ICSV14
Cairns • Australia
9-12 July, 2007

NEURAL TECHNIQUES FOR SOURCE IDENTIFICATION

Neil Mc Lachlan[1], Marco Paviotti[2], Stylianos Kephalopoulos[2] and Dinesh Kant Kumar[3]

[1]School of Behavioral Science, University of Melbourne,
3052, Victoria, Australia
[2]European Commission, DG JRC, 21020 Ispra, Italy
[3]School of Electronic and Computer Engineering
RMIT
Melbourne, Australia
mcln@unimelb.edu.au

Abstract

Noise source separation is a key issue in environmental noise assessment and of particular interest in the implementation of the European Environmental Noise Directive (END), since the contribution to the overall noise level from each single source should be evaluated separately. This, because concerning noise reduction measures each noise source should be eventually reduced independently from the other sources and the effect of the single noise source reduction should be readily compared to the corresponding noise of other sources, and to the health benefits. A technique which is based on wavelet analyses was tested to evaluate how far these concepts could be developed to effectively assist reaching these objectives. This technique is applied here and evaluated using noise long-term measurement data of campaigns performed in Italy in the context of the HARMONOISE and IMAGINE projects funded by the European Commission. Moreover, a proposal is made for using the same techniques to assess the physiological human response to specific noise sources. A metric is introduced which considers each single event in relation to the environment where this happens and to the subjective feeling that this might evoke in the person.

Brief time periods of the noise recordings obtained during the experimental campaigns in a real urban environment were used to test the technique. Subsequently, an attempt was done to separate the noise of the cars, motorbikes, buses, airplanes, trains and that produced by the local human voices. Characteristics of the sound signal will be shown, which are used to discern a specific sound source signal from another; these could be used in the future as alternative noise indicators.

1. INTRODUCTION

Increasing annoyance due to noise sources in Europe has prompted the Environmental Noise Directive (EU regulation, 2002/49/EC) which requires that all major environmental noise sources within urban environments be mapped. This is a challenge for both the noise mapping software developers and for noise measurements in urban environments. The first has to tackle the difficulty of complex reflections and propagation, and the correct modelling

of the noise sources, the second, have to guarantee the precision of their measurements even in complex situation with mixtures of road, railway, aircraft, industrial and other local noise sources. The measurement campaign undertaken within the European IMAGINE project showed how difficult it could be to get reliable values for several simultaneous noise sources, which may occur at very low levels in urban environment such as 50 dB or 55 dB L_{DEN} (Level day-evening-night as defined by EU regulation). Therefore, it was seen as necessary to supply measurement teams with software that is able to maximise information relative to different sources occurring in mixtures.

More research on human health assessment is a basic need for the protection of the European citizens' health, and so for the Institute for Health and Consumer Protection of the European Commission as well. Environmental noise measurements in urban environments require that the specific sources under assessment are separated from other, simultaneously occurring sources such as cars, scooters, motorbikes, air conditioning systems, people talking, birds, cars standing under the microphone, etc.

This paper then explores the possibility of developing a noise monitoring system that evaluates a metric for annoyance based on the acoustic signal and a set of parameters that relate the human response to that signal in terms of annoyance. Such a noise monitoring system will require many new features, which could both be used for noise separation and for assessment of the psycho-acoustic properties and health effect, like it is foreseen if better epidemiological studies are to be conducted [1],[2]. Firstly, the signal will need to be segmented into sound classes based on source types, as listeners will have a different psychological response to each sound source based on their social relationships to the source. This segmentation will also allow the computation of a range of psycho-acoustic properties associated with each specific sound source and the prevalence of each sound source over time to be determined.

It is proposed that annoyance may be modelled as the level of distraction caused by sound sources and the level of acceptance individuals have for each source. The level of distraction caused by a sound event will also depend on the type of cognitive activity that the individual is undertaking at the time. Distraction may be related to annoyance and to stress on the individual by considering the added cognitive load associated with cognitively evaluating each event. Changes in the subject's environment cause reflexive shifts in attention to identify the source associated with that change and in the first instance decide on whether it represents a threat or an opportunity. This reflex is activated when changes in the acoustic environment exceed that to which the subject has become habituated. Therefore the first factor we can associate with annoyance is the number of time per minute sound events other than the 'background' occur. Approaches to defining background sound will be explored in the theory section.

Some events will cause stronger arousal if an emotional relationship with the source has been previously established. The most obvious example is if the source represents a danger to the subject, but the valence and strength of the emotion will generally depend on a wide variety of factors. A statistical spread of acceptance in any population can be determined by a questionnaire on the strength of approval or disapproval of a range of common sources across a population in a similar fashion to noise annoyance surveys. However the proposed metric would use this information in relation to a specific recording rather than relying on the long term memory of subjects to recall their annoyance with respect to given sources over large lengths of time.

2. THEORY AND METHOD

2.1 Sound classification

A short description of the classification algorithm, Cyber Ear©, follows and more details can be found in earlier papers [3]. Wavelet filter banks are used as they have many features in common with the multi-resolution frequency-time filtering of the human middle ear [4]. They also have the advantage over Fast Fourier Transforms of high frequency resolution at low frequencies and high temporal resolution at higher frequencies [5]. In keeping with the findings of Gygi et al. [6], the feature vectors used for classification were the RMS amplitude and variance of short frames of wavelet coefficients over 12 decomposition levels.

Supervised back-propagation neural networks with three layers were used for feature classification. 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. 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 recordings of training and test files are time framed for analysis and the length of frame is user defined. 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.

2.2 Choice of classification features

As computational power increases finer filter channel resolution and the inclusion of features such as direction, pitch strength and height and envelope shape are likely to further extend the capability of the classification algorithm in real-time monitoring. However the choice of features needs to be considered carefully so as not to overload the NN with noisy data. For example 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 between the training and test files unless the signals are first normalized. Therefore the user should decide whether the distance of the source is likely to vary at close ranges to the microphone and use normalization accordingly. Table 1 presents a number of features that could be used for sound classification and the conditions under which they may be useful or not. It appears likely that humans are able to learn which features are constant and so specific for various sounds that they learn to recognise.

Table 1. Classification Features

Feature	**Likely to be useful**	**Not likely to be useful**
Absolute RMS amplitude	Road traffic passing close to a fixed recording location	Mobile phone rings in street

Direction	Industrial noise source	Mobile phone rings in street
Pitch strength	Human voice amongst other sources	Musical instrument classification
Pitch height	Particular alarm (fixed pitch)	Human voice or music
Envelope shape	Animal calls	Traffic monitoring
Spectral shape	Emotional arousal in speech	Speaker recognition

2.3 Background sound characterization

There are a number of issues raised by the use of the classification algorithm in environmental noise monitoring. Firstly the algorithm will need to perform with similar (or better) sensitivity and specificity as the human ear. While this may seem ambitious, the results presented here and in previous papers show that for a limited set of up to about 8 sounds the present algorithm is already able to do this. Accuracy rates of 95% have been achieved for both road traffic classification and broader environmental sound classification trials in real-time monitoring.

A second important issue for the present method of sound classification is the characterization of background sound. Sound classification performance was found to improve when undertaken with respect to the sound that is continuously present at relatively constant levels on the site at the time of recording. This sound could be added to other training sound files that were initially recorded in quiet conditions. If background sound was then included as a sound class in training the system, the NN specifically responded to differences in features due to target sound alone (since background is always present).

Transient sources were 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 for transient sounds that it has not been trained for. It is therefore crucial to develop automated algorithms to segment examples of background sound and to use them to update the NN during monitoring.

A simple model was created in Matlab using well-established empirical data from the literature [7] to better understand the acoustic properties of background sounds when they arise from many uncorrelated sources. Outdoor sound intensity is expected to decay at 3 dB with every doubling of distance to a sound source. Depending on sound attenuation factors such as humidity and vegetation, 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 1 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. In this model, 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. The summation over time of intermittent 1-second sources the same as those used in Figure 1 for light vegetation showed the background sound tended 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.

Background sound may also be characterized from a psychological perspective. Habituation is a well-known phenomenon in both vision and auditory research that results in a loss of arousal as the subject becomes more familiar with a particular stimulus or a set of stimuli that comprise their environment. It is a form of learning that happens autonomously

over a relatively short time. At a neurological level, habituation has been shown to occur in neurones in a range of brain nuclei in the auditory brain stem and even in the auditory nerve. Habituation in this context is a loss of sensitivity to repetitive stimuli both as a natural product of neural mechanics and as a result of active inhibition [8]. Background sound may then be characterised as that sound that is constant in absolute spectral levels, variability and periodicity. Sound events that are not within the usual variance from mean spectral levels cause reflexive attention and arousal. If this occurs too often then repeated arousal can lead to chronic stress [9].

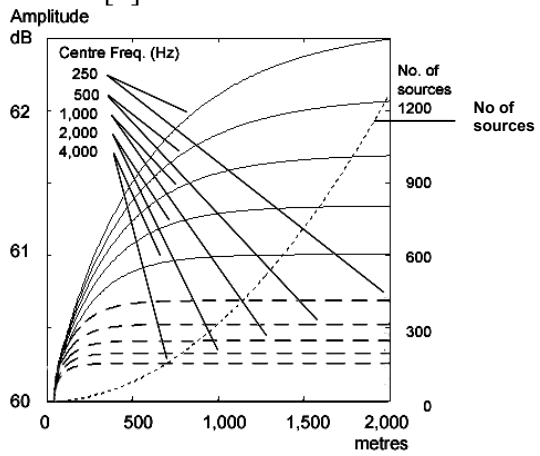

Figure 1. An approximation of the summation of white noise sources (amplitude = 65 dB when measured at 1m) when evenly distributed 100m apart up to a radius of 2km in 5 octave bands with the centre frequencies as shown. Solid lines=light vegetation and broken lines=heavy vegetation. Dotted line = number of sources within each radius.

2.4 Prevalence and annoyance

We have described a sound classification algorithm that is able to recognise the presence of examples of a dictionary of sounds in real world recordings similar to many noise monitoring situations. This paper now focuses on how such algorithms might form the basis of a new paradigm in noise monitoring. A paradigm in which a metric for annoyance is calculated for a particular acoustic environment, for an idealized listener undertaking a particular type of activity.

Results are presented showing how a classifier is able to report the time over which each sound class is likely to be audible for a given recording. This is used to calculate the prevalence of each sound class over a given time period. However, simply counting the number of events, or the total time that a particular sound class was present, does not model the number of times that a listener's attention was drawn to this sound class. For example a series of short rapid knocks should count as just one event and a long sound should not always be given more weight than one short sound. Therefore an "attentional" time window is applied to the classifier output such that 1 event is recorded if any number of events occurred in that window for each sound class. Events that are longer than the 5-second window will therefore be recorded more than once. The correct length and shape of the window for a given listener's activity will require further research, but for the present experiments an approximate length of human 'echoic' working memory of 5 seconds is used.

The number of recorded events for each sound class is then divided by the total recording time to determine the prevalence. The prevalence of each sound class can then be weighted by the mean RMS amplitude of the signal above the RMS of the background (1).

$$P(i) = N(i) \cdot \frac{\overline{\left[\log_{10}\left(\frac{RMS(a_i)}{RMS(a_b)}\right)\right]}}{T} \tag{1}$$

where P(i) is the prevalence of class i, N(i) is the number of positive classifications of class i, T is total time of recording, a_i is the amplitude of the noise event of class i, and a_b is the amplitude of the background.

Having determined the prevalence of each sound class the annoyance metric can be calculated by weighting the measured prevalence by predetermined factors describing the social, attentional and psycho-acoustic impact of sounds of this class. The calculation of these factors is beyond the scope of the present paper and will be the subject of future research. Likely approaches to calculating the psycho-acoustic factor include a combination of psycho-acoustic metrics such as roughness and sharpness.

Social factors which capture the subjects' relationship to the control of these sources could be determined from noise annoyance surveys for extensively studied sound classes such as aircraft and road and rail transport noise. Finally, attentional factors that capture the listener's mental activity will need to be included in the model. Penalties applied to long-term sound levels for measurements made at night attempt to capture listeners' increased sensitivity to noise whilst preparing for sleep (presumably rather than whilst actually sleeping). This increased sensitivity may also be due to lower noise floors at night leading to a greater impact on overall levels for any given source and more chance of it causing an attentional reflex. More detailed research on habituation to noise during a range of cognitive activities will enable better modelling of the prevalence of noise attentional reflexes and their cognitive loads. For the present experiments these three factors are varied over a small range to explore the sensitivity of the annoyance metric to them.

3. RESULTS

Figure 2 shows the successful classification of just 4 minutes of a sound recording made in the Italian city of Pisa. It should be stressed that this example is taken in the worst possible situation, when several sources contribute to the noise with overall levels which do not differ more than 3-5 dB. Examples of each sound class were manually concatenated into 45 second training files from examples of target sounds recorded at other times (apart from plane and dog where there was only one example of each). The ventilator was constantly on during the recording. A combination of this sound and traffic on a major road (200m away) comprised for most of the time the background sound. Examples of cars were vehicles passing on a closer road (50m away) that were distinguishable aurally from the background. Apart from cars the target sounds were generally easily distinguished from the background aurally. The plane was a commercial jet airliner.

Figure 3A compares the calculated loudness of each sound class relative to the background sound with the effect of multiplying this vector by the psycho-acoustic and acceptance weighting factors. Row 1 of Table 2 gives the values chosen for these factors for each of the sound classes. Note that peaks for cars are not seen in Figure 3 because the RMS values were very similar to that of the background. Only small changes in the heights of the peaks relative to other classes are observed when the weighting factors are applied apart from the class of dog barking, which is assumed to be quite sharp and tonal and to have a higher non-acceptance value. Figure 3B shows the final values of the annoyance metric computed by multiplying strength by prevalence over a 60-second moving integration window. The plane and talking classes are now the highest annoyance due to their higher maximum prevalence values (7 and 3 respectively). Figure 3B also shows that the changes in the assumed weighting factors for cars and talking according to Table 2 had very small effects on the annoyance metric compared to that of prevalence.

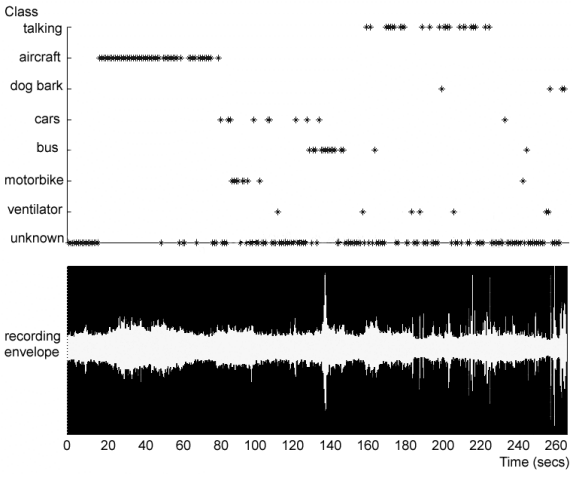

Figure 2. Classification results for an example sound recording. Each star represents a positive classification in a 1 second frame.

A **B**

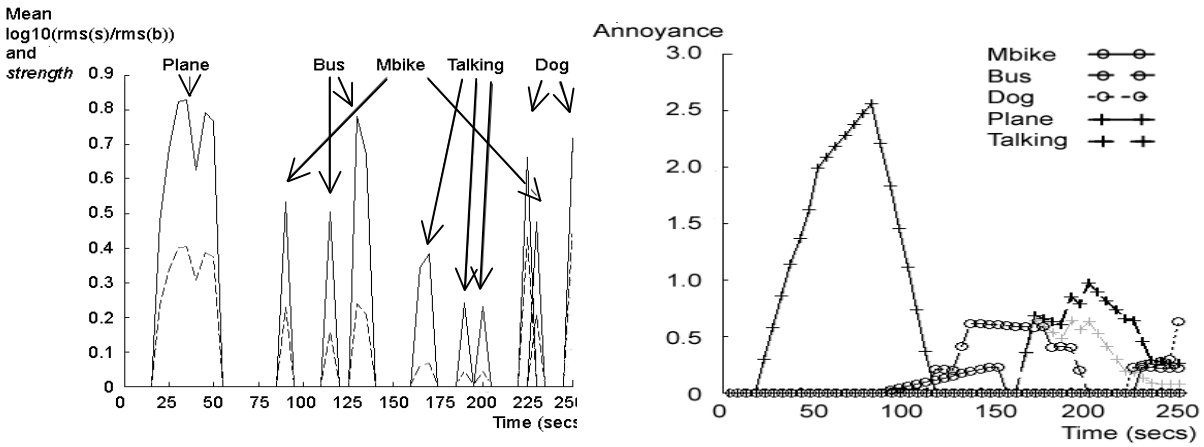

Figure 3A. Calculation of approximate loudness of each sound class relative to background (solid line) compared to strength (dashed line), in which loudness has been multiplied by psycho-acoustic and non-acceptance weighting factors. B. Annoyance for each sound class calculated by multiplying strength by prevalence (the number of 5 sec. frames containing events per minute). Two trials of the weighting factors shown in Table 2 are compared; black lines are Trial 1 and grey are Trial 2.

Table 2. Assumed weighting factors of psycho-acoustic qualities and non-acceptance.

	Factors	Mbike	Bus	Car	Dog	Plane	Talk
Trial 1	Psycho-acoustic	0.6	0.6	0.2	0.8	0.6	0.7
	Non-acceptance	0.7	0.5	0.4	0.8	0.8	0.8
Trial 2	Psycho-acoustic	0.6	0.6	0.6	0.8	0.8	0.4
	Non-acceptance	0.7	0.5	0.6	0.8	0.8	0.4

4. CONCLUSION

A hypothetical metric for the direct measurement of annoyance was proposed in this paper. The first problem in developing this metric, the ability to automatically and rapidly segment a sound recording into a library of sound classes has been shown to be largely achievable. However, there remain a number of other challenges to fully develop this metric as outlined below.

The background sound needs to be defined in terms of human habituation to the acoustic environment. Neurological models of the human auditory system are rapidly

developing and models of auditory habituation currently exist. It is likely to be possible in the near future to use these models populated by experiment data to model human habituation to natural acoustic environments and therefore calculate the background in a sound recording. This may not be possible in loud, dynamic environments, but it is not likely to be a concern under these conditions.

In the current model psycho-acoustic features such as tonality, roughness and sharpness are suggested to be important weighting factors for annoyance as they have been used in various noise monitoring paradigms. However it is possible that these features add to the annoyance of a sound because they increase the human ability to discriminate these sounds from background, and so in the present model they increase the prevalence of each sound source. This can be tested in future studies by correlating these factors with prevalence. The annoyance metric is also directly proportional to loudness above the background. This may only hold in environments where the background is not at such high levels that it acoustically interferes with activities such as speech. The proposed annoyance metric was most sensitive to prevalence as this could vary between 0 and 12, whereas the other weighting factors could only vary between 0 and 1. This is not unreasonable as the frequency of disturbance by any source should be the principle factor in annoyance. However, the correct ratio of importance between factors contributing to the metric will need to be properly determined by experimental studies. One of the strengths of the proposed annoyance metric is that the non-acceptance factors for a range of sources can be determined for a given population in one short questionnaire, allowing annoyance to be then directly measured from recordings in the environment over extended periods of time. This procedure, while performing environmental noise assessment studies in the future, might help better qualifying specific environments and performing more sophisticated epidemiological studies, but more precise as well. Having a better understanding of the noise contribution is a help in epidemiological studies, where several confounders are already present (e.g.: social status, air quality).

REFERENCES

[1] Babisch, W., “Transportation noise and cardiovascular risk – Review and syntesys of epidemiological studies”, Environmental Federal Office, Berlin, 2006.

[2] Rasche, F., “Arousal and aircraft noise – Environmental disorders of sleep and health in terms of sleep medicine”, Noise & Health, 6; 22, pp. 15-26, 2004.

[3] McLachlan, N. M., Kumar, D. K. and Becker J., “Wavelet Classification of Indoor Environmental Sound Sources”, International Journal of Wavelets, Multiresolution and Information Processing, 4 (1), p 81-96, 2006.

[4] Agerkvist, F. T., “A Time-Frequency Auditory Model Using Wavelet Packets” J. Audio Eng. Soc. 44 (1-2), pp 73-50, 1996.

[5] Szu, Harold H., Kadambe, Shubha, 'Neural Network Adaptive Wavelets for Signal Representation and Classification', Optical Engineering, 31 no. 9 pp 1907-1916, 1992

[6] 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.

[7] Smith, B. J., Peters, R. J., and Owen S., Acoustics and Noise Control, pp. 64-67, (Longman, Essex, 1985).

[8] Jorris, X. P., Schreiner, C. E., and Rees, A., ‘Neural Processing of Amplitude Modulated Sound’, Physiol. Rev., 84, pp 541-577, 2004.

[9] Spreng, M., “Cortical excitations, cortisol excretion and estimation of tolerable nightly over-flights”, Noise & Health, 4; 16, pp. 39-46, 2002.