Uninformed Abnormal Event Detection on Audio

Rolf Bardeli, and Daniel Stein

Fraunhofer Institute for Intelligent Analysis and Information Systems
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Email: `surname.name@iais.fraunhofer.de`

Abstract

Although abnormal events in an audio stream are by their nature hard to define, a continuous monitoring of audio surveillance data can detect crucial information in, e.g., train engines that might require critical maintenance. Our method detects abnormal events without being trained on a certain situation, by building a model of the expected sound environment given a continuously adapting history of past audio material within a limited time interval. We evaluate the precision of this method on recordings from train rides.

1 Introduction

Especially for heavy-duty engines emitting certain kinds of sounds, continuous monitoring for abnormal events of pre-installed audio surveillance can submit crucial information for an early maintenance. This holds true especially for public transport, where newly introduced brands of tires or damaged tracks often produce abnormal sounds long before being noticed, and where an in-time mending of broken components could prevent both minor damages and major desasters. However, the sounds produced by different types of engines are as heterogeneous as are their accompanied sound situations. A suitably-tailored solution for each and every scenario is not always feasible.

In this paper, we present a procedure for continuous abnormal event detection in an audio stream which does not require special training material to define its events. Rather, it constantly builds up a theory of surrounding sounds and noises and compares newer audio fragments with an adaptive history.

This paper is organized as follows: in Section 2, we review existing techniques on audio event detection. In Section 3, we present our technique of uninformed abnormal event detection. Section 4 describes the data that we used for evaluation, and Section 5 presents the outcome of our experiments. We conclude the paper in Section 6.

2 Related Work

In many applications, abnormal event recognition still has to be performed by human listeners. Any approach helping to significantly reduce the amount of data to be inspected by humans is of great help.

Systematic technical solutions exist only for pre-defined sounds or sound classes, such as gun shots [1]. Models for these sounds have to be learned from training data which often leads to implicit modeling of the background sounds of the domain of the training data and therefore restricts generalisation capabilities. Typical models in this context include Gaussian mixture models and hidden Markov models [2, 3].

Approaches for abnormal audio event detection without prior knowledge of the abnormal sounds typically need training data to describe the whole range of sounds which are to be classified as normal [4].

The few existing entirely uninformed approaches are based on support vector machine novelty detection [5]. Current approaches typically have to be adjusted significantly for even slight changes in the application domain or the recording equipment.

The public transport domain has been identified before as an important field of application for abnormal event detection [6, 7].

3 Abnormal event detection from audio

While public transport is full of noise which does not carry crucial information, other noise that deviates from the usual background may indicate damage on roads, tracks, or engines. Statistics on such abnormal audio events can point public transport managers to parts of their infrastructure that need repair or improvement. Ideally, audio event detection can point out problems at such an early stage that accidents can be prevented long before they would have occurred. In the following, we describe a model-free algorithm for abnormal event detection from audio based on adaptive dictionary learning.

In order to find abnormal sounds, we follow an approach that is based on constantly learning a representation of the acoustic situation in a past time interval and trying to explain the current acoustic situation based on a spectral dictionary derived from this data as well as possible (see Figure 1). If there is considerable energy left in the signal that cannot be explained from the dictionary, this will be classified as an abnormal event and reported.

More formally, we model our approach as follows. First, we fix a dictionary size s. This is the number of spectral vectors to be used in the dictionary. Moreover, we choose a history length h, giving the number of spectral vectors preceding each position in an audio file. This history is used to learn an updated dictionary for each position in the audio by performing s-means clustering on it. Each cluster centre forms one entry in the dictionary. The latter is represented by a matrix $D \in \mathbb{R}^{s \times b}$, consisting of s rows each of which contains a cluster centre. From the audio spectrum, we choose b Mel scaled bands. The cluster centres from the previous time step are used to initialise the clustering in the next step in order to track the acoustic situation. After the dictionary is learned, it is used to describe the current analysis interval of the audio stream consisting of k vectors of Mel bands given by a matrix $V \in \mathbb{R}^{k \times b}$. We find a matrix W of weights such that $WD + E = V$ where E is an error matrix and the weights W are chosen such that the norm $\|V - WD\|_2$ of E is minimised. The row vectors w of W are found from the row vectors v of V by minimising the norm of each individual row vector e of E:

$$w = \underset{x}{\operatorname{argmin}} \|v - xD\|_2^2,$$

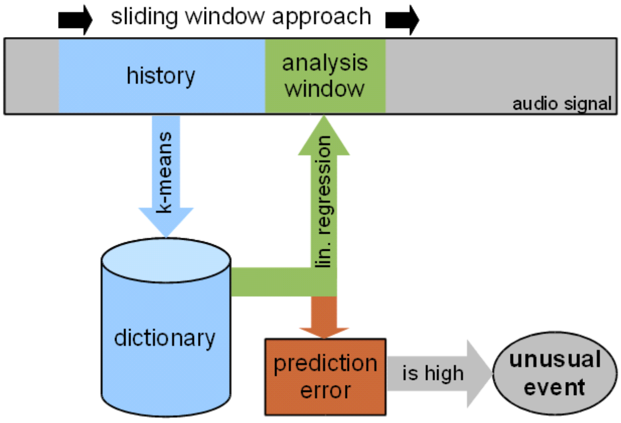

Figure 1: Overview of the event detection process. The spectrogram in each analysis interval is modelled by an adaptive dictionary. Events are detected based on the energy in the residual left by describing the analysis interval by the dictionary.

subject to the condition that the components of w are non-negative. This reduces to an ordinary least squares problem:

$$w = \operatorname*{argmin}_{x}(xDD^{\top}x^{\top} - 2vD^{\top}x^{\top}).$$

It can be solved by applying constrained quadratic programming with respect to x implying the non-negative constraint. Now the resulting w is a sparse vector of the weights for each dictionary entry from D.

Finally, each position in the audio stream is mapped to an error vector, i.e., a column of E. Each column is further reduced to a single number by computing its ℓ_1-norm. Further, the harmonic spectral spread (HSS) of the error signal is computed. The harmonic spectral spread S of a row vector $v = (v_0, \ldots, v_{b-1})$ of V is defined as follows:

$$S = \sqrt{\sum_{n=0}^{b-1} (\log_2(f_n/1000) - C)^2 v_n / \sum_{n=0}^{b-1} v_n}$$

where f_n is the centre frequency of bin n. Larger values for HSS indicate that the energy in the residual covers a significant amount of the entire mel spectrum rather than some specific frequency band. Abnormal event detection is then performed by peak picking on a resulting sequence of numbers given as a weighted sum of these two measures. This sequence will be further referred to as detection function.

4 Data

Audio data from public transport is hard to obtain via transport providers. We therefore used the opportunity to collect data from various means of public transport used by one of the authors on his daily commute to and from work. On a route of about 75 km we collected synchronised audio recordings and GPS data. Audio recordings were conducted using a mobile recorder standing on the floor, close to the engine level of the train. By wrapping the recorder inside a backpack, it was ensured that utterances from fellow travelers are too muffled to be intelligible, in order to avoid privacy concerns. GPS data was logged synchronously. Each recording contains approximately 2 hours and includes bus, tram and train commutes. On the content level, the data contains a wide variety of acoustic events including engine and vibrational noise, (muffled) utterances from fellow travelers and public announcements of stops and connection options, as well as background noises like ambulance alarms.

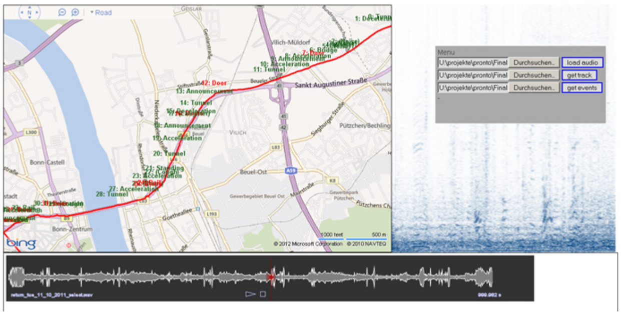

Figure 2: Visualising events on a map. Based on synchronous recording of audio and GPS data, the events can be localised. *Map material (c) Microsoft Corp.*

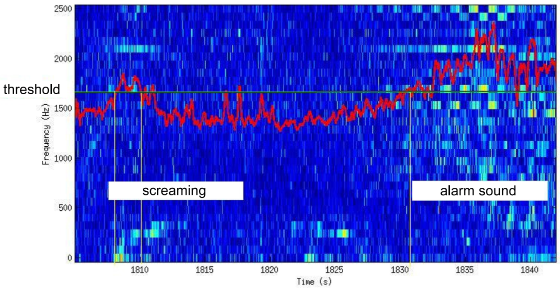

Figure 3: Example of a residual spectrogram and the detection function. Events are detected by thresholding this detection function.

Synchronous recording of audio and GPS data allows to visualise events on a map. Figure 2 shows an example of a combined visualisation and playback system.

5 Experiments

The parameters in the above algorithm have been chosen as follows. The dictionary consists of $s = 5$ vectors with $b = 24$ Mel bands and a cutoff frequency of 2500 Hz. The history size used to learn the dictionary consists of a number h of spectral vectors amounting to 100 seconds of audio. The window size w of the analysis interval is chosen such that it amounts to 0.5 seconds of audio. This is also the step size of the algorithm, i.e., a decision whether or not an abnormal event is happening is taken every 0.5 seconds. These parameters have been found experimentally on a development set. With these settings, the algorithm runs in real-time.

Tests have been performed on recordings from three consecutive days on the same train route. Each recording lasts for about 16 minutes. It is difficult to get a complete ground truth annotation for this kind of data. We have therefore followed an approach that gives an impression of the precision of our algorithm. However, an exact judgement of the recall cannot be made, yet.

We have first run the event detector on the data and then annotated all events reported by the algorithm. Here,

Table 1: Event classes detected in the test set and their relevance for the application scenario.

Event	Relevant
Acceleration	yes
Announcement	yes
Bridge	yes
Change in sound of tracks	yes
Click	yes
Cough	no
Deceleration	yes
Doors	yes
Engine	yes
Laughter	no
Noise	no
Standing	yes
Steps	no
Tunnel	yes

an event is characterised by a time interval in which the detection function is above a fixed threshold found on a development set. An example for this thresholding is given in Figure 3. A human annotator has listened to all these portions of the audio data and has described the contents of such portions by an event name. Table 1 gives a list of all event names that have been attributed to the detected events and states whether or not we judge them as relevant for a public transport scenario. The guiding principle in this has been that only those events may be interesting for public transport management that come from either the train or the supporting infrastructre such as tracks. Most of the events that we regard irrelevant are caused by passengers. These could be avoided entirely by a different placing of the microphone.

Based on this annotation, we have counted the number of events per recording and compared those to the number of relevent events, giving the precision of our method on this data set. Each portion of the audio in which the detection function is steadily above the threshold has been counted as an event and has been annotated by an event name. Due to the nature of our algorithm, the beginning of an event is marked by changes in the acoustic situation: the current spectrogram cannot be well predicted by the previous spectral vectors anymore. The end of an event may have two different causes. First, an event may end because its sound ceases to be physically present. Second, the algorithm may become accustomed to the event. As an example, when a train enters a tunnel, the sound scene changes accordingly. This is recognised by the algorithm because the new spectrogram cannot be explained by spectral vectors from earlier in the recording where the train has not been in the tunnel. If the train stays in the tunnel sufficiently long, considerable parts of the history used to learn the dictionary will contain the sound characteristic of the tunnel. This will make it possible to explain this sound from the history. Thus, even though the tunnel sound may persist, it has become normal and is no longer reported as an abnormal event. In this way, the size of the history influences what is detected as an event and when a persistant event is no longer detected.

Table 2 gives an overview of the number of events detected in the data set and the number of relevant events. In each of the recordings the algorithm has detected about 100 events of which between 70% and 92% were confirmed to be relevant. This gives a mean alarm rate of one alarm in every 10 seconds. In total, over three days, 79% of the events detected in the test set are relevant. This shows, that the algorithm has good precision. Qualitative listening tests show that recall is also high but this cannot be quantified, yet.

Table 2: Number of events and relevant events detected per recording in the test set.

Recording	Day 1	Day 2	Day 3
Number of events	100	103	108
Relevant events	92	73	83
Precision	92%	70%	76%

6 Conclusion

We have provided first promising results for an uninformed abnormal event detection procedure that does not rely on suitably tailored training material but rather operates on an adaptive fixed-size history of an audio stream. In its current state, the algorithm allows to detect abnormal events in a train with rather high precision.

Giving an alarm every 10 seconds in average, the alarm rate is too high to be used as direct feedback for a human surveillance operator. Therefore, in a next step, we plan to correlate the events with position information derived from GPS data, in order to accumulate evidence from several sources. This has the advantage that known event sources like tunnels or train stops can be identified as such and human operators can easily focus on relevant places or track segments. Already, we can robustly derive the locations of train stops automatically from the sound of closing doors accumulated from multiple rides on the same route. By placing the microphone in a more suitable position for infrastructure monitoring, events generated due to sounds from passengers can be avoided completely.

Beyond this, relevance feedback can be incorporated to reduce the number of events not indicating issues with the transport infrastructure. On the one hand, this allows to train model-based event recognisers for irrelevant events. On the other hand, using example-based or model-based approaches can facilitate finding more events of a specific type that have once been reported by the event recognition algorithm.

Finally, our method is entirely generic. Thus it would be very interesting to evaluate its behaviour in other application domains.

Acknowledgments

This work has been partly funded by the European Community's Seventh Framework Programme (FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics – STREP) under grant agreement n° 231738.

We would like to sincerely thank Georgi Dzhambazov for the valuable suggestions and implementational work.

References

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 1306–1309, july 2005.

[2] A. Sasou, K. Tanaka, S. Tanaka, and M. Tanimoto, "Acoustic based abnormal event detection using robust feature compensation," in *TENCON 2011 - 2011 IEEE Region 10 Conference*, pp. 255–258, nov. 2011.

[3] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 634–637, july 2005.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Identification of abnormal audio events based on probabilistic novelty detection," in *INTERSPEECH*, pp. 2218–2221, 2010.

[5] S. Lecomte, R. Lengelle, C. Richard, F. Capman, and B. Ravera, "Abnormal events detection using unsupervised One-Class SVM - Application to audio surveillance and evaluation," *Advanced Video and Signal Based Surveillance, IEEE Conference on*, vol. 0, pp. 124–129, 2011.

[6] V. T. Vu, F. Bremond, G. Davini, M. Thonnat, Q. C. Pham, N. Allezard, P. Sayd, J. L. Rouas, S. Ambellouis, and A. Flancquart, "Audio-video event recognition system for public transport security," in *Proceedings of ICDP*, (London, UK), June 2006.

[7] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *9th International IEEE Conference on Intelligent Transportation Systems (ITSC'2006)*, (Toronto, Canada), 2006.