AUTOMATIC SURVEILLANCE OF THE ACOUSTIC ACTIVITY IN OUR LIVING ENVIRONMENT

Aki Härmä, Martin F. McKinney and Janto Skowronek

Digital Signal Processing Group, Philips Research,
Prof. Holstlaan 4, 5656AA, Eindhoven, The Netherlands

ABSTRACT

We report an experiment with an acoustic surveillance system comprised of a computer and microphone situated in a typical office environment. The system continuously analyzes the acoustic activity at the recording site, separates all interesting events, and stores them in a database. All interesting acoustic events over a duration of more than two months were recorded. A number of low-level signal features are computed from the audio signal and used to classify and identify sound events. The analysis reveals interesting patterns and activities which would be difficult to find by any other means.

1. INTRODUCTION

One may playfully argue that the largest conceivable multimedia database is our living environment. There are many different types of acoustic events present in our daily environment. For example, in the office room, the most typical acoustic events come from people talking, opening and closing doors, and using computers and other devices. We adapt easily to many of those sounds and after a short while may not even notice them anymore. In some sense, our natural acoustic environment is at the same time very familiar to us, but still a rich and continuously evolving unexplored terrain. In this paper we demonstrate that systematic classification and log-keeping of acoustic events indeed reveals fascinating patterns and structures in the ordinary acoustic activity around us.

An acoustic surveillance system could have many different applications including various types of automatic or semi-automatic security, safety, and monitoring applications. Recognition of the sounds of vehicles [1], machines [2], and sounds of closing doors, dropping or breaking objects [3, 4, 5] have been studied earlier. In addition, the new MPEG7 multimedia standard provides an interesting framework for sound recognition [6]. Most of the earlier studies are based on supervised training with preselected training material. In this study, the goal is to develop methods for a scenario where the system is installed in an unknown environment and it is expected to adapt in a meaningful way. The standard method in this situation would be to use *unsupervised* classification methods. However, there may also be some possibilities to collect training data for a *supervised* classifier, e.g., through user input. In this article, an experimental system was designed and implemented where we can compare the two alternative methods.

2. THE EXPERIMENT

The system used in our study has two components, a real-time recording system and classification/logging system. The real-time recording system runs on a PC and continuously analyzes the microphone input signal. When an *interesting* event is detected, it stores it as an audio file. The recording system is designed such that once set up, it can run unattended as long as there is disk space available to store new entries. The maximum duration of a recording is limited to four seconds.

In the current experiment the actual classification and log-keeping system runs off-line. It reads the entries and computes their parametric representations. Parametric representations are used to form a classification model that can be used to compute different types of logs and statistical representations related to the activities in the environment.

In a more realistic system, these two components could be combined so that the system could be continuously adaptive to the changes in the environment.

3. EVENT DETECTION AND SEGMENTATION

The first steps in the recording system are to detect and segment auditory events. Detection and segmentation must be adaptive to the changes in the environment. The proposed system is based on a continuously updated background noise spectrum profile. The short-term complex FFT spectrum $S(n,t)$ at the frame-time t is averaged over time using a first-order integrator applied to the square of the spectrum using the following recursion:

$$S(n) = (1-\gamma)|S(n,t)|^2 + \gamma S(n-1), \forall n = 0, \cdots N-1 \quad (1)$$

0-7803-9332-5/05/$20.00 ©2005 IEEE

p1	RMS value	p2	Length
p3	Spectrum centroid	p4	Bandwidth
p5	Band energy ratio	p6	Delta spectrum magn.
p7	Pitch	p8	Coherence
p9	Envelope flatness	p10	Number of events

Table 1. Audio features used in the current study

The FFT size was 1024, the sampling rate of the data was 16 kHz, and the temporal smoothing parameter was set to $\gamma = 0.998$. With this value the noise estimate is moderately elevated in the case of a car passing slowly by the building but for most sounds from humans and animals, the fluctuations of the noise estimate are small.

In this study, the detection of activity is based on two alternative criteria. The first measure is designed to detect sufficiently loud onsets and transients in the environment. This was obtained with a simple full-band difference-to-noise value given by

$$T1_{\mathrm{dB}} = 20 \log_{10} \frac{E[D(n)]}{E[S(n)]}. \quad (2)$$

where the difference spectrum $D(n) = |S(n,t)|^2 - S(n)$.

The threshold value is typically set close to 35 dB, which allows the system in the office to collect the sound of a pen dropping to the floor but not all typing sounds from the computer keyboard.

The second criteria was especially developed to improve the detection of narrow-band sound events. The following criteria was used

$$T2_{dB} = 20 \log_{10} \left[\max(D(n)) - \sqrt{\frac{1}{N} \sum_n D(n)^2} \right]. \quad (3)$$

This measure simply compares the maximum peak in the difference spectrum to its variance. It turned out that the value of 35 dB was also useful as the detection threshold with this measure. In an office environment this is sufficient for detecting a casual whistler a short distance down the hallway but not normal speech sounds from a neighboring office room.

4. PARAMETRIZATION

We characterize individual sound events by a small number of descriptive features, listed in Table 1. The first nine features in the table have been shown to be useful in several audio classification applications [7]. Some of the features are calculated in the time domain and some of them from the spectrum of the signal. All parameters are related to stationary properties of the signal observed over a brief time window. In the current study the parametrization is performed independently of the recording and off-line. Each recording is segmented with a finer temporal resolution such that only

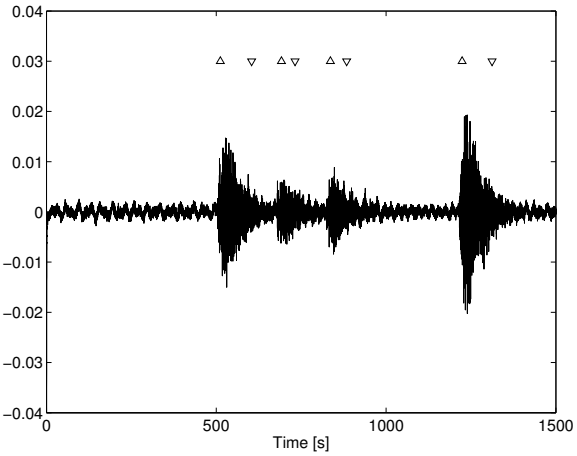

Fig. 1. Fine segmentation of a sound of a computer keyboard.

the part of the recorded signal that triggered the activity detector in the recording phase is preserved. It is common that an event that triggers the detection is immediately followed by other sound events. This is illustrated in Fig. 1, where four brief sound events are segmented from a recording entry. The borders of the events are indicated by small triangles.

The features *p1-p9* are computed separately from each event. The last parameter *p10* is common to all events in a recording and it represents the number of events per second in a recording. The ten parameters were chosen from a larger set of candidates by the analysis of the covariances.

5. CLASSIFICATION

More than 140000 interesting events were recorded in the office room during the surveillance period of 48 days. Manual classification of a random selection of recordings showed that 65% of the recordings are sounds of the computer keyboard, 20% are speech sounds, 6 % are high frequency sounds of whistling and the printer in the hallway. The rest of the sounds include sounds of doors (2%) and office furniture, coughing and sneezing, unidentified rumble, and ringing telephone.

In this article, we test the performance of the classical k-means algorithm for unsupervised clustering of the data. It aims to position K code-vectors at the local maxima of the data distribution in the feature space. Another classifier was designed by hand so that the cluster centers were selected from the training data by subjective criteria. This simulates the supervised training scenario where the classifier is based on user input.

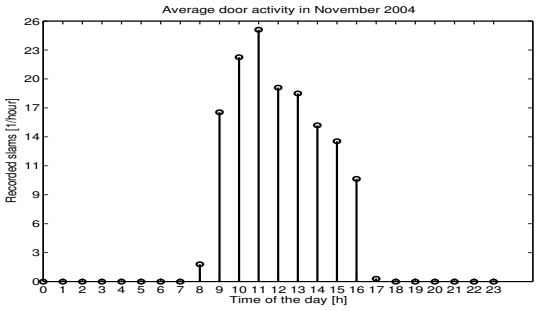

Fig. 2. Sounds of a closing door on the opposite side of the hallway. The graph shows average activity per hour of the day on work days averaged over a period of two working weeks.

5.1. Manual selection of cluster centers

The selection was performed by a subject who is not the resident of the office room. He picked up 25 sounds including various types of vocal expressions, sounds of doors, and devices in the office and the hallway. The selected entries can be used directly as a basis of classification and for building logs and statistics. In practice, the feature vector of a *representative* example of a sound is selected and its Euclidean distances to all other entries are computed. Then all entries within a certain threshold distance are considered to be associated with the prototype sound. In the following examples the threshold was determined manually for each prototype.

A particular door close to the office room is used quite frequently. The closing of the door made a slamming sound until it was apparently fixed near the end of the observation period. The histogram of the sounds produced by the door in an observation period of two weeks is shown in Fig. 2. The highest peak of activity (almost 26 slams per hour) is found between 11AM and noon. The activity is at a higher level before lunch than after the lunch. The activity between 5PM and 7AM is low because the door from the office to the hallway is usually closed at those times. It was possible to find a damped version of the same sound event from the data for the times when the door was closed but that is not included in the current analysis.

Another interesting finding is shown in Fig. 3. The data represents coughing sounds caused by a sudden respiratory infection of one of the authors in October, 2004. The coughing starts gradually on Friday and is at the peak on Monday finally fading away by the end of the week.

5.2. Comparison to automatic clustering

There are many different methods for unsupervised clustering of data. In this article we use one of the most classical algorithms, often called the k-means. Many different configurations were tested such as different initializations and

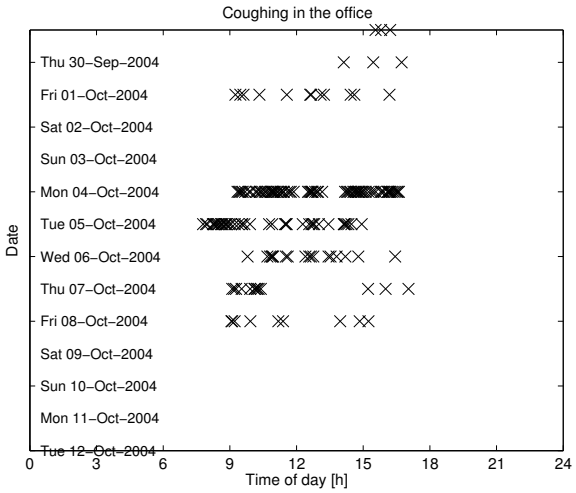

Fig. 3. Log for the sound of coughing in the office over a period of two weeks in October, 2004. The author had a respiratory infection over the weekend.

numbers of initial cluster centers. In this article we show the results of a typical case with 25 classes.

The k-means and manual classification are compared in Fig. 4. The top left panel gives the numbers of data points per class determined by the Euclidean distances. In both manual (solid) and k-means clustering (dotted line) the largest number of data points represent a sound of keyboard (normal typing). Manual inspection of these clusters showed that the 50 closest entries to the cluster center were also keyboard sounds in both cases.

The analysis of the closest hits of classes showed that in both classifiers there were several classes containing mainly different keyboard and speech sounds. The percentages for four different types of sounds are plotted in the top right panel of Fig. 4. The percentage of k-means classes representing different keyboard, door, and whistling sounds (asterisks) are almost identical to the statistics obtained with random sampling of the data (circles). In manual classification (triangles) there are significantly less classes for the keyboard sounds. In the case of speech sounds both classifiers give a larger percentage than the actual statistics suggests. This mainly reflects the fact that there are typically more individual events in a speech recording than, e.g., a recording containing a door slam. The manual clustering was found unreliable in the inspection. In 36% of the classes the most of the entries in the class were actually other type of sounds than the chosen prototype sound.

The bottom left panel shows the squared distances for 10% of the closest data points in each class. This value roughly measures the compactness of a cluster. The value behaves very differently in the manual and k-means clustering: the clusters are, on average, more compact when k-means clustering is used. For example, class 8 in manual

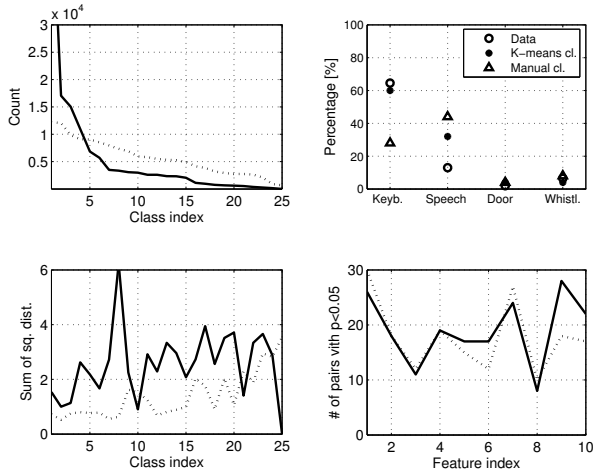

Fig. 4. Comparison of the k-means clustering (dotted) and the manual (solid curve) clustering. The panels are explained in the text.

clustering representing a mixed class of sounds gets a very high value.

The bottom right panel compares how well any single feature of Table 1 separate pairs of classes. This was computed by two-way ANOVA analysis of pairs of classes separately for each feature. The y-axis gives the number of pairs of classes where the separation by one feature only was found statistically significant ($p < 0.05$). The average number of such pairs is similar in both clusterings but there are some differences. The largest difference is in the envelope flatness which is a more significant factor in manual clustering.

6. CONCLUSIONS

Development of ubiquitous and pervasive systems requires technologies that can automatically adapt to changes in the environment. To achieve this goal an acoustic surveillance system should monitor the environment and organize the observed activity in a meaningful way. Here, we reported on an experiment with an automatic recording system that was used to collect over 140000 interesting sound events in a typical office room over a continuous observation period of more than two months. One important component of the system is a detection algorithm based on continuous estimation of background noise and two different criteria for triggering the recording of sound events. The proposed algorithm performed in a predictable way for several weeks in a varying environment.

Classification of the recordings was based on descriptive parametric representations of the collected sound events. The training of the classifier can be performed in many different ways. In this article we used the classical k-means algorithm for unsupervised learning from the data. Secondly, we designed another classifier where a small number of subjectively interesting sound events were identified from the data and the parameter vectors of those were used directly to define the classification regions. The unsupervised classifier did find many interesting events from the environment. However, the clustering was dominated by the large number and variability of keyboard and speech sounds. The supervised method based on user's input produced the desired class centers but it often led to a classifier with many fuzzy or mixed classes. The results of the current experiment may suggest that a system combining both supervised and unsupervised training methods could be potential for practical applications. But, before that, we need to determine what are actually the interesting sounds in our environment.

Acknowledgments

The authors are grateful to Dr. Denteneer for valuable discussions and suggestions.

7. REFERENCES

[1] M. E. Munich, "Bayesian subspace methods for acoustic signature recognition in vehicles," in *Proc. EUSIPCO-2004*, (Vienna, Austria), September 2004.

[2] L. Atlas, G. D. Bernard, and S. B. Narayanan, "Applications of time-frequency analysis to signals from manufacturing and machine monitoring sensors," *Proc. IEEE*, vol. 84, pp. 1319–1329, September 1996.

[3] J. P. Woodard, "Modeling and classification of natural sounds by product code hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 1833–1835, July 1992.

[4] R. S. Goldhor, "Recognition of environmental sounds," in *Proc. of IEEE Int. Conf. on Acouctics, Speech, and Signal Processing*, vol. I, pp. 149–152, 1993.

[5] H. Watanabe, Y. Matsumoto, S. Tanaka, and S. Katagiri, "A new approach to acoustic signal monitoring based on the generalized probabilistic descent method," *IEEE Trans. Signal Processing*, vol. 47, pp. 2615–2618, September 1999.

[6] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Trans. Circ. Systems for Video Tech.*, vol. 11, pp. 737–747, June 2001.

[7] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR'2003)*, (Baltimore, USA), October 2003.