

Context Awareness using Environmental Noise Classification

L. Ma, D.J. Smith and B.P. Milner

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

{ling.ma, dan.smith, b.milner}@uea.ac.uk

Abstract

Context-awareness is essential to the development of adaptive information systems. Environmental noise can provide a rich source of information about the current context. We describe our approach for automatically sensing and recognising noise from typical environments of daily life, such as office, car and city street. In this paper we present our hidden Markov model based noise classifier. We describe the architecture of the system, compare classification results from the system with human listening tests, and discuss open issues in environmental noise classification for mobile computing.

1. Introduction

The reduction in size of computing devices combined with the widespread deployment of mobile radio systems has resulted in an expansion of mobile computing. Users of such systems expect more and more information services to be made available to them. To deploy such a range of services it is necessary to automate these systems through the use of various input and output modalities. These may include automatic speech recognition input, text input, pointing devices and so on. To improve the usability of these information services it is desirable for the system to gain awareness of the context in which the user is in. It is further desirable to identify context implicitly rather than through the user explicitly providing contextual details. Information service offerings can then be adapted according to the user's context to provide a more intelligent set of responses. A knowledge of context can also aid automatic speech recognition or handwriting recognition performance by restricting the search space to consider only phrases or commands likely to be associated with a particular context. Up-to-date context information is especially important in the mobile computing area where the mobility of users means that context and requirements can change rapidly.

The first notions of context-aware computing were proposed in 1992 in the form of the Active Badge system [1]. The term "context-aware" was introduced in 1994 [2] with many definitions of context having been proposed since. These include location, time, weather, co-located objects, noise, recent events, etc. Recent surveys [3] [4] [5] have shown that the most common contextual inputs are location, user-identity and time. Sensors used to gather location information are mainly short range Infra-Red (IR), Radio Frequency (RF) signals and Global Positioning System (GPS). These require the use of additional hardware and processing to gain an indication of location.

A challenge for context-aware computing is to reduce the complexity of capturing, representing, processing and adapting to contextual information. In context-aware applications, contexts can be captured through many different sensors. User

feedback can also be used to confirm context and adapt the current contextual model if necessary. Services on offer can be provided according to the current context and user preferences. It may also be useful to tag context information so that it can be retrieved at a later time [6].

This work proposes the use of environmental noise as a contextual cue. Environmental noise can provide a rich source of information about the current context. For example, humans often infer the location of a respondent in a mobile phone conversation by identifying the background noise and adjusting their response accordingly. The system described in this work is able to classify a number of different noise contexts found in environments typical to daily life, such as office, bar and city street. Once the environment has been identified it can be used to predict the user's needs and adjust the mode of operation accordingly. This work forms part of a larger investigation into the integration of multiple sources of context information in a unified framework.

Previous work has focused on recognising single sound events. Couvreur [7] introduced three classifiers for use in separate noise event recognition (car, truck, airplane etc.). Gaunard [8] implemented a HMM-based classifier to recognise five noise events (car, truck, moped, aircraft, train) and observed that the frame length in noise recognition should be larger than in speech recognition. Best results came from a five-state HMM using LPC-cepstral features, which gave better results than human listeners. Much work has also been done on separating speech from background noise, following [9]. Only a few classification systems have been proposed to recognise auditory scenes. Peltonen [10] demonstrated that mel-frequency cepstral coefficients outperformed other feature representations. Sawhney [11] classified five everyday noise types, comparing several approaches, of which filterbank features outperformed others. El-Maleh [12] used four pattern recognition frameworks to design noise classification algorithms. Five commonly encountered noises in mobile telephony (car, street, babble, factory, and bus) were considered and experimental results showed that line spectral frequencies (LSFs) were best at distinguishing between the different classes of noises. Other work (e.g. [13], [14]) has been directed at recognising both the scene and sound subjects in it, focusing on identifying sound events and their relationships.

This paper describes the development and evaluation of an environmental noise classifier. Section 2 describes the design of the environmental noise classification system. Experimental results are presented in section 3 together with an analysis of several different kinds of environmental noise. In section 4 a human listening test is described. Finally a conclusion is made in section 5 together with some suggestions for further work.

2. HMM-based Noise Classification

This section describes a hidden Markov model (HMM) framework for classifying a range of different environmental noises. Classification is based on the successful approach of combining digital signal processing (DSP) technology with pattern recognition methods that has been central to progress in automatic speech recognition [15] over the last 20 years. However, it is important to note that there are significant differences between recognising speech and identifying an environment from a sample of acoustic noise. For example speech is produced from a single point source (namely the human speech production mechanism) which is reasonably well modelled. The character of sounds which speech can adopt is restricted through physical limits on the movement and rate of change of the vocal chords and vocal tract. Speech is also constrained to emanate from a single location in an environment. Environmental noise, however, has none of these constraints and is a complex sound made up from a mixture of different acoustic events. There is no constraint on the form in which these sound components take and they may emanate from many different localities in the environment. For example consider an office environment; a stationary component may come from an air conditioning fan, a quasi-stationary component from keyboard clicks, and non-stationary events from people moving around, opening doors and talking. The fan noise can be modelled quite accurately while the other components are much less stationary and need more sophisticated models.

In this paper we are only interested in modelling the slow-changing attributes of environmental noise in the acoustic signal. This means the focus is on identifying the environmental context as opposed to analysing and interpreting discrete sound events as is the case with speech recognition. The three phases which have been used to construct a set of noise models suitable for identifying the underlying environment are environmental noise database capture via portable recording devices, feature extraction and finally training and testing a set of hidden Markov models (HMMs). These experiments have been performed using the HTK (Hidden Markov Model Toolkit) developed at the Speech, Vision and Robotics Group of the Cambridge University Engineering Department (CUED) [16].

2.1. Data Collection

A high quality microphone and portable recording device were used to capture background noise samples from a range of different environments. The recordings were designed to cover everyday environments from where users would be likely to access information services from mobile devices. The recordings took place in and around the University of East Anglia (UEA) during the spring and summer of 2002. In total ten different noise environments were used to form the database and these are shown in table 1.

For each noise environment a total of 80 examples were collected. Each example was recorded as a 3 second duration audio file as this was expected to be a reasonable estimate of the likely length of noise data from which a practical system would operate. The audio was collected at a sampling frequency of 22.050kHz using 16-bit quantisation.

<i>Scene</i>	<i>Location</i>
Bar	Graduate student bar in UEA
Beach	Great Yarmouth beach
Bus	Across a range of buses
Car	Small car in urban driving
Football match	Football match at Norwich City
Laundrette	Laundrette at UEA
Lecture	Taken at UEA
Office	Wolfson Lab at UEA
Railway station	Norwich railway station
Street	Norwich city centre on a Saturday

Table 1: Description of environmental noise database

Of the 80 examples from each of the 10 noise environments, 60 were used for training and 20 for testing. Figures 1-a and 1-b show example spectrograms of two different environmental noises.

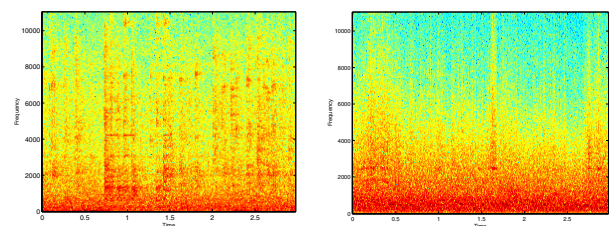


Figure 1: Spectrograms of a) Office noise, b) Bus noise

The spectrogram in figure 1-a is taken from a sample of office noise and displays several distinctive characteristics. For example the vertical lines result from the impulsive noise of keyboard clicks – a feature reasonably characteristic of an office environment. The continuous horizontal line comes from the air conditioning fan which provides another important cue to underlying environment. The spectrogram shown in figure 1-b is bus noise. The darker regions show the low frequency noise caused by the bus engine.

2.2. Feature Extraction

The purpose of feature extraction is to extract useful discriminative information from the time-domain waveform which will result in a compact set of feature vectors. Many different feature extraction techniques have been proposed [7][15][10]. One of the most successful for speech recognition applications is Mel-Frequency Cepstral Coefficients (MFCCs). As it is likely that features used for noise classification will be used by integrated systems that also perform speech recognition it was decided to adopt the same MFCC features for noise classification. This is particularly useful given the recent work on distributed speech recognition (DSR) by the European Telecommunication Standards Institute (ETSI) Aurora committee which has defined a standard MFCC-based feature extraction algorithm for implementation on mobile devices [17].

Feature extraction was performed by first pre-emphasising the audio signal and then applying a Hamming window to extract 25ms duration frames. These frames were extracted every 10ms. A 23-channel mel filterbank was applied to the resultant magnitude spectrum of the audio frame. This was transformed into a 12-D MFCC vector which was augmented by a log energy term to give a 13-D static feature vector. To improve classification accuracy both the velocity and acceleration

derivatives were computed augmented onto the feature vector. This gave a 39-D feature vector used for training and testing.

2.3. HMM-based Noise Modelling

To model the different environmental noises a set of HMMs was generated. These provide a powerful statistical method of dynamically characterizing a time-varying signal in time and frequency and have been successfully deployed in the area of speech recognition [18]. For this work a left-right topology was used. An important parameter to determine is the number of states in the model. To a certain extent this depends on the amount of training data available and the typical duration of the signal to be classified. For the time-varying nature of environment noise, more states are generally required. For the three second noise signals used in this work the number of states has been varied from 3 to 21 states.

2.4. Context-Awareness Using Noise Classification

The HMM-based environmental noise classifier has been integrated into a client-server context-aware system. The server uses the database (offline) to produce a set of noise models which are then used for classification (online). The client can communicate with the server to update the noise models. The architecture of the distributed system enables the central control on server side and lightweight processes on client side which can easily run on limited capacity devices. The current system senses two contexts, environmental noise and time. A microphone forms the environmental noise sensor and the built-in clock is the time sensor. These sensors are sampled periodically and the identified contexts continually fed into an extensible log in XML format.

3. Experimental Results

Each HMM was trained using the 60 noise examples from each of the 10 different noise scenes to give a set of 10 HMMs. A preliminary test was performed to identify the optimal number of states needed for the models. Ten different numbers of states were tested; 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 states. The noise classification performance for each of these configurations is shown in figure 2.

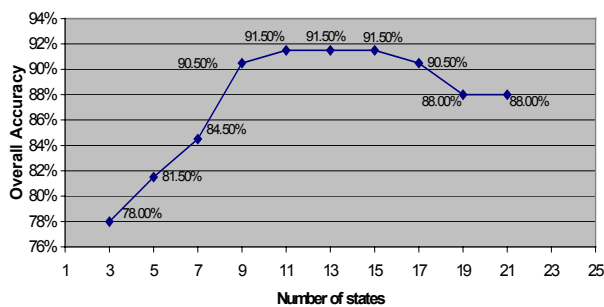


Figure 2: Overall accuracy of 10 scenes using 3 to 21 states

The figure shows that noise classification accuracy using a set of 3-state HMMs is 78.0%. As the number of states increases, classification accuracy also increases up to a peak of 91.5% with 11, 13 and 15 states. As the number of states is increased still further, classification accuracy begins to reduce. This indicates that the number of states in the HMM has a significant impact on the overall classification accuracy. In practical applications it is important to consider the likely

duration of the noise segment as this will affect the optimal number of states in the set of HMMs.

Using the result from the 11-state HMM configuration, figure 3 shows a confusion matrix which gives the individual classification accuracy of the set of noise environments.

Accuracy, %	Bar	Beach	Bus	Car	Football M.	Laundrette	Lecture	office	Rail Station	Street
Bar	85	0	15	0	0	0	0	0	0	0
Beach	0	100	0	0	0	0	0	0	0	0
Bus	0	0	95	0	5	0	0	0	0	0
Car	0	10	0	85	0	0	0	0	0	5
Football M.	0	0	0	0	100	0	0	0	0	0
Laundrette	0	0	0	0	0	100	0	0	0	0
Lecture	0	0	0	0	0	0	85	0	0	15
Office	0	0	0	0	0	0	0	100	0	0
Rail Station	0	0	0	0	0	0	0	0	90	10
Street	0	0	0	0	0	10	0	0	15	75
Overall accuracy: 91.5%										

Figure 3: Confusion matrix of noise classification

Classification accuracy ranged from 75% to 100%, with office, football match, beach and laundrette giving 100% classification accuracy for the 20 examples tested of each. Lowest performance was obtained with street noise which attained only 75% accuracy. This may be attributed to the fact that street noise is one of the most diverse noise types and may contain noise components found in other environmental categories which results in misclassification.

The accuracy of noise classification depends on a number of factors such as the amount and coverage of the training data, the feature extraction component, the allowable computational complexity, and the model parameters. The noise classifier designed in this system is not capable of recognising multiple and simultaneously occurring environmental noises. For example sitting in an office with a car passing by would cause conflict. Further problems also occur when attempting to identify acoustically similar noise scenes.

For simultaneous classification, and to get improved discrimination when classifying similar scenes, we may consider how humans solve these sorts of classification problems. Peltonen's [19] human listening test shows that humans distinguish similar scenes by identifying a number of sound components which make up the overall noise (e.g. music, clinking glasses and conversation in a bar). The short sampling strategy we use does not require a dedicated audio channel and can obtain noise samples during periods of speech inactivity. The use of a dedicated noise channel would facilitate alternative recognition strategies.

4. Human Listening Tests

A human listening test was conducted using the same test dataset in order to obtain a baseline for evaluating the classifier. The test was performed in a small listening room free from external distractions. A supervisor controlled the test. A set of 14 subjects was employed which comprised 8 males and 6 females between 21 and 45 years old, all with

normal hearing and no previous experience of this type of test. The test was presented on a PC connected to an external hi-fi amplifier and noise-cancelling headphones; playback volume was adjusted to a comfortable volume by each subject.

Each subject was required to listen to 30 randomly ordered noise samples selected from the test dataset (3 samples of each scene). The subjects listened to each sample and then decided upon the noise category from a list of the ten scenes. Subjects were not given any examples of the noise environments before the tests. They were allowed to repeat the listening and refine their answers. After completing the test, their submissions were inserted into the database for subsequent analysis. Figure 4 shows the confusion matrix for the human listening tests which gave an overall accuracy of 35.0%.

Accuracy, %	Bar	Beach	Bus	Car	Football M.	Laundrette	Lecture	office	Rail Station	Street
Bar	59.5	2.4	0	4.8	19	2.4	0	2.4	2.4	7.1
Beach	4.8	16.7	21.4	28.6	0	16.7	0	0	7.1	4.8
Bus	7.1	2.4	35.7	19	0	9.5	0	0	23.8	2.4
Car	2.4	16.7	7.1	40.5	0	14.3	0	0	11.9	7.1
Football M.	19	7.1	4.8	0	14.3	0	0	0	7.1	47.6
Laundrette	0	7.1	14.3	7.1	0	35.7	7.1	9.5	14.3	4.8
Lecture	14.3	0	2.4	4.8	14.3	7.1	35.7	0	16.7	4.8
Office	4.8	0	0	2.4	0	4.8	16.7	71.4	0	0
Rail Station	0	4.8	28.6	14.3	9.5	7.1	0	0	31	4.8
Street	52.4	4.8	7.1	2.4	7.1	11.9	0	0	4.8	9.5
Overall accuracy: 35.00%										

Figure 4: Confusion matrix of human listening test

This is considerably worse than the HMM-based system which attained 91.5%. Subjects reported that the recognition was very difficult as the samples were very short and many sounded similar. Some noises were very distinct, such as the office, while other noises proved very difficult to identify such as the street. This correlates with results from obtained from the HMM-based system.

5. Conclusions and Future Work

We have described our HMM-based environmental noise classifier which was trained on a database comprising 10 different noise environments. The overall classification accuracy of the ten environments is 91.50%. The recognition accuracy of individual scenes ranged from 75% to 100%. A human listening test was also performed on the same test set, but only yielded 35% classification accuracy. This indicates that our classifier has the advantage in recognising environmental noise by short samples.

We intend undertaking further experiments, using more samples and more scenes from different locations, to compare the result of using different features, classification methods and algorithms. We are developing context-aware system to provide a range of customisable output responses and modalities appropriate to the user's situation. To aid this we are also extending the range of input modalities and increasing the range of context data.

6. References

- [1] Want R., Hopper A., Falcao V., Gibbons J., *The Active Badge Location System*, ACM Transactions on Information Systems, 10(1) 1992.
- [2] Schilit B., Adams N., Want R., *Context-Aware Computing Applications*, IEEE Workshop on Mobile Computing Systems and Applications, 1994.
- [3] Chen G., Kotz D., *A Survey of Context-Aware Mobile Computing*, Research Dept. of Computer Science, Dartmouth College, 2000.
- [4] Dey A.K., Abowd G. D., *Towards a Better Understanding of Context and Context-Awareness*, CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness, 2000.
- [5] MIT media Lab, <http://cac.media.mit.edu:8080/contextweb/jsp/projects.jsp>
- [6] Brown P. J., *STICK-E NOTES: changing notes and contexts -- the SeSensor module and the loading of notes*, EP-odd, January 1996.
- [7] Couvreur C., *Environmental Sound Recognition: A Statistical Approach*, PhD thesis, Faculte Polytechnique de Mons, Belgium, June 1997.
- [8] Gaunard P., Mubikangiey C. G., Couvreur C. and Fontaine V., *Automatic Classification Of Environmental Noise Events By Hidden Markov Models*, Applied Acoustics, 1998.
- [9] Brown G.J. and Cooke M. P., *Computational Auditory Scene Analysis*. Computer Speech and Language, 8, pp. 297-336, 1994.
- [10] Peltonen V., Tuomi J., Klapuri A., Huopaniemi J. and Sorsa T., *Computational Auditory Scene Recognition*. In Proc. International Conference on Acoustic, Speech and Signal Processing, Orlando, Florida, May 2002.
- [11] Sawhney N., *Situational Awareness from Environmental Sounds*, 1997.
- [12] El-Maleh K., Samouelian A., Kabal P., *Frame level noise classification in mobile environments*, In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999.
- [13] Clarkson B., Sawhney N. and Pentland A., *Auditory Context Awareness via Wearable Computing*, Proc. of the 1998 Workshop on Perceptual User Interfaces (PUI'98), San Francisco, CA, USA, November 1998.
- [14] Ellis D., *Prediction-Driven Computational auditory Scene Analysis For Dense Sound Mixtures*, ESCA Workshop on Auditory Basis of Speech Perception, Keele UK, 1996.
- [15] Huang X., Acero A. and Hon H., *Spoken Language Processing*, Prentice Hall, 2001.
- [16] The HTK Book Version 3.1, Cambridge University Engineering Department, December 2001, <http://htk.eng.cam.ac.uk>.
- [17] ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression, 2000.
- [18] Rabiner L. R., *A tutorial on hidden Markov models and selected application in speech recognition*, Proc. IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.
- [19] Peltonen V.T.K., et al, *Recognition of Everyday Auditory Scenes: Potentials, Latencies and Cues*, 110th Convention Audio Engineering Society, 2001.