Acoustic environment classification

Article *in* ACM Transactions on Speech and Language Processing · July 2006

DOI: 10.1145/1149290.1149292 · Source: DBLP

Dan Smith
University of East Anglia

Acoustic Environment Classification

LING MA
University of York
and
BEN MILNER and DAN SMITH
University of East Anglia

The acoustic environment provides a rich source of information on the types of activity, communication modes, and people involved in many situations. It can be accurately classified using recordings from microphones commonly found in PDAs and other consumer devices. We describe a prototype HMM-based acoustic environment classifier incorporating an adaptive learning mechanism and a hierarchical classification model. Experimental results show that we can accurately classify a wide variety of everyday environments. We also show good results classifying single sounds, although classification accuracy is influenced by the granularity of the classification.

Categories and Subject Descriptors: H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, systems*

General Terms: Algorithms, Design, Experimentation, Human Factors

Additional Key Words and Phrases: Sound classification

1. INTRODUCTION

The acoustic environment or background noise is a rich and underexploited source of information. We present an adaptive model and experimental results for identifying and classifying acoustic environments and outline some ways in which this can be exploited, primarily in context-aware computing systems and video classification.

We are interested in obtaining and exploiting location and activity information through the analysis of the acoustic environment. This is a rich information source for deriving information about a user's current situation, potentially enhancing the description of a location and a user's activities. It is almost always available and can be readily sampled by a wide range of devices.

Authors' addresses: L. Ma, Department of Computer Science University of York, Heslington, York, YO10 5DD, UK; D. Smith, B. Milner, School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK; email: djs@cmp.uea.ac.uk.

In order to test the feasibility of acoustic environment recognition in mobile devices, we have developed and evaluated a Hidden Markov Model (HMM) classifier. Our initial work intended to classify the typical environments of our daily life, such as an office, car, and street has been reported elsewhere [Ma et al. 2003a, 2003b]. In this article, we summarize those experiments and report further experiments which address requirements for building applications on limited capability devices and single sourced sound event classification. We also establish confidence measures and mechanisms for adaptive learning of new environments.

Acoustic environments can be accurately classified using sound samples acquired from widely available consumer devices. Applications that exploit context information from acoustic environments can be built on current PDAs as part of a wide mobile computing infrastructure. In the low bandwidth experiments described here, we classified 12 kinds of acoustic environment based on a user's daily routine. By classifying short duration (3 second) samples, using Mel-Feature Cepestral Coefficient (MFCC) features and a HMM classifier, the application can rapidly recognize an environment. Recognition based on lower-quality data (8kHz, 8-Bit PCM) is useful for limited capability devices and low bandwidth communications. Our overall accuracy in each of these series of experiments varies between 93% and 96%. The complete database of our sound recordings is available for research use at http://www.cmp.uea.ac.uk/Research/noise_db.

The remainder of this article is organized as follows. In Section 2, we describe related work in environmental noise recognition, video indexing, and context-aware computing. Our classification experiments are described in Sections 3 and 4. In Section 5, we describe our adaptive learning approach and, in Section 6, the sound event model. Section 7 describes our application. Our conclusions are presented in Section 8.

2. RELATED WORK

Much work on environmental noise has been done to separate speech from background noise for robust speech recognition, but few context-aware applications have attempted to use environmental noise. Work at MIT [Sawhney 1997, Clarkson et al. 1998] concentrated on distinguishing 15-second environmental sound samples into six predefined classes, using near-real time classification techniques and an HMM-based structure that relied on unsupervised training for segmentation of acoustic environments.

Many research approaches on audio data can be explored from the perspective of content-based audio classification or audio information retrieval. Foote [1999] and Wold et al. [1996] reviewed audio classification systems using multiple features and a variety of classification techniques including static modeling and dynamic modeling. Content-based audio classification and retrieval research has been mainly based on speech and music, focusing on searching and indexing. Much work has been done on speech recognition, word spotting, speaker identification, speech interface, and music information retrieval.

MFCCs are the most common feature representation for nonspeech audio recognition. Peltonen et al. [2002] implemented a system for recognizing 17-sound events using 11 features individually and obtained best results with the MFCC features. In a comparison of several feature sets, MFCCs perform well [Tzanetakis and Cook 2002], although the classification granularity affects the relative importance of different feature sets. In the TREC-10 video track, IBM obtained highly-ranked results using a 24-element MFCC (including the energy coefficient) feature vector for audio [Smith et al. 2003]. Other work [Sawhney 1997] classified five everyday noise types, comparing several approaches, of which a filterbank frontend with a Nearest Neighbor classifier clearly outperformed the others. Cai et al. [2003] used a combination of statistical features and labels describing the energy envelope, harmonicity, and pitch contour for each sample. Wu et al. [2003] also used a combination of features. CueVideo [Srinivasen et al. 1999] used Zero Crossing Rates and Gaunard et al. [1998] used LPC-cepstral features. None of these representations show clear performance or conceptual advantages over MFCCs.

Nonadaptive statistical classifiers have a major drawback in their sensitivity to variations in utilization conditions; classifiers that adapt to changes in spectra can solve this problem, while HMM-based classifiers provide a dynamic solution. The Differential Swap Rate algorithm [Bakker and Lew 2002] gives good results classifying movie soundtrack features and recorded music samples. Its misclassification rate for automobile sounds, which is the most direct point of comparison with our approach, is 10% which is better than the performance reported by Ma et al. [2003] but worse than later results. Gaussian Mixture Models (GMMs) have been used by Wu et al. [2003] and Scheirer and Slaney [1997]. A neural net model trained with 44.1kHz samples from small sets of sound events from the RWCP database [Nishiura et al. 2003] and a room environment performed well in experiments focusing principally on collision sounds [Toyoda et al. 2004]. Couvreur and Laniray [2004] used a hybrid HMM/MLP model to detect passing scooters and car horn sounds in an urban environment with up to 95% accuracy.

A HMM classifier which recognizes five noise events (car, truck, moped, aircraft, and train) with better performance than human listeners is described by Gaunard et al. [1998] Nishiura et al. [2003] have classified 92 kinds of environmental sounds using a HMM framework for robust speech recognition. Their experimental results show how the HMM can accurately identify and classify single-occurrence environmental sound sources, speech, and speech with sounds.

A variety of approaches have been used to distinguish speech from music [Vendrig et al. 2003; Wu et al. 2003; Scheirer and Slaney 1997]. Work at Microsoft Asia with 457 environmental noise samples of between 2-8 seconds duration gave an average confusion rate between 12 types of environmental noise speech of 5% and 3.5% with music, noting that discriminating environmental sounds from speech and music is challenging [Liu et al. 2001]. In other experiments based on 1-second samples, they reported 84% accuracy in identifying environmental sound, confusing environmental sound samples with speech in 10% and with music 5% of the cases; their performance in correctly identifying speech and music was better (97% and 93%, respectively) [Liu et al. 2002].

The CLIPS group in TRECVid 2002 obtained better than average results using separate GMMs for speech and acoustic environment recognition [Quénot et al. 2003]. Tzanetakis and Cook [2002] have used an audio hierarchy of genres to classify short samples of music, speech, and sports commentary (i.e. speech with strong background noise). The accuracy of their classification is heavily influenced by the musical heterogeneity of the categories, achieving an overall average of 61%. Comparing the contribution of several feature sets to the overall classification, MFCCs performed well, although there are differences in the relative importance of different feature sets in classification tasks of different granularities. Srinivasen et al. [1999] reported classification accuracy of over 80% using an approach based on Zero Crossing Rates, although the principal goal of the CueVideo system was to identify useful audio features for MPEG-7 description schemes. Work in the MPEG compressed domain for speech detection, music detection, and genre classification is described in Browne et al. [2003].

Cai et al. [2003] propose an audio classification based on a combination of statistical features and labels describing the energy envelope, harmonicity, and pitch contour for each sample in a database containing approximately 600 high quality samples ranging from 1 second to 30 seconds in duration. A similarity measure is calculated from a combination of the labels and statistical features. The labels were most effective in improving recall beyond the top 20 results where the average improvement was approximately 25%. Recall at 100 samples retrieved was 90%. Their approach is the opposite to our hierarchical model described in Section 6 as they perform an initial broad classification before attempting a precise identification. Our approach makes the most precise identification first and only uses the higher levels of the hierarchy when the initial assignment is uncertain. The most direct points of comparison in the work described with our work suggest that our current approach performs better in many situations.

A number of projects have experimented with audio for context-aware computing. ComMotion [Marmasse and Schamdt 2000] is a location-aware computing environment using GPS, but the core set of reminder creation and retrieval can be managed completely by speech. The Nomadic Radio project [Sawhney and Schamdt 2000] developed interaction techniques for managing voice and text-based messages in a nomadic environment. It employs an auditory I/O, with speech and nonspeech audio, for navigating and delivering messages. Gerasimov and Bender [2000] describe using sound for device-to-device and device-to-human communication and have explored the possibility of using the existing audio capability in many commonplace devices. Audio Aura [Mynatt et al. 1998] explored the use of background auditory cues to provide information coupled with people's location in the workplace. The Everywhere Messaging project [Schmandt et al. 2000] considered aspects of message delivery including minimizing interruption, adaptation to the user, location awareness, and unintrusive user interfaces. Early work by Hindus and Schmandt [1992] focused on ubiquitous audio to support acquiring and accessing the contents of office discussions and telephone calls.

The work described in this section shows that a variety of approaches to environmental noise classification have been tried with varying degrees of success. The relative immaturity of research in this area is typified by the poorer than expected results in TRECVid [Smeaton and Over 2003].

3. INITIAL SOUND CLASSIFICATION

In this section, we describe a HMM framework for classifying a range of different acoustic environments. Classification is based on combining digital signal processing (DSP) technology with pattern recognition methods that have been central to progress in automatic speech recognition [Huang et al. 2001] over the last 20 years. However, there are important differences between recognizing speech and identifying acoustic environments: notably, speech is produced from a single source, constrained to a single location, which has limitations on the character of sound it can produce. An acoustic environment, however, has none of these constraints and is a complex sound made up of a mixture of different events. For example, consider an office environment: a stationary or quasistationary component may come from air conditioning fans and keyclicks, and nonstationary events from people moving around, opening doors, and talking. Every environment has its characteristic consistent and periodic background noise. In this section, we concentrate on modeling all the attributes of the acoustic environment in the audio signal as our primary focus is on the overall acoustic environment. In Section 6, we describe our work on single-source sound events.

There are four phases in the construction of a set of acoustic environment models suitable for classification. These are first, acoustic environment database capture (via portable recording devices), second preprocessing (digitization, segmentation, and labeling), third feature extraction, and, finally, training and testing a set of HMMs.

3.1 Acoustic Environment Categories

The classification of acoustic environments has a number of inherent difficulties. First, to be useful, the classification must reflect differences which are meaningful for a particular application. A way to think about this problem is to look at the perceptual organization of real world sounds based on both acoustic properties and human labels. Second, the classes chosen should be sufficiently distinct that the classifier can achieve an acceptable accuracy. Third, within each class the sounds should be sufficiently homogeneous to allow a classifier to be trained without requiring infeasibly large volumes of data.

Every environment can be viewed as having a characteristic consistent background noise and periodic or quasiperiodic sounds which are often the primary cues for recognizing the environment; discrete sound events can then be seen as intrusions or interruptions to the acoustic environment. At any given scale, the background noise in an acoustic environment is composed of sound events at finer granularities. For instance, we may regard the engine noise from a passing car in the street as a sound event, but to an automotive diagnostic system, that same noise is in itself an environment which contains many lower-level recognizable types of sound event. At another scale, many examples of engine noise may be merged to create the background hum of a busy road.

From the review of sound classification applications, we can summarize that the choice of the categories really depends on the objectives of the application. The categories can be defined in several ways such as the environment source or the environment name. The primary motivation for the work discussed in this article is to investigate the use of acoustic environment information in context-aware computing. Therefore our focus is the acoustic environment of people's daily activities, and the acoustic environments used in our experiments have been chosen as representative of typical daily activities rather than for their acoustic properties.

3.2 Data Collection, Preprocessing and Feature Extraction

A high quality microphone and portable recorder were used to capture the audio examples from a range of different environments. The recordings, made in and around Norwich during the spring and summer of 2002, covered a range of everyday environments where users would likely to be accessing information services from mobile devices. For these experiments, each type of acoustic environment was recorded in a single location. Based on the length and quality of the recordings, 10 different acoustic environments were chosen for the initial classification experiments. These are office, lecture, bus, urban driving, railway station, beach, bar, laundrette, soccer match, and city center street. To these we added silence, the state in which there is no input signal (e.g., microphone switched off).

Following data collection, the audio data was digitized and divided into short duration segments. Digitization was achieved by sampling the analogue data at 22.050kHz using 16-bit quantization. The recordings were then partitioned into 3-second isolated audio segments. This duration was chosen as it is likely to be the length of noise data from which the system would operate in practical applications using a single audio channel for all functions. Associated with each of these 3-second audio segments is a manually created label that indicates which of the environments the data was collected in. For each environment, this gave a set of 80 examples of which 60 were used for training and 20 for testing. Therefore, for the 10 environments and silence, 880 noise examples were used for the initial experiments (660 examples for training and a another 220 for testing). Spectrograms from some of these environments are presented in Ma et al. [2003a]; many of the salient features can also be seen in the spectrograms presented in Section 4.

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. We conducted a series of preliminary experiments comparing linear predictive coding (LPC), filterbank, and MFCCs in which MFCCs clearly outperformed the other methods; these results are consistent with experiments performed elsewhere and summarized in Section 2. MFCC features have also been adopted as the standard feature for distributed speech recognition (DSR) on mobile devices through the European Telecommunications Standards Institute (ETSI). This means that future deployments of environmental classification could be incorporated into such a DSR architecture. For the experiments in this work, the audio signal was first preemphasized and then a Hamming window used to extract 25ms duration frames of audio. These frames were extracted every 10 ms. From the magnitude spectrum of this frame, a 23-channel mel filterbank was applied. Alogarithm, Discrete Cosine Transform (DCT) and truncation converts this into a 12-dimensional MFCC vector. This was augmented by a log energy term to give a 13-dimensional static feature vector. To improve performance, both the velocity and acceleration were computed and this resulted in a 39-dimensional feature vector.

3.3 HMM Training and Testing

The training and testing of the acoustic environment HMMs has been performed using HTK [Young et al. 2001]. Within each state of the HMMs, a single Gaussian mixture component is used together with a diagonal covariance matrix. The assumption of a diagonal covariance matrix is justified through the use of MFCC features which are decorrelated by the DCT in the feature extraction process. A HMM was trained for each of the 10 environments using the 60 training data examples available in each class. The HMMs were trained in isolation with initialization performed using Viterbi segmentation and state-based clustering. The models were then refined using Baum-Welch re-estimation applied to each model separately.

An initial experiment was performed to determine a suitable topology for the HMM by comparing the classification attained by first a 9 state left-to-right HMM and then a 9 state ergodic HMM. The ergodic HMM attained a classification accuracy of 71%, while the left-to-right HMM achieved 92%. This substantial increase in accuracy using the left-to-right topology may be explained by the observation that many of the sound events in these acoustic environments have a distinct temporal structure. The left-to-right HMM is more effective at modeling this temporal evolution of sounds, while the less restrictive structure of the ergodic HMM makes the model less effective at following temporal patterns. Based on these results, the left-to-right HMM was selected as the HMM topology for all remaining experiments.

The next experiment was designed to investigate the effect that different numbers of states have on classification accuracy. A series of experiments was performed using from 1 to 19 emitting states in the left-to-right HMMs. Figure 1 shows the classification accuracy for each different number of states in the HMMs. Using only 1 state, the classification accuracy is 80% but this increases rapidly as more states are added, allowing more detailed modeling of the temporal structure of the signal. Classification accuracy is broadly constant from 7 to 15 states, with a slight peak at 9 states where accuracy is 92%. Beyond 15 states classification accuracy begins to fall. This is likely to come from both state-based data sparseness as data is shared among more states and also from the models becoming too constrained in terms of their temporal structure.

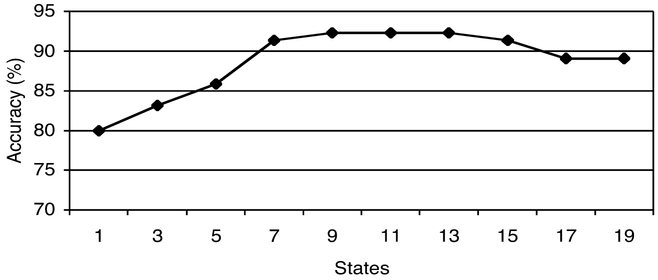


Fig. 1. Acoustic environment classification using left-to-right HMMs with 1 to 19 states.

Table I. Confusion Matrix of Acoustic Environment Classes

Results and Confusion Matrix for 10 Environments										
Accuracy %	Bus	Office	Soccer Match	Bar	Beach	Car	Laundrette	Rail Station	Street	Lecture
Bus	95	0	5	0	0	0	0	0	0	0
Office	0	100	0	0	0	0	0	0	0	0
Soccer Match	0	0	100	0	0	0	0	0	0	0
Bar	15	0	0	85	0	0	0	0	0	0
Beach	0	0	0	0	100	0	0	0	0	0
Car	0	0	0	0	10	85	0	0	5	0
Laundrette	0	0	0	0	0	0	100	0	0	0
Rail Station	0	0	0	0	0	0	0	90	10	0
Street	0	0	0	0	0	0	10	15	75	0
Lecture	0	0	0	0	0	0	0	0	15	85
Overall accuracy: 92% Overall accuracy (including silence):92%										

The classification accuracy within each of the sound environments using the 9 state HMMs is shown in the confusion matrix in Table I. Classification accuracies range from 75% to 100%, with the office, soccer match, beach, and laundrette giving 100% classification accuracy. The worst performance was obtained for identifying the street noise which is confused with the laundrette and railway station. This may be explained by the wide variety of sounds which were collected in the street sample which gives the model a large variance that causes other sounds to be incorrectly classified as street noise.

3.4 Human Listening Tests

A human listening test was conducted to find out how well humans are able to recognize environments based on short audio recordings with no prior training in classifying the sounds, using the same test dataset as the experiments described previously. The subjects were eight males and six females, all aged between 21 and 45 years old, with normal hearing and no previous experience of this type of test. The test was performed in a soundproof room with playback adjusted to a comfortable volume by each subject. Subjects were told the 10 acoustic environment categories but they received no prior audio examples of these sounds before the test. During the test, each subject was required to listen to 30 randomly ordered noise samples selected from the test dataset.

Following the test, the subjects were interviewed to discover the cues they used for classification.

The human listening tests gave an overall accuracy of 35%. The classification accuracy for individual environments ranged from 10% to 71%. This is considerably worse than the HMM-based system which attained 92% overall accuracy and shows that humans have difficulty recognizing environments when restricted to short duration samples without any prior training. Subjects reported that the classification was very difficult as the samples were very short (only 3 seconds) and that many sounded similar. Some acoustic environments were very distinct such as the office, while others such as the street noise proved very difficult to identify. This correlates with results obtained from the HMM-based system. The subjects were asked to describe what cues they used in the recognition process. The cues were most often described in terms of familiar sound sources or events such as the sharp brake noise from the bus environment or speech in the lecture environment. Humans are oriented towards analyzing acoustic environments into distinct sound sources. That is, human classification is often not based on the global properties of the signal but on some specific cues. This suggests that the addition of sound event classification to acoustic environment classifiers may allow a larger number of acoustic environment classes to be recognized.

3.5 Discussion

The initial HMM-based acoustic environment classifier has been shown to give satisfactory results for the environments tested. However, these experiments have two important limitations. First, the examples from each type of environment have limited variability as they were made from single recordings which made the recognition task easier. Second, the database is too small to allow the evaluation of recognition algorithms that require large amounts of training data.

In this series of experiments, we evaluated only continuous HMMs for the classification of acoustic environments and implemented only single Gaussian HMMs. In the speech recognition literature, it has been found that Gaussian mixture HMMs usually outperform single HMMs when properly trained. However, more complex Gaussian mixture HMMs require substantially greater amounts of training data than are available for acoustic environments and sound events. Our solution is a trade-off between performance and the computational cost.

HMM-based classifiers take into account the temporal structure of the signal and use different states to model different stationary parts of the signal. Ergodic (fully connected) HMMs and left-to-right HMMs are two candidate topologies to model acoustic environments. Theoretically, left-to-right models are suited to model signals with well-defined temporal structure and ergodic HMMs are better suited to model signals with a less well-defined temporal structure. In choosing the model topology, we must consider the structure of the acoustic environment signal to find whether it has a particular temporal signature or not. In our experiments, there are large variances between the different acoustic environments. For example, the noise of a busy street with cars passing by has a readily identifiable temporal structure, but some environments with only industrial background noises are composed mostly of stationary sounds. The segmentation also affects the structure of the signal; we have taken a long recording and chopped it into 3-second examples at equal time intervals. We used this segmentation method because, in practice, the recorded 3-second signal could be taken at any time. This segmentation does no harm to stationary and quasistationary signals but may not allow longer nonstationary signals to exhibit the same inherent temporal structure. Left-to-right HMMs can encompass the time-ordering structure even if there are some random variations and can also model stationary signals in which fewer states are required. Moreover, ergodic HMMs require very careful initialization to ensure that different states are initialized with different stationary parts in the signal. Since the target application domain constrains us to analyze short duration signals, it is not possible to perform this kind of initialization, making left-to-right HMMs the better choice.

This hypothesis is supported by our experimental results. The left-to-right topology yields better results than the ergodic topology. Better classification results are obtained when more than one state are used in the HMMs and when shorter analysis frames are used for the feature extraction. It should be noted that many apparently stationary sounds, such as keyclicks [Zhuang et al. 2005] and the majority of sounds in the RWCP database, have a temporal structure. The acoustic complexity of different environments varies considerably and using left-to-right HMMs, it is possible to have a different number of states to model different environments.

For simultaneous classification and to get improved discrimination when classifying similar environments, we may consider how humans solve these problems. Peltonen et al.'s [2001] human listening test shows that humans distinguish similar environments by catching pieces of sound events in them (e.g. music, clinking glasses, and conversation in a bar) and distinct sound events have been successfully modeled by Gaunard et al. [1998] and Peltonen et al. [2002]. The short sampling strategy we use does not require a dedicated audio channel and can obtain noise samples during periods of speech inactivity. A dedicated sound recognition channel would facilitate alternative recognition strategies.

4. LOW BANDWIDTH EXPERIMENTS

To deploy context tracking applications on mobile devices, three issues must be addressed. First, since users are limited to low bandwidth connections because of communications infrastructure and small footprint devices (mobile phones and PDAs), we need to implement and test a HMM-based acoustic environment classifier that is appropriate for low-bandwidth connections. Second, the classifier is designed for real-time processing devices which typically have only low-quality recording facilities. Third, the environments to be classified are designed to cover a range of everyday situations where users would be accessing information services from mobile devices. The experiments described in the

Table II. List of Sounds Collected from a Range of 12 Different Environments

No	Routine	Environment	No	Routine	Environment
1	Walk to bus stop	Street (traffic)	7	Shopping in mall	Shopping mall
2	Take bus to office	Bus	8	Walk in city	Street (people)
3	Pass a building site	Building site	9	Shopping in supermarket	Supermarket
4	Work in office	Office	10	Laundrette	Laundrette
5	Listen to a presentation	Presentation	11	Driving (long distance)	Car (highway)
6	Urban driving	Car (city)	12	Local or express train	Train

remainder of this section are designed to test the feasibility and limitations of using a PDA for acoustic environment tracking. Our focus is on modelling slow-changing attributes of the environmental noise in the audio signal. Some of the issues in recognizing discrete sound events are discussed in Section 6.

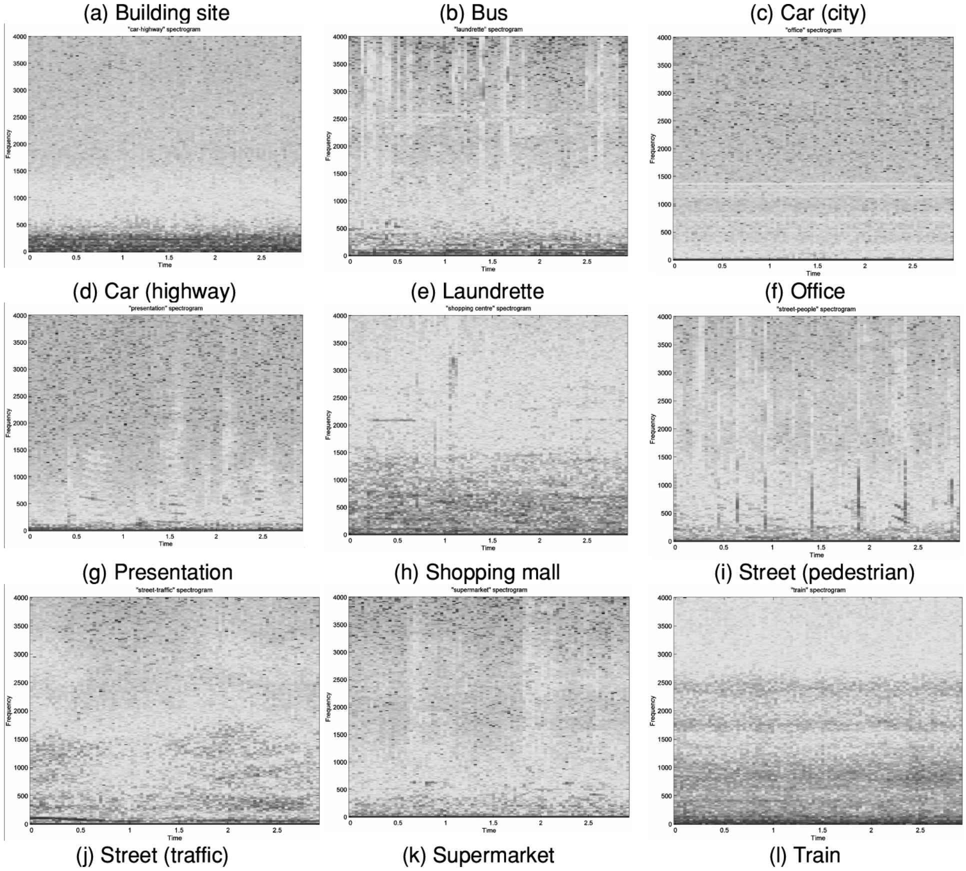
4.1 Low-Bandwidth Data Collection

Data collection was made using an MP3 player/recorder (with the resulting audio sampled at 8kHz.) attached to the strap of a shoulder bag as the recording device to capture the acoustic environments from typical daily activities. The recordings were made in and around Norwich over three months during the spring and summer of 2004. Sounds from 12 different environments were collected as shown in Table II. For each environment, several recording sessions were conducted over a range of different times and locations.

The data were partitioned into three sets, collected on different occasions. Each set contained 100 3-second isolated examples of each of the 12 environments to give a total of 1,200 examples. Associated with each audio example was a label file which indicated the environment from which the data was collected. For classification experiments, Dataset 1 was used for training, while Dataset 2 was used for testing. Dataset 3 is reserved for further experiments in Section 5 which examine the effectiveness of confidence measures. The distinctive features of the different environments are seen in Figure 2, which shows 3-second duration spectrograms taken from each of the 12 sounds. For example, the office (Figure 2f) is characterized by dominant tones at 700Hz and 1400Hz which are produced by an air conditioning fan. Even relatively similar sounding environments such as the two car environments (city and highway) and the bus are shown to be distinct. The energy of the bus noise is much higher than that of the car, and the car in the city is associated with short duration street-type noises in comparison to the more stationary noise encountered on the highway.

4.2 Low Bandwidth Classification Experiments

This section examines the accuracy of acoustic environment classification using the low bandwidth data. From each of the 3-second duration audio files, MFCC vectors (augmented with velocity and acceleration derivatives) were extracted as described in Section 3. A set of 12 HMMs, corresponding to the 12 environments, was trained from the 1,200 sound examples in Dataset 1. The HMMs used the 9 stage, left-to-right topology that gave the best results in the experiments discussed in Section 3. Dataset 2 was then used for testing which gave an overall classification accuracy of 96% on the 1,200 samples. This is comparable


(a) Building site

(b) Bus

(c) Car (city)

(d) Car (highway)

(e) Laundrette

(f) Office

(g) Presentation

(h) Shopping mall

(i) Street (pedestrian)

(j) Street (traffic)

(k) Supermarket

(l) Train

Fig. 2. 3 second duration spectrograms of showing each of the 12 environments.

to the accuracy shown in Section 3 which used wide bandwidth audio samples and reveals no significant loss in classification accuracy at the lower sampling frequency. Table III shows the confusion matrix for these experiments.

The recognition accuracy of individual environments ranged from 81% to 100%, with the building site, car (city), office, presentation, pedestrian street, and supermarket, giving 100% classification accuracy for the 100 examples of each tested. The worst performance was obtained for identifying bus noise which achieved only 81% accuracy. However, the confusions of bus noise were reasonable because the incorrectly classified samples were all recognized as other vehicles, 13% as car (city), 3% as car (highway) and 3% as train. Other reasonable confusions can also be seen; for example, 11% of shopping mall samples were incorrectly classified as supermarket. This is due to the acoustic similarity of the two environments, that is, sounds from trolleys, people, tills, etc.

5. ADAPTIVE LEARNING

To be useful in realistic situations, the acoustic environment classification must be adaptable to changes in the character of environments and to the

Table III. Confusion Matrix of Acoustic Environment Classes

Accuracy %	Building site	Bus	Car (city)	Car (highway)	Laundrette	Office	Lecture	Shopping mall	Street (people)	Street (traffic)	Super market	Train
Building site	100	0	0	0	0	0	0	0	0	0	0	0
Bus	0	81	13	3	0	0	0	0	0	0	0	3
Car (city)	0	0	100	0	0	0	0	0	0	0	0	0
Car (highway)	0	1	1	98	0	0	0	0	0	0	0	0
Laundrette	0	0	0	0	90	0	0	4	0	3	3	0
Office	0	0	0	0	0	100	0	0	0	0	0	0
Lecture	0	0	0	0	0	0	100	0	0	0	0	0
Shopping mall	0	0	0	0	0	0	0	89	0	0	11	0
Street (people)	0	0	0	0	0	0	0	0	100	0	0	0
Street (traffic)	0	2	0	0	1	0	0	0	4	93	0	0
Supermarket	0	0	0	0	0	0	0	0	0	0	100	0
Train	0	0	0	0	0	0	0	0	0	1	0	99

introduction of new environments. It is also desirable to know the confidence of the environment classification. Here, we describe a confidence measure for our classifier and its use in a set of continuous learning experiments. The common notion of continuous learning has been described by Vega and Bressan [2003] who have shown that an algorithm which gives a good initial performance will improve its performance if it is given its own classified data as further training examples.

5.1 Confidence Measure

The classifier operates by each acoustic environment HMM outputting a log likelihood score according to how well it matches the unknown test environment sample. For classification, the HMM outputting the highest log likelihood is selected as the output class. The likelihoods from the two highest scoring HMMs can be used to form a confidence measure. Let L_1 be the log likelihood of the best matching model, and L_2 be the log likelihood of the second best matching model. The confidence measure for the classification result can then be computed as:

$$conf = \left| \frac{L_1 - L_2}{L_1} \right| .$$

In these classification experiments, the accuracy was 96% on the test set which was comprised of 1,200 examples, meaning that 50 examples were incorrectly classified. Of these, 38 had the correct model as the second best matching model. This shows that the inclusion of uncertain samples from Dataset 2 degrades the performance; the result is similar to results given by Vega and Bressan [2003].

Further analysis of the classification results revealed that the confidence scores of incorrectly classified samples were all below 0.07. By selecting an appropriate threshold, the confidence measure computed after classification can be used to modify the output of the classifier. If the confidence is below the threshold, the result can be labeled as uncertain. If the confidence is greater

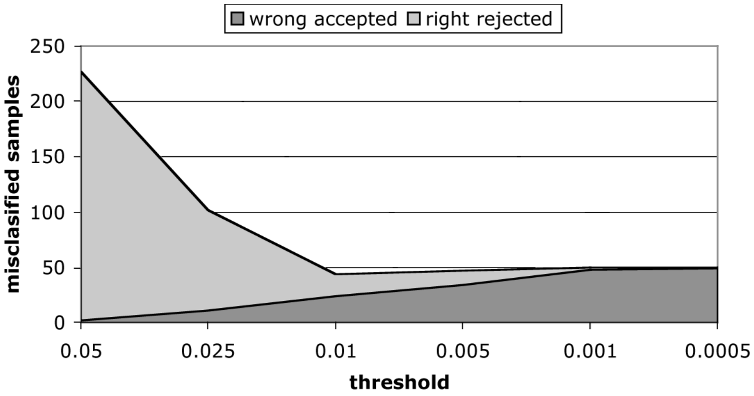


Fig. 3. Misclassifications at different confidence levels.

than the threshold, the result can be labeled as certain. Figure 3 shows the effect that varying the threshold has on the number of rightly rejected (Type 1 error) and wrongly accepted (Type 2 error) classification results.

These misclassifications are minimized at a threshold of 0.01. This shows that the threshold can be used as an uncertain recognition indicator in the application. The appropriate threshold value for a particular application depends on the relative cost of rejecting correctly identified samples against that of accepting wrongly classified samples.

5.2 Learning

Over time, the characteristics of an environment may change and also new environments may be encountered. In some instances, the environment changes incrementally such as new sounds in the street or shopping mall. In other cases, the environment may change very suddenly such as moving from a quiet individual office to a large open office. Therefore, adapting to changes of environment and continuously learning new environments are important requirements for acoustic environment classifiers. These requirements mean that it is necessary to continuously update existing models with new training data and to adaptively build more robust models. Several methods have been successfully developed to adapt the models for speech recognition to either a particular speaker or acoustic environment. These include maximum likelihood linear regression (MLLR) [Leggetter and Woodland 1995] and maximum a-posteriori (MAP) adaptation [Duda et al. 2001] which utilize relatively small amounts of training data to reestimate the speech model. However, in this work, we examine the use of a relatively large amount of adaptation data that has been collected from many 3-second examples.

In order to investigate the impact of continuous learning, we conducted a series of experiments using Dataset 3 which we split into two equal parts: Dataset 3A and Dataset 3B, each containing 50 samples of each of the 12 environments. The sound environments in Datasets 3A and 3B are similar since they were recorded in the same month. Dataset 3B was used as the testing set and Dataset 3A was used as the new training set.

Table IV. Learning Experiments

Training Set	Testing Set	Accuracy
Dataset 1	Dataset 3B	75%
Dataset 1 and Dataset 3A	Dataset 3B	94%
Dataset 1 and 'certain' samples in Dataset 3A	Dataset 3B	85%

To examine the effect of changes to the acoustic characteristics of the environments the 600 samples in Dataset 3B were tested using the models trained on Dataset 1. The overall classification accuracy is 75% (Table IV). This is considerably lower than that obtained when testing using Dataset 2 and arises from the variation between Dataset 1 and Dataset 3. Because of the limited size of the database and the time differences between recordings, some variants of an environment may end up being present in the training set but not in the test set or vice versa. Clearly this is detrimental to the performance.

To adapt the models to the environment characteristics in Dataset 3B, the 600 sound examples in Dataset 3A were combined with Dataset 1, and the models were retrained and Dataset 3B was then reclassified. This gave a classification accuracy of 94% (Table IV) which is substantially higher than that achieved with the mismatched models trained using only Dataset 1. The accuracy on Dataset 3B is now comparable to that attained on Dataset 2 in Section 4.2. This demonstrates how the models can be adapted to new environmental conditions. However, when retraining the HMMs, the examples in Dataset 3A were manually labeled to ensure that they were associated with the correct HMM. This is not a realistic scenario in the deployment of such a system.

To examine the effect of not having the reference labels with the adaptation data, a new experiment was performed. Using the models trained from Dataset 1, the training data samples in Dataset 3A were classified and the confidence measure for each sample computed. This allows the sound examples in Dataset 3A to be divided into certain (where the confidence measure is above a threshold) and uncertain (where the confidence measure is below a threshold). Based on the analysis from previous experiments (Figure 3), the confidence threshold was set at 0.075. Following this procedure, 343 samples of the 600 in Dataset 3A were certain and 257 uncertain. The 343 certain samples from Dataset 3A were then combined with Dataset 1 and a new set of models trained. These were tested using Dataset 3B which gave a classification accuracy of 85% (Table IV). This is 10% above that obtained using only Dataset 1 for training and 9% below using all of Dataset 3A. However, the adaptation procedure is now unsupervised as no hand labeling of the adaptation data is necessary. Instead the classification result followed by a confidence measure-based decision is used to determine which examples are used to form the training data set.

Theoretically, it is possible to train a classifier with data covering all the variation of an acoustic environment. However, such learning may not be practical for context-aware applications. First, it requires an enormous amount of training data since the training data must be representative of the variability of the patterns for all possible classes of environments; this is a limiting factor. Second, it is not possible to build a generic model for an environment due to the difference between the acoustic properties and the human labels. We may use the office environment as an example. Room A is a large, open plan room with air conditioning, printers, photocopiers, many telephones, and background music, and room B is a small room in a library building. From the human label point of view, office can be defined as a room or other area in which people work. Thus, both room A and room B can be classified as office. In fact, there are no similar sounds in room A and room B so from the view of acoustic properties, room A and room B are completely different.

An alternative solution to designing and training a classifier to be as general as possible is to train a specific classifier for each new operating situation and combination of environments encountered. For context tracking applications, we may train a personalised classifier whose goal is to accurately classify a user's own daily routines.

6. SOUND EVENT MODEL

The acoustic environment models are vulnerable to incorrect classification when an intermittent acoustically-dominant sound event occurs in an environment. An example is a telephone ringing in an office. This would lead to a 3-second recording full of ringing sounds rather than the general office background sound, potentially causing an incorrect classification. A potential solution to address the problem of classifying multiple simultaneous environmental noises is to build models of those sound events which can be combined to provide a description which can be used to classify the acoustic environment. A second use for sound event recognition is to identify sound events of interest to an application or provide additional information about a particular acoustic environment, possibly incorporated into a single classification framework with the acoustic environment models described previously.

The experiments using sound event models were conducted using the RWCP Sound Environment Database in Real Acoustic Environment [Nishiura et al. 2003]. The RWCP-DB includes types of 105 sound events and approximately 100 samples for each type. The types of sound event include collision sounds of wood, plastic, and ceramics; sounds of human activities such as spraying, sawing, and clapping; the sounds of coins, books, pipes, telephones, toys. They are grouped into three main categories and 14 subcategories. There are a total of 9,685 samples in the database of which we used 6,776 for training and 2,909 for testing. The data are in 16kHz, 16-bit, Mono RAW format.

Models for sound events can be obtained using our HMM classifier, but there are subtle differences between recognizing isolated sound events and identifying acoustic environments. An acoustic environment is a complex sound made up of a mixture of different events. There is no constraint on what these sounds can be, and they may emanate from many different localities in the environment. All the sound events in this database are produced from a single source and are constrained to a single location. A typical sound event can be divided into five states, silence-begin-middle-end-silence. Experiments suggest a 5-state HMM best models the sound events, compared with a 9-state model for acoustic environments.

Table V. Results of Single Sound Event Classification (HMM)

Group	Category	Sound Source Examples	No. Test Samples	Accuracy(%)
Collision	wood	wood board, wood stick	347	76
	metal	metal board, metal can	300	72
	plastic	plastic case	165	90
	ceramic	glasses, china	240	46
Action	article dropping	dropping articles in box	60	84
	gas jetting	spray, pump	60	95
	rubbing	sawing, sanding	150	93
	Bursting/breaking	breaking stick, air cap	60	97
	clapping	hand clap, slamming clap	249	81
Characteristic	small metal articles	small bell, coin	327	91
	paper	book, tearing paper	120	75
	instruments	drum, whistle, bugle	316	100
	electronic sound	phone, toy	215	100
	mechanical	spring, stapler	300	99
Overall accuracy: 85%				

The results of these experiments show that the 5-state model gave the best performance. We obtained an overall 85% accuracy classifying the 105 kinds of sound event from the testing samples. The classification results of the categories are shown in Table V. Results show that the HMM classification works very well with the characteristic sounds, especially instruments, electronic, and mechanical sounds. It yields poor results with the collision sounds as a consequence of the finer subdivisions used in the collision sounds that disappears when a coarser classification granularity (i.e. similar to that of the other groups) is used. The two-stage neural network model of Toyoda et al. [2004] gives better results on the collision sounds but lower accuracy on other types of sound (e.g., 80% accuracy on typing sounds). Their model is limited to recognizing 10 or fewer sound types as it does not converge with more sound types. Our model shows no signs of degraded performance as the number of classes increases.

To test the hypothesis that the poor classification results of some categories of sound event in the RWCP database arose from the acoustic similarity of sounds, we classified the 9,685 samples, using a k-means clustering, with the same features as the HMM classifier. We conducted three series of experiments. The measure of accuracy of the classification is calculated as the number of samples of the largest category in each cluster normalized for the varying category sizes.

First, we classified all 105 sound source types into clusters, allowing one cluster for each type. The overall accuracy is 37%. The top 10 most accurately classified sound source types were all from the characteristic sounds group (Table VI). An analysis of variance of the classification accuracy shows significant differences between the groups ($p < 0.01$). For the second set of experiments, we reduced the number of clusters to 14 and classified at the granularity of the category to examine the between-categories cohesiveness. The classification accuracy for the 14 categories of sound sources varied between 0% (i.e., the category was not the largest in any cluster) and 81% (i.e., 81% of examples were in the largest category in a cluster). In a further set of experiments to test the stability of the classification, we divided the RWCP database into eight subsets

Table VI. K-means Clustering of Single Sound Events

RWCP Hierarchy			Sound Source Classification		Sound Category Classification	
Group	Sound Category	Sound Surce Types	Accuracy % (aggregated to categories)	Accuracy % (aggregated to groups)	Accuracy % (categories)	Accuracy % (aggregated to groups)
Collision	wood	12	0	21	82	17
	metal	10	45		0	
	plastic	6	5		0	
	ceramic	8	19		0	
Action	article dropping	2	51	26	69	19
	gas jetting	2	48		0	
	rubbing	5	0		0	
	bursting/ breaking	2	0		0	
	clapping	10	13		0	
Characteristic	small metal articles	15	39	49	74	42
	paper	4	24		0	
	instruments	11	56		23	
	electronic sound	8	57		36	
	mechanical	10	60		48	
Overall accuracy			37%		44%	

and clustered each separately; these gave similar results. Another measure of the acoustic similarity of different categories is given by the group of the second largest category in each cluster. Only 30% of these come from the same group as the largest category in each cluster, providing further evidence that the sound event groups are not clearly distinguished in the MFCC domain.

The rank order of categories is similar to that from the HMM classifier (r_s significant at 0.05), indicating that the k-means and HMM classifiers are classifying in similar ways and that many categories are heterogeneous at least in the MFCC domain (which is a reasonable approximation to the human aural response to sounds).

These results show that many of these manually-constructed sound event classes are based on criteria other than acoustic similarity. The analysis of our acoustic environment misclassifications and the evidence of acoustic heterogeneity in the RWCP hierarchy suggest that successful hierarchical models of acoustic environments and sound events need to ensure adequate acoustic cohesion as well as semantic distinctions useful to the applications which use them. There is a need for substantial further work on sound taxonomies which is beyond the scope of this article.

6.1 Hierarchical Model

The problems caused by the confusion of noise environments with similar characteristics described in Sections 3 and 4 and the variable granularity of classifications such as the RWCP database described previously show the need for an extensible sound model.

The simple hierarchical noise model we have described supports hierarchies of environments. It allows environments to be grouped using the heuristic that a good grouping of two classes should maximize the number of misclassified instances of each class having the other as the second most likely class. The use of suitable hierarchies allows us to assign samples with poor confidence scores to a more general class. In this case, the model chooses the lowest class in the hierarchy that subsumes the first and second matches (i.e., those used in calculating the confidence score). For example, in the experiments reported in Sections 3 and 4, most of the misclassifications are covered by grouping the car, bus, and train environments into a more general transport class, and grouping supermarket and shopping mall into a shopping class.

Grouping environments with very different acoustic characteristics will result in poor performance (e.g., grouping train and office under a class of places where I read papers). However, our experience is that the acoustic environment is a sufficiently good predictor of activity that users' natural classifications are often reasonable from the perspective of our model. However, the analysis of the RWCP data described previously shows that constructing meaningful acoustically-cohesive sound event hierarchies requires more sophisticated approaches than those in our prototype.

The use of the hierarchical model allows us to merge the acoustic environment and sound event data. The models have been divided into two levels, environments (office, supermarket, street, etc.) and sound events (telephone ring, clapping, etc.). The classifier first compares the unknown sample against the set of environment models. If the classification is uncertain (low confidence score), the sample will be compared against the set of sound events to find a cue. If confusion still exists, other contextual information might aid in reducing ambiguities although that is outside the scope of this article.

It is possible to define a hierarchy of environmental sound recognition problems with increasing levels of complexity. The simplest recognition problem is the classification of isolated sound event if there are only a limited number of distinct classes.

7. APPLICATIONS

To show how the acoustic environment can be exploited in context-aware applications, we have developed a context tracker application using a microphone with periodic audio buffering of acoustic environments; this is fully described in Smith et al. [2005]. We have implemented and evaluated the tracker application on a PDA [Steward 2005], although the power consumption of many PDAs gives rise to issues of limited availability and connectivity which are beyond the scope of our work but have been addressed by Stäger et al. [2004] at ETH Zurich.

The tracker is able to recognize a user's current environment rapidly by classifying short duration sound recordings together with a confidence score. The recording duration, interval, and recording quality are user-controllable. If the classification is uncertain (i.e., the confidence measure is below the threshold) samples are taken more frequently until either an environment is recognized (i.e., the confidence measure is above the threshold) or a time interval for classification is exceeded and the application marks the samples as unknown and prompts for a label for the new environment. Training can be performed manually by the user to train the personalised model or it can be started automatically when an uncertain environment is found. The samples with the high confidence score can also be sent to the classifier to refine models. The design of the application allows the model to be continuously updated with new training data in different locations and adaptively build more robust models over time. Users are able to create and train personalised models based on their daily routines. All recognized and confirmed environments are recorded in a tracking file or database for later retrieval and are annotated for use by other components in multimodal systems.

8. CONCLUSIONS

We have described a HMM-based acoustic environment classifier which was trained on data from 11 acoustic environments encountered in daily routines. The classification accuracy of individual environments ranged from 75% to 100% with an overall accuracy of 92%. A human listening test performed on the same test set yielded 35% accuracy which indicates that our classifier has the advantage of recognizing acoustic environments from short samples. The classifier is optimized for use in low-bandwidth communication and in limited capability devices. Overall recognition performance using these devices to classify everyday acoustic environments is 96%. The classifier also performs well on sound event data.

The experimental results show that wrongly identified samples are most often from similar environments to their correct classification; this is particularly true of the fine-grained classification used in parts of the RWCP single-sound database. We have developed a confidence measure which allows us to classify sounds in a hierarchical model at the most specific level consistent with high confidence. The construction of semantically-useful cohesive acoustic environment and sound event hierarchies has several open issues which are beyond the scope of this work.

These results have clearly shown the potential of using the acoustic environment as a context indicator. The classifier gives good results over a wide range of sample qualities and environments and an adaptive learning strategy allows the model to learn new environments and recognize the specific environments of an individual. We have described the model's use in a client-server context-aware prototype tracker application and demonstrated the feasibility of capturing environmental noise using current PDAs, although there are several implementation issues that need to be addressed before integrated systems can be deployed.

This work shows that the acoustic environment is a rich source of context information which can be recognized with a high degree of accuracy and can be used as a good indicator of current activity.

ACKNOWLEDGMENTS

The authors wish to thank the University of East Anglia for supporting Ling Ma's work, colleagues in the University of East Anglia Language and Speech Laboratory, particularly Stephen Cox, for their help and comments, and the anonymous reviewers for their comments and suggestions.

REFERENCES

BAKKER, E. M. AND LEW, M. S. 2002. Semantic retrieval using audio analysis. In *Proceedings of the Conference on Image and Video Retrieval*. London UK. Lecture Notes in Computer Science, vol. 2383. 271–277.

BROWNE, P., CZIRJEK, C., GURRIN, C., JARINA, R., LEE, H., MARLOW, S., MCDONALD, K., MURPHY, N., O'CONNOR, N. E., SMEATON, A. F., AND YE, J. 2003. Dublin City University video track experiments for TREC 2002.

CAI, R., LU, L., ZHANG, H. J., AND CAI, L.-H. 2003. Using structure patterns of temporal and spectral feature in audio similarity measure. In *Proceedings of the ACM Multimedia Conference*. Berkeley, CA. (Nov.). 219–222.

CLARKSON, B., SAWHNEY, N., AND PENTLAND, A. 1998. Auditory context awareness via wearable computing. *Workshop on Perceptual User Interfaces*. 37–42.

COUVREUR, L. AND LANIRAY, M. 2004. Automatic noise recognition in urban environments based on artificial neural networks and hidden Markov models. *Inter-noise2004*. Prague, Czech Republic.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification*, 2nd Ed. Wiley, New York, NY.

FOOTE, J. 1999. An overview of audio information retrieval. *Multimedia Syst. 7*, 1, 2–11.

GAUNARD, P., MUBIKANGIEY, C. G., COUVREUR, C., AND FONTAINE, V. 1998. Automatic classification of environmental noise events by hidden Markov models. *Appl. Acoustics 54*, 3, 187.

HINDUS, D. AND SCHMANDT, C. 1992. Ubiquitous audio: Capturing spontaneous collaboration. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*. Toronto, Canada (Nov.). 210–217.

HUANG, X., ACERO, A., AND HON, H. 2001. *Spoken Language Processing*. Prentice Hall, Englewood Cliffs, NJ.

LEGGETTER, C. J. AND WOODLAND, P. C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang. 9*, 171–185.

LIU, L., JIANG, H., AND ZHANG, H.-J. 2001. A robust audio classification and segmentation method. In *Proceedings of the ACM Multimedia Conference*. Ottawa, Canada. 203–211.

LIU, L., ZHANG, H.-J., AND JIANG, H. 2002. Content analysis for audio classification and segmentation. *IEEE Trans. Speech Audio Process. 10*, 7, 504–516.

MA, L., SMITH, D. J., AND MILNER, B. P. 2003. Context awareness using environmental noise classification. In *Proceedings of Eurospeech*. Geneva, Switzerland, 2237–2240.

MA, L., SMITH, D. J., AND MILNER, B. P. 2003. Environmental noise classification for context-aware applications. In *Proceedings of the International Conference on Database and Expert Systems Applications (DEXA)*. Lecture Notes in Computer Science, vol. 2736. 360–370.

MYNATT, E. D., BACK, M., WANT, R., BAER, M., AND ELLIS, J. B. 1998. Designing audio aura. In *Proceedings of Conference on Human Factors in Computing Systems (CHI'98)*. 566–573.

NISHIURA, T., NAKAMURA, S., MIKI, K., AND SHIKANO, K. 2003. Environment sound source identification based on hidden Markov model for robust speech recognition. In *Proceedings of EuroSpeech*. 2157–2160.

PELTONEN, V. T. K., ERONEN, A. J., PARVIAINEN, M. P., AND KLAPURI, A. P. 2001. Recognition of everyday auditory environments: Potentials, latencies and cues, 110th Convention of Audio Engineering Society.

PELTONEN, V., TUOMI, J., KLAPURI, A., HUOPANIEMI, J., AND SORSA, T. 2002. Computational auditory environment recognition. In *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*. Orlando, FL.

QUÉNOT, G. M., MORARU, D., BESACIER, L., AND HULHEM, P. 2003. CLIPS at TREC-11: Experiments in video retrieval. *TREC-2002*.

SAWHNEY, N. AND SCHMANDT, C. 2000. Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput. Human. Interact. 7*, 3, 353–383.

SAWHNEY, N. 1997. Situational awareness from environmental sounds. Tech. rep. for *Modeling Adaptive Behavior (MAS 738)*. MIT Media Lab.

SCHEIRER, E. AND SLANEY, M. 1997. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 1331–1334.

SCHMANDT, C., MARMASSE, N., MARTI, S., SHAWHNEY, N., AND WHEELER, S. 2000. Everywhere messaging. *IBM Syst. J. 39*, 3–4, 660–677.

SMEATON, A. F. AND OVER, P. 2003. TRECVID: Benchmarking the effectivenss of information retrieval tasks on digital video. Lecture Notes in Computer Science, vol. 2728. 19–27.

SMITH, D., MA, L., AND RYAN, N. 2005. Acoustic environment as an indicator of social and physical context. *Person. Ubiquitous Comput. 10*, 1 (DOI: 10.1007/s00779-005-0045-4).

SRINIVASEN, S., PETKOVIC, D., AND PONCELON, D. B. 1999. Towards robust features for classifying audio in the CueVideo system. In *Proceedings of the ACM Multimedia Conference*. 393–340.

STÄGER, M., LUKOWITZ, P., AND TRÖSTER, G. 2004. Implementation and evaluation of a low-power sound-based user activity recognition system. *International Semantic Web Conference*. 138–141.

STEWARD, J. 2005. Using a PDA for audio capture. BSc Project, University of East Anglia, Norwich, UK.

TOYODA, Y., HUANG, J., DING, S., AND LIU, Y. 2004. Environmental sound recognition by multilayered neural networks. In *Proceedings of the 4th International Conference on Computer and Information Technology (CIT '04)*. 123–127.

TZANETAKIS, G. AND COOK, P. 2002. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process. 10*, 5, 293–302.

VEGA, V. S B., BRESSAN, S. 2003. Continuous naive bayesian classifications. In *Proceedings of the International Conference on Asian Digital Libraries*. 279–289.

VENDRIG, J., DEN HARTOG, J., VAN LEEUWEN, D., PATRAS, I., RAAIJMAKERS, S., VAN REST, J., SNOEK, C., AND WORRING, M. 2003. TREC feature extraction by Active learning, *TREC-2002*.

WOLD, E., BLUM, T., KESLAR, D., AND WHEATON, J. 1996. Content-based classification search and retrieval of audio. *IEEE Multimedia 3*, 3, 27–36.

WU, L., GUO, Y., QIU, X., FENG, Z., RONG, J., JIN, W., ZHOU, D., WANG, R., AND JIN, M. 2003. TRECVid 2003. *TREC-2003*.

YOUNG, S., EVERMANN, G., KERSHAW, D., MOORE, G., ODELL, J., OLLASON, D., VALTCHEV, V., AND WOODLAND, P. 2001. *The HTK Book 3.1*. Cambridge University Engineering Department, Cambridge, UK. http://htk.eng.cam.ac.uk.

ZHUANG, L., ZHOU, F., AND TYGER, J. D. 2005. Keyboard acoustic emanations revisited. In *Proceedings of the ACM Conference on Computer and Communications Security*. Alexandria, VA (Nov).

Received November 2004; revised October 2005; accepted April 2006 by Enrique Vidal