# Audio Surveillance through Known Event Classification

*František GRÉZL, Jan ČERNOCKÝ*

Dept. of Computer Graphic and Multimedia, Brno University of Technology, Božetěchova 2, 612 66 Brno, Czech Republic

{grezl,cernocky}@fit.vutbr.cz

**Abstract.** *The way of audio surveillance through known event classification is presented introducing simple yet efficient framework. The use of the proposed system for unknown event detection is also suggested and evaluated. Further, a specific audio event is detected with use of audio classification, which helps the detection to focus on a signal of specific behavior. Thus it is shown that the system can be used in several applications.*

## Keywords

Sound classification, event detection, audio surveillance.

## 1. Introduction

Currently, audio signals are making their way to public surveillance. For example, microphones are installed in locations with high crime done by street gangs [6]. The basic advantages and disadvantages of audio surveillance can be demonstrated on this use-case: the advantage of audio is, that the sensors are quite small and can be installed practically anywhere, their low power-consumption allows local power-supply including wireless data transfer. The downside is that the signals from many microphones cannot be presented to one operator simultaneously which is easily doable for video by mosaic screens.

In case of large number of sensors, automatic processing is needed to either point out potentially interesting parts of a signal or to directly analyze the signal and trigger an appropriate action. The first case may include the speech-nonspeech segmentation, or, more generally, pointing out segments which are different from usual background. This can be done, for example, by auditory attention levels [3] which are designed to capture sudden and unexpected changes of audio texture and highlight this sound to the operator. The second case can be represented for example by gunshot detector [4], which is able to localize the source of the sound, focus the camera in the direction and in addition, it may call the police.

The processing we are introducing here is more related to the analysis of the signal. Our goal is to classify events in audio signal captured in public area. We make use of our experience from automatic speech recognition and employ supervised learning to this task. With this system, the usual events appearing in normal operation are detected. The way of using this approach also for identification of exceptional sounds is suggested and evaluated.

## 2. Definition of the Experiment

The CARETAKER project focused on audio-visual surveillance of public transportation, namely the metro system. The audio surveillance should complement the video one in such cases that cannot be captured by video or are difficult to detect from it.

In the first stage of the project, 298 hours of audio signal captured by two microphones simultaneously (596 hours together) were collected. The signal was sampled at 16 kHz and stored in 16 bit PCM samples. These data were recorded over several days in normal metro operation ensuring large variety of the acoustic scene and possibly capturing most of its normal behavior.

In addition to these live recordings, two acted scenarios were recorded. Here, gun shots, loud screams and hard bangings were recorded after the closing hour of the metro, which means that these sounds appear on silent background.

### 2.1 Labeling of Data

Our task in CARETAKER project was to classify usual audio events and thus monitor the audio scene to complement the video surveillance, and also to help to identify unexpected sounds.

As mentioned above, supervised training approach was used to train our classifier. This requires the selection and labeling of training and test data.

Twelve classes were chosen: three of them linked to the metro train: `train moving`, `breaks` and `horn`, two linked to the station operation: `announcement` and `beep` and other six were general audio events occurring in the recording: `banging`, `click`, `hiss`, `pop`, `whistle` and `whiz`. When none of these events occurs, the respective part of the signal was labeled as `background`, which actually covered large variety of audio scenes from an empty and quiet station, to a crowded station and station with train in it (which significantly increased the level of the signal). This is in contrast with video surveillance where the background is well defined, especially in interiors with constant lighting. In video analysis, there are objects and their interactions, but in audio, we are able to register only objects that produce sounds (i.e. `moving train` or `announcement`) or the interaction between objects, which is manifested by a sound (i.e. `banging` which means that one object hits another, or `whiz` which may occur when someone opened door that made a whiz sound).

Co-occurrence of audio events is possible and actually very frequent. During the labeling process, such co-occurrence was either discarded from the labeled data (mostly cases when two short events like `banging` and `click` occur at the same time) or the stronger event was labeled (i.e. `horn` sound when `train moving` is around).

Twenty minutes of audio data from each microphone were annotated using the above labels. Then a classifier (see Sec. 2.3) was trained. The trained classifier was used to automatically label another four hours of data, two hours from each microphone. These data were manually corrected. Large confusion between labels `whiz` and `whistle` was observed in such a way, that almost all events of both labels were assigned to `whistle`. For this reason, we dropped the label `whiz` from our label set. The corrected four hours of data were used to train the final classifier. Additionally, twenty minutes of audio data were taken from each microphone, were processed in the same way as the training data and served for testing. Note that when taking audio for training/testing, the selected segments from different microphones did not overlap in time.

## 2.2 Acoustic Features

As acoustic parameters, critical band energies (CRBE) [1] – a de-facto standard for speech recognition – were chosen. These parameters were designed according to psycho-acoustic studies on human sound perception in general, so we believed that these features would be suitable also for recognition of general audio events.

The Critical band energies are computed using toolkit HTK [1] and their main characteristic are:

- The bands are spaced linearly along the Mel axis,
- triangular bands are used with half-band overlap,
- the number of bands is 23,
- the length of analysis window is 25 ms, parameterization is done every 10 ms.

## 2.3 Classifier

Hidden Markov Models (HMM) were used to transcribe an audio signal in terms of chosen labels. A model with three states was constructed for each audio event. All states for one event shared the same probability distribution, the number of states ensures minimum duration of a given event and prevents fast oscillation between overlapping events or on the boundary of events.

The probability distribution was estimated using a neural network (NN). The NN was preferred over Gaussian mixture model because its outputs are directly interpretable as probabilities. It provides the opportunity to directly see the probability of given class independently of the HMM classifier. Further, we can obtain the probability of the event almost immediately (the delay is given by the parameter stacking on the input to the NN, the forward pass through NN is very fast), whereas the classification by HMM is presented only when the sequence of events ends.

The input to the NN is composed from several consecutive frames of CRBE. The time evolution of energy in each critical band is processed independently similarly to [5], namely there is Hamming windowing and Discrete Cosine Transform (DCT). In this setup, 31 frames of CRBE were used, 15 frames ahead and 15 frames after the current frame. After the Hamming windowing, 16 DCT bases were applied including the $0^{\text{th}}$ one. Together, there were $16 \times 23 = 368$ inputs to the NN. A three layer neural network was trained with standard back-propagation algorithm and it has about 10 000 parameters in total.

The scheme of the whole processing is depicted in Fig. 1.

## 2.4 Definition of Test Sets

As mentioned above, 40 minutes of audio data were automatically transcribed and manually corrected to serve as a test set for known event classification. This set is referred to as *test A*.

However, we also aimed to test the proposed techniques for detecting unknown events. For these purposes, acted scenarios were recorded with gun shots, screams and a pretended vandalism on a vending machine manifested by loud banging sounds. The two records have lengths of 10 minutes and 7 seconds and 6 minutes and 40 seconds, respectively. The recordings were concatenated (with small gap) and expanded to 20 minutes. The parts of the resulting audio signal which was not covered by the acted recordings were taken from *test A* signal. Finally, the audio was manually annotated and label `unexpected` was used for screams and gun shots. This set is referred to as *test B*.

The acted scenarios were recorded at night after closing hour of the metro which makes the background unrealistically quiet. To have the unknown events in normal audio ambiance of a metro station, the acted audios were summed with *test A* signal. Values that overflow the valid dynamic range are limited to the maximum/minimum values. The test set was transcribed starting from *test A* transcription and manually corrected as the new signal was added. Again, the transcription contains `unexpected` events. This set is referred to as *test C*.

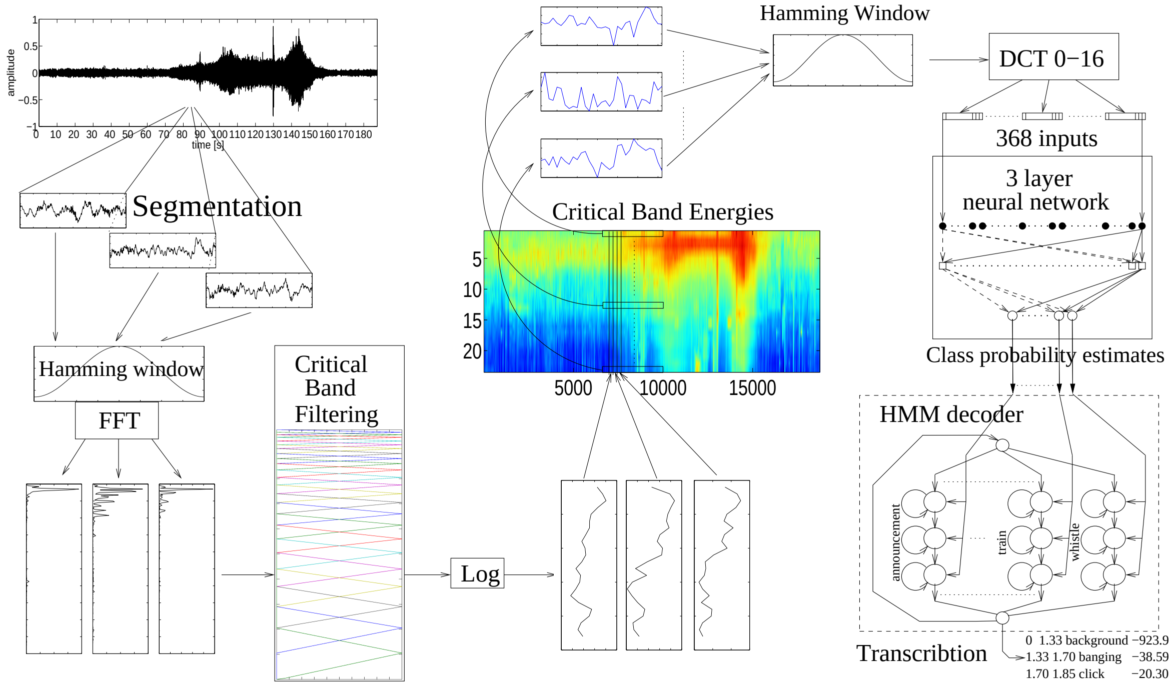

**Fig. 1.** Scheme of audio signal processing from the feature extraction and NN input preparation to class probability estimation and classification.

## 3. Known Event Classification

The evaluation of known event classification was done on 40 minutes of manually labeled data, 20 minutes from each microphone (*test A*). The standard metric for automatic speech recognition, word error rate (WER), was used (in our case "word" means audio event):

$$WER = 1 - \frac{N - Del - Sub - Ins}{N} \times 100[\%]$$

where $N$ is the total number of events in test set, $Del$ is the number of not recognized events, $Sub$ gives the number of substitution errors and $Ins$ stands for the number of inserted events. The obtained results are shown in Tab. 1.

The results for tests B and C are only indicative, as here, we are evaluating the performance of known event recognition and there are unknown events in the audio. Note, that the recognizer parameters such as *word insertion penalty*[1] were tuned on the test set.

## 4. Unexpected Event Detection

The unexpected event detection is usually based on learning the background and spotting events which somehow differ from it. For example in [3], the time-frequency representation is computed first, in a very similar way to our CRBEs, mean, variance and the third and fourth order statistics are computed together with other spectral statistics over the audio samples, and delta features are added. These coefficients are normalized over 3 or 10 seconds of the signal. Finally, each segment is thresholded to decide if there is an unexpected event or not.

Another possible approach was designed for spotting *out of vocabulary* words in automatic speech recognition [2]. It uses the full power of Large Vocabulary Continuous Speech Recognition (LVCSR) system and combines its posteriors with posteriors obtained from a weakly constrained phone classifier.

The approach suggested in this paper makes use of the trained classifier although it is far less powerful then classifiers used for automatic speech recognition. The motivation comes from the fact that outputs of the NN can be considered as posteriors of individual events. Then, if a previously unseen input is presented to the NN, its output should be uncertain and it should assign more or less the same probability to all classes. The measure of uncertainty is the entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i), \qquad (1)$$

where $n$ is the number of classes. The higher the value of $H(X)$ is, the more uncertain the classifier is about its input.

| Microphone | 1 | 2 | whole set |
|---|---|---|---|
| test A | 46.7 | 32.2 | 38.2 |
| test B | 59.5 | 54.6 | 56.5 |
| test C | 52.8 | 41.3 | 45.4 |

**Tab. 1.** WER [%] for audio signals from individual microphones and over whole test set.

[1] Penalty that is added to the likelihood in the decoder when switching from one event to another.

## 4.1 Enhancing the Entropy Evolution

When the frame-by-frame entropy of the test signal is plotted, it appears to be very noisy. To smooth the values, two simple techniques were employed:

**filtering** – the entropy function was filtered by a simple low-pass filter. This operation removed most of the high-frequency components (noise).

**normalization** – this operation removes the dependency on the absolute levels and lets the detector concentrate on the changes in dynamics. This is relevant especially when the sound ambiance changes (from quiet station to crowd of normally behaving people). Technically, this is done by $H[n] = H_{50}[n]/H_{1000}[n]$ where $H[n]$ is the normalized entropy, $H_{50}[n]$ is an average of entropy over 50 past frames and $H_{1000}[n]$ is similar value but computed over a history of 1000 frames. We have also experimented with the on-line estimation of mean values with simple recursive filters but the results were about the same.

Note that both techniques use only the past values to allow on-line processing and real-time detection of unexpected events.

## 4.2 Evaluation of Unknown Event Detection

Detection of unknown events is a basic classification task and as such it was evaluated using standard metrics for bi-valent results, based on frames (e.g. each 10 ms). The results are presented by Detection Error Trade-off (DET) curves that show the dependency of the probability of false positives on the false negative errors with the detection threshold as parameter. All tests are performed on merged *test B* + *test C* set which is closed to real operating conditions of the system (*test B* represents quiet station and *test C* a crowded one). The DET curves obtained with entropies are shown in Fig. 2.

## 4.3 Repeated Banging

During the annotation phase, we encountered a moment, which sounds like if someone is trying to call a service person by hard knocking on a kiosk. This actually lasted quite long and according to the audio, no one appeared and the metro user left unsatisfied.

The detection of such event could be valuable, because it may call an authorized person to come when his presence is demanded.

We approach this problem from the point that we know what we would like to detect – `repeated banging`. A `banging` event can be observed at three points in our system: increased energy of input signal, the probability value of corresponding NN output and the classifier output.

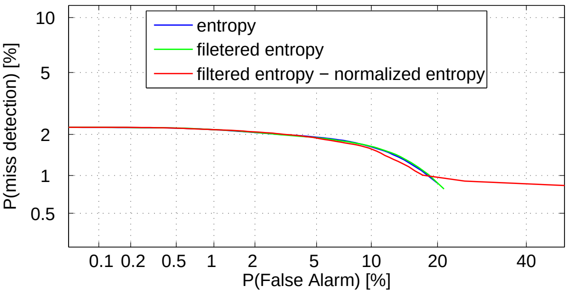

**Fig. 2.** DET curves obtained with entropy based unknown event detection.

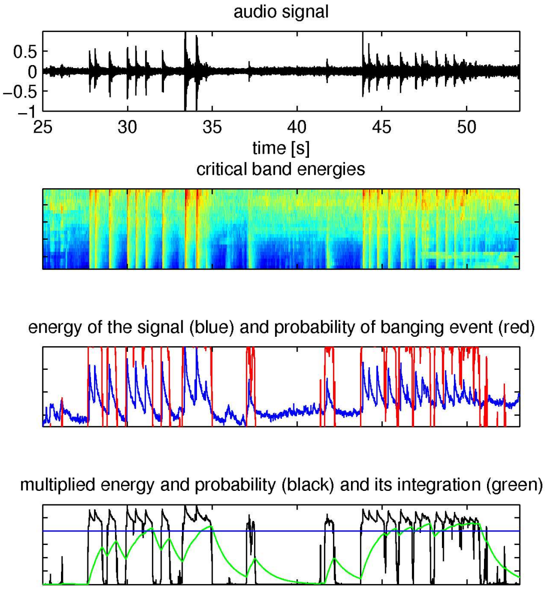

**Fig. 3.** The `repeated banging` event. The audio signal is shown on top, the critical band spectrogram is presented below. Third pane shows the evolution of energy and probability of the `banging` event and the lower pane shows energy multiplied by the probability and its integrated version.

The idea of using classifier output and count number of `banging` events was dropped because the same result could be produced by rolling of a suitcase wheels on the metro tile floor. To prevent triggering this alarm based on week `banging`, the energy has to be employed. Instead of final classification, which gives a hard decision about the output class, the probability of the class `banging` was used to soften our function. The probability of the class was then multiplied with the energy and resulting signal was filtered by and integration filter with long time constant. This prevents the detection of short and hard banging such as shutting the train doors. The threshold was set so that only long duration of co-occurrence of high signal energy and `banging` event probability will trigger the desired action.

The case study is depicted in Fig. 3.

# 5. Conclusions

A simple but efficient framework for analyzing general audio by known event classification was introduced. The main concern of the work was to develop the known event detection as a part of the CARETAKER project framework. The presented results are satisfactory for the purpose of audio surveillance, mainly when used in such a variable environment such as metro hall.

The system performance could be further improved if the classifier would be allowed to output two parallel hypotheses when it is not sure about the correct class. This would help specially in cases of overlapping events or when one event is slowly changing into another as discussed in Sec. 2.1.

The possibility of using the classifier in a mode of unknown event detector was introduced. The fact that such event is unknown to the classifier is utilized and entropy of neural network outputs is computed. This measure then serves for decision whether there is an unknown signal at the input. Note, that the decision is made for each frame independently, which in our case covers about 300 ms. This is far less than other proposed techniques are using.

Finally a case study of detection of more specific event – `repeated banging` – was presented. In this scenario, the advantage of using an event classifier which allows to concentrate only on a signal with specific behavior, can be clearly seen. Such analysis would not be possible without knowledge of the signal class.

# Acknowledgements

This work was partly supported by European project CareTaker (FP6- 027231), by Grant Agency of Czech Republic project No. 102/08/0707, and by Czech Ministry of Education project No. MSM0021630528. František Grézl was supported by post-doctoral project of Grant Agency of Czech Republic, No. 102/09/P635.

# References

[1] YOUNG, S., JANSEN, J., ODELL, J., OLLASON, D., WOODLAND, P. *The HTK book*. Cambridge: Entropics Cambridge Research Lab., 2002.

[2] BURGET, L., SCHWARZ, P., MATĚJKA, P., HANNEMANN, M., RASTROW, A., WHITE, C., KHUDANPUR, S., HEŘMANSKÝ, H., ČERNOCKÝ, J. Combination of strongly and weakly constrained recognizers for reliable detection of OOVs. In *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP*. Las Vegas (USA), 2008, p. 4081 - 4084.

[3] COUVREUR, L., BETTENS, F., HANCQ J., MANCAS, M. Normalized auditory attention levels for automatic audio surveillance. In *Proc. Second International Conference on Safety and Security Engineering SAFE*. Malta, 2007.

[4] VALENZISE, G., GEROSA, L., TAGLIASACCHI, M., ANTONACCI, F., SARTI, A. Scream and gunshot detection and localization for audio-surveillance systems. In *IEEE Conference on Advanced Video and Signal Based Surveillance AVSS*. London (UK), 2007, p. 21 - 26.

[5] GRÉZL, F., ČERNOCKÝ, J. TRAP-based techniques for recognition of noisy speech. In *Proc. $10^{th}$ International Conference on Text Speech and Dialogue TSD 2007, vol. 9*. Pilsen (Czechia), 2007, p. 270 - 277.

[6] O'BRIEN, T. *Public Audio Surveillance Hits London*. [Online] Cited 2009-10-21. Available at: http://www.switched.com/2007/10/23/london-police-love-their-surveillance/

# About Authors...

**František GRÉZL** (1977) received the Ing. (MSc.) and Ph.D. degrees from Brno University of Technology (BUT) in 2000 and 2007 respectively. He has worked with OGI (USA), IDIAP (Switzerland) and ICSI (USA) and is now with BUT's Faculty of Information Technology (FIT) as assistant professor. His research interests include feature extraction for speech recognition, mainly based on neural networks.

**Jan ČERNOCKÝ** (1970) received Ing. (MSc.) degree from BUT and Dr. (Ph.D.) degree jointly from Université Paris XI and BUT. He has been with ESIEE (France) and OGI (USA). Currently, he is associate professor and head of Department of Computer graphics and multimedia at FIT BUT. His research interests span numerous topics from speech and signal processing.