance and crash severity (extracted from [8])

VIII. Conclusion

In 2008, 13 per cent of the 1464 people killed on Australian roads were pedestrians, with pedestrians often being at fault. MP3 players and mobile phones are a recipe for pedestrian disaster, according to insurance company research (AAMI Insurance survey, 2010).

A novel vehicle detection method based solely on audio from a microphone has been proposed. The extraction of sound features from a noisy environment is a challenging task. Audio features were mined, selected for utility, and combined heuristically using existing statistical tools.

The main advantage of our approach is that it doesn't require heavy machine learning classification techniques which makes the processing lightweight and portable on PDAs and mobile

phones. The research results are promising as were able to detect approaching vehicles while incurring a very low rate of false positives under a wide range of environmental conditions.

Our approach has a number of limitations. The experiments conducted in this paper used an unobstructed iPhone microphone for recording audio. However, in real application, music players are often placed in a bag or pocket. This would have the effect of creating more microphone brushing noises, which in the conducted experiments showed up as weak false positives, as well as dampening many frequency components. The absence of publicly available data recordings, as a benchmark, to compare the performance of our system against others and to test the general applicability of our method prevented us from demonstrating the true advantages of our system. However the sample of data used in our experiment are realistic and general enough for a pilot study.

Our system doesn't differentiate between passing vehicles and approaching ones especially when passing vehicles are very close. Both are classified as vehicles approaching the AHD system. This could lead to a considerable number of unwanted and distracting warnings about vehicles which do not pose direct threats to the pedestrian. However the system was designed for walking pedestrians who could collide with approaching vehicles or passing vehicles should the pedestrian is stepping towards the passing vehicle's trajectory. In the later case the role of our system is to prevent the pedestrian from walking towards the passing vehicle.

Further work needs to be conducted to assess performance improvement when using a moving standard deviation, as a static recording-length standard deviation is only feasible in offline tests. This system could also be enhanced by incorporating GPS location awareness information, and motion information (from the iPhone's internal gyroscopes) to filter out situations where crash risk are very low (e.g in-door or pedestrian not moving). Our future work will concentrate on detailed validation and improvements of the AHD system which includes Human Machine Interface (HMI) studies, differentiating approaching and passing by vehicles, objective measurement of road safety benefits and further performance measures on different type of mobile phones and PDAs.

REFERENCES

[1] Silvia Allegro, Michael Bchler, Stefan Launer; Automatic Sound Classification Inspired by Auditory Scene Analysis, Eurospeech, Aalborg, Denmark, 2001.

[2] Martin Azizyan, Romit Roy Choudhury, SurroundSense: Mobile Phone Localization using Ambient Sound and Light, Poster, pp. 69–72, ACM MobiCom 2008.

[3] John Baras, Sound Classification and Localization Based on Biology Hearing Models and Multiscale Vector Quantization, DARPA Air-Coupled Acoustic Microsensors Workshop, August 24 and 25, 1999 in Crystal City, VA. http://www.darpa.mil/mto/archives/workshops/sono/index.html

[4] Judith C. Brown, Olivier Houix, Stephen McAdams; Feature dependence in the automatic identification of musical woodwind instruments, The Journal of the Acoustical Society of America 109 (3), pp. 1064–1072, March 2001.

[5] Selina Chu; Unstructured Audio Classification for Environment Recognition, Proceedings of the 23rd national conference on Artificial intelligence - Volume 3 (AAAI 2008), Chicago, Illinois, pp. 1845-1846, Association for the Advancement of Artificial Intelligence, 2008.

[6] Selina Chu, Shrikanth Narayanan, C.C. Jay Kuo; Environmental Sound Recognition with Time-Frequency Audio Features, IEEE Transactions on Audio, Speech and Language Processing, 2009.

[7] Michael Cowling; Non-Speech Environmental Sound Classification System for Autonomous Surveillance, PhD Thesis, Faculty of Information Technology, Griffith University, March 2004.

[8] Bruce Corben, Angelio D'Elia, David Healy; Estimating Pedestrian Fatal Crash Risk. Proc of road safety research policing and education 2006, Available at http://www.rsconference.com/RoadSafety/detail/676.

[9] Michael Cowling, Renate Sitte; Comparison of techniques for environmental sound recognition, Pattern Recognition Letters 24, pp. 2895–2907, Elsevier, 2003.

[10] Du Deng, Christian Simmermacher, Stephan Cranefield; Finding the right features for instrument classification of classical music, Proc of the International Worskshop on Integrating AI and Data Mining (AIDM'06), IEEE 2006.

[11] Antti Eronen "Comparison of features for musical instrument recognition". In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 19-22, WASPAA 2001.

[12] Derry FitzGerald and Jouni Paulus, Unpitched Percussion Transcription, in "Signal Processing Methods for Music Transcription", Anssi Klapuri and Manuel Davy (Ed.s), Springer Science and Business Media LLC, 2006

[13] David Gerhard; Audio signal classification: history and current techniques, Temko, C. Nadeu, Pattern Recognition, pp. 682 694, 39 (2006).

[14] Anssi Klapuri; Audio Signal Classification, Technical Report, ISMIR Graduate School, 2004. Available at http://mtg.upf.edu/ismir2004/graduateschool/people/Klapuri/classification.pdf.

[15] Mingchun Liu, Chunru Wan; A study on content-based classification and retrieval of audio database, International Symposium on Database Engineering & Applications, pp. 339–345, 2001.

[16] Johannes Luig, Feature extraction from audio signals, Lecture Notes "Digitale Audiotechnik II (Summer Term 2007)", Signal Processing and Speech Communication Lab, Graz University of Technology, Austria, April 2007. www3.spsc.tugraz.at/courses/dat2/download/Feature_extraction.pdf, accessed 2009.

[17] Martin Mckinney and Jeroen Breebaart; Features for Audio and Music Classification, Proceedings of the International Symposium on Music Information Retrieval, pp. 151–158, 2003.

[18] Geoffrey Peeters, A large set of audio features for sound description (similarity and classification) in the CUIDADO project, CUIDADO I.S.T. Project Report 2004.

[19] William H. Press, Brian P. Flannery, Saul A. Teukolsky, William T. Vetterling; NUMERICAL RECIPES IN C: THE ART OF SCIENTIFIC COMPUTING (2nd Ed.), pp 545-546, Cambridge University Press, 1992.

[20] Jouni Pohjalainen, Methods of Automatic Audio Content Classification, Licentiates Thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology, November 16, 2007.

[21] Roger Stafford, Calculating a moving standard deviation for time series, Comp.Soft-Sys.Matlab, 2006. http://newsgroups.derkeiler.com/Archive/Comp/comp.soft-sys.matlab/2006-05/msg00833.html, Frankfurt, Germany.

[22] Masaru Teranishi, Sigeru Omatu, Toshihisa Kosaka; Classification of Bill Fatigue Levels by Feature-Selected Acoustic Energy Pattern Using Competitive Neural Network, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)-Volume 6, pg. 6249, IEEE Computer Society, 2000.

[23] Aik Ming Toh, Roberto Togneri, Sven Nordholm; Spectral entropy as speech features for speech recognition, Proceedings of PEECS, 2005

[24] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti; Scream and gunshot detection and localization for audio-surveillance systems, Proceedings of The IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007), pp. 21–26, 5-7 September 2007.

Justin Lee Dr Justin Lee completed a Ph.D. at Queensland University of Technology on the topic of Morphogenetic Evolvable Hardware. Worked for National ICT Australia on a Vision-based anti-collision project for rail maintenance vehicles. Currently working as a postdoctoral fellow in Intelligent Transportation Systems at Queensland University of Technology - Australia.

Andry Rakotonirainy Associate Professor Andry Rakotonirainy directs the Intelligent Transport System - Human factor research program within Centre for Accident Research Road Safety Queensland (CARRS-Q) Queensland University of Technology - Australia.