IMPROVING EVENT DETECTION FOR AUDIO SURVEILLANCE USING GABOR FILTERBANK FEATURES

Jürgen T. Geiger and Karim Helwani
juergen.geiger@huawei.com
Huawei European Research Center, Munich, Germany

ABSTRACT

Acoustic event detection in surveillance scenarios is an important but difficult problem. Realistic systems are struggling with noisy recording conditions. In this work, we propose to use Gabor filterbank features to detect target events in different noisy background scenes. These features capture spectro-temporal modulation frequencies in the signal, which makes them suited for the detection of non-stationary sound events. A single-class detector is constructed for each of the different target events. In a hierarchical framework, the separate detectors are combined to a multi-class detector. Experiments are performed using a database of four different target sounds and four background scenarios. On average, the proposed features outperform conventional features in all tested noise levels, in terms of detection and classification performance.

Index Terms— Audio surveillance, event detection, Gabor features, noise robustness

1. INTRODUCTION

Automatic surveillance systems are becoming more and more ubiquitous in public spaces. Audio analysis can complement video-based systems, which are exposed to several vulnerabilities, such as occlusions. Systems that analyse audio signals can successfully be combined with video solutions, or used in a stand-alone manner [1]. Relevant tasks that can be solved by audio analysis are abnormal event detection (such as gunshots or explosions) and classification, as well as source localisation and tracking. The problems that an audio analysis system has to face include high amounts of non-stationary background noise and a strong diversity of potential interesting sound events.

This paper deals with sound event detection in highly realistic noisy environments. Several previous studies addressed the problem of detecting and classifying acoustic events such as gunshots, explosions, or screams. Most of the proposed systems rely on the traditional approach of modelling Mel-frequency cepstral coefficient (MFCC) features with Gaussian mixture models (GMMs) or hidden Markov models (HMMs) and explore different system setups or different additional audio features. Our work goes in the same direction, with the goal of creating a robust system that can operate in realistic environment.

1.1. Related Work

Over the last years, several studies evaluated systems for event detection in surveillance scenarios. Several of the proposed systems use classical spectral features in a GMM or HMM framework. In [2], six sound event classes (including human screams, explosions, and gunshots) are detected with a median filter and classified using linear spectral band features and either a GMM or HMM classifier. The system showed solid recognition rates in white and musical background noise. Clavel *et al.* used MFCCs and other spectral features (spectral centroid and spread) together with a GMM classifier to detect gunshot sounds in recordings of public places [3]. A similar system is used in [4] to detect scream and gunshot sounds, and small improvements were obtained by adding more features, most notably spectral distribution features (e. g. spectral slope or spectral roll-off) and correlation-based features. A two-stage approach is proposed in [5]: an audio signal is first classified as normal or abnormal, followed by a maximum-likelihood classification to determine the class. This work relies again on MFCC features and an HMM classifier. In [6], different gunshot detection algorithms are compared, with the conclusion that correlation and wavelet-based detection algorithms give higher performance. A bag of aural words classifier was used in [7] to classify acoustic events in surveillance scenarios. In [8], wavelet features are proposed for environmental sound classification. The general problem of event detection in surveillance scenarios is that almost no realistic databases are available. In all of the mentioned studies, databases were created by mixing target sound events into background recordings. Furthermore, most of the previous studies rely on techniques that were originally designed for speech processing. There is still a lack of features and classification models that are specifically tailored to the underlying problem.

Acoustic event detection systems are also used in other environments. In the CLEAR [9] and D-CASE [10] evaluations, the goal was to detect acoustic events in a domenstic environment. In [11], acoustic event detection was per-

The research leading to these results has received funding from the European Commision Union Seventh Framework Programme (FP7/2007/2013) under grant agreement 607480 LASIE.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

formed on real-life recordings. The UrbanSound dataset, another database of real-life recordings is described in [12]. An interesting approach for event detection that goes beyond the classical spectral features are the spectrogram image features proposed by Dennis *et al.* [13]. Relevant information is extracted by regarding the spectrogram as an image. These features achieved good results in the similar problem of noise-robust acoustic event classification.

1.2. Contributions

The goal of the present study is to construct a noise-robust event detection system for surveillance scenarios. While most of the previous studies in the field rely on classical spectral and cepstral audio features (mostly MFCCs), we investigate the suitability of a Gabor filterbank feature set. The employed Gabor filterbank features are physiologically inspired and were originally proposed for noise-robust speech recognition [14]. These features extract spectro-temporal modulation frequencies from the signal by filtering the Mel spectrogram with different Gabor filters. The use of such features is motivated by the finding that a similar processing is performed in the primary auditory cortex of mammals [15]. In a recent challenge for acoustic event detection in an office environment, these features achieved a good detection performance [16].

Acoustic events are modelled with GMMs, and single-class detectors for noisy environments are created. A hierarchical system setup is used to distinguish between different event classes, in order to arrive at multi-class detection system. Experiments are carried out using recordings of *breaking glass*, *explosion*, *gunshot*, and *scream* sounds. Target sounds were mixed into realistic background scene recordings. The experimental evaluations reveal that the proposed GBFB features achieve better results than MFCC features, in terms of event detection and classification.

The rest of the paper is organised as follows. In Section 2, the framework of the event detection system is delineated. The employed audio features are described in detail in Section 3. Experimental results are presented in Section 4, followed by some conclusions in Section 5.

2. EVENT DETECTION SYSTEM

The proposed event detection system is composed of single-event detectors. For each target event, a detector is trained. Each event detector consists of a two-class GMM classifier, one model, θ_1, for the target event and one, θ_2, for the background noise. The GMMs are trained with diagonal covariances, and the number of mixture components is fixed to 16 (following preliminary experiments). For a given unknown sample $X = x_1, \ldots, x_T$, where T is the length in frames, the log-likelihood for both models

$$L_i = \log P(X|\theta_i), i \in \{1, 2\} \quad (1)$$

is evaluated. The log-likelihoods are used to derive a detection score

$$\phi = L_1 - L_2. \quad (2)$$

Together with a threshold, this score can be converted to a detection decision.

The same detection framework can be used for single-event and multi-event detection. For single-event detection, a detection score is obtained as described above. In order to perform multi-event detection, a hierarchical system setup is used. After obtaining the scores for each single-event detector, maximum-likelihood classification between all target event models is performed to obtain one result.

The given problem of event detection in surveillance scenarios differs from other acoustic event detection scenarios. In most other event detection scenarios, precise timing information is important, regarding the onset and offset of events. On the other hand, in surveillance scenarios, exact timing (in the order of several frames) is not required. Therefore, we evaluate the system with pre-segmented recordings, instead of detecting events in longer background recordings.

3. AUDIO FEATURES

As a baseline, MFCCs are used as features. In previous studies about event detection for acoustic surveillance, MFCCs achieved a good performance in combination with GMM classifiers. 13 MFCC coefficients are computed for each frame of 25 ms length (frame shift 10 ms). Together with delta and delta-delta coefficients, this results in a 39-dimensional feature vector per frame.

As an alternative, we propose to use Gabor filterbank (GBFB) features. This feature set models the spectro-temporal modulation frequencies in the signal and it was recently proposed for noise-robust speech recognition [14].

The selection of features based on Gabor filterbank is motivated by the fact that it provides a systematic approach to describe the spectro-temporal characteristics of a signal and it offers the benefits of the wavelet analysis by analysing the signal at different scales. This results in an optimal time-frequency localization [17]. Therefore, global as well as localised characteristics of the temporal as well as the spectral structure of the signal can be gathered. Since the Gabor filterbank is defined in the spectro-temporal domain, it can be adjusted to capture features in the time domain only, the spectral domain or features related to dependencies of the spectral excitation with respect to the time. This is useful to characterise many sound events where the frequency excitation structure follows a specific chronology. For example, a detonation causes shock waves which have specific spectro-temporal structure caused by increased pressure of the air, its local temperature and the local speed of sound. Hence, in an initially plane sinusoidal wave of a single frequency, the peaks of the wave travel faster than the troughs, and the pulse becomes cumulatively more like a sawtooth wave [18]. As

another example, although the sound of a gun shot depends on the gun, however, the evolvement of the excited frequencies has similar time dependency for the majority of guns.

To extract GFB features, first, the log Mel-spectrum of a signal (25 ms frames with 10 ms frame shift) is computed. This spectrum is filtered by a Gabor filterbank. Each Gabor filter is defined as the product of a 2-dimensional sinusoid carrier (3) with corresponding temporal modulation frequency ω_k and spectral modulation frequency ω_n, and an envelope function (4):

$$s_\omega(x) = \exp(i\omega x), \tag{3}$$

$$h_b(x) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi x}{b}\right) & -\frac{b}{2} < x < \frac{b}{2} \\ 0 & \text{else.} \end{cases} \tag{4}$$

The parameter b controls the width of the carrier function. Each of the sinusoid carriers corresponds to a specific temporal and spectral modulation frequency. The maximum size of the filters is limited to 69 frequency channel and 40 time frames. The filterbank is designed to consist of 41 Gabor filters (with different temporal and spectral modulation frequencies). Each of these filters can be applied to each of the 23 frequency channels. From the 943 possible combinations, a number of representative channels is selected. This reduces the filterbank output to 311 dimensions. These settings correspond to the original definition of the GBFB features [14] and are used throughout the present work.

Figure 1 illustrates the log Mel-spectrogram and the output of one Gabor filter for two exemplary recordings of the classes *breaking glass* and *gunshot*. The *breaking glass* recording has only few low-frequency components, while the *gunshot* recording reveals a considerable amount of low-frequency components. In addition to the spectrograms, the output of the Gabor filter corresponding to a spectral modulation frequency of 0.06 cycles per channel and a temporal modulation frequency of 6.2 Hz is shown in order to illustrate the extracted features. Considerable differences between the two different classes are visible in the figure in terms of the spectral distribution, as well as characteristic properties within the recording of the same class.

Applying the 2-dimensional Gabor filterbank can also be understood as an image filtering process on the spectrogram. With the spectro-temporal extent of the filters, spectral and temporal context is incorporated in the resulting features. Spectral modulation frequencies of up to 0.25 cycles per channel and temporal modulation frequencies of up to 25 Hz are captured with the filterbank. Exploiting this information seems to be promising for the task of event detection in surveillance scenarios, since the target sounds are not assumed to be stationary.

In order to compare GBFB features to MFCCs, a principal component analysis is applied to reduce the dimensionality of the GBFB features to 39. As a consequence, the same model order can be applied for both feature sets. In the experiments, the PCA basis is always computed from the training data, and test data are projected onto this basis.

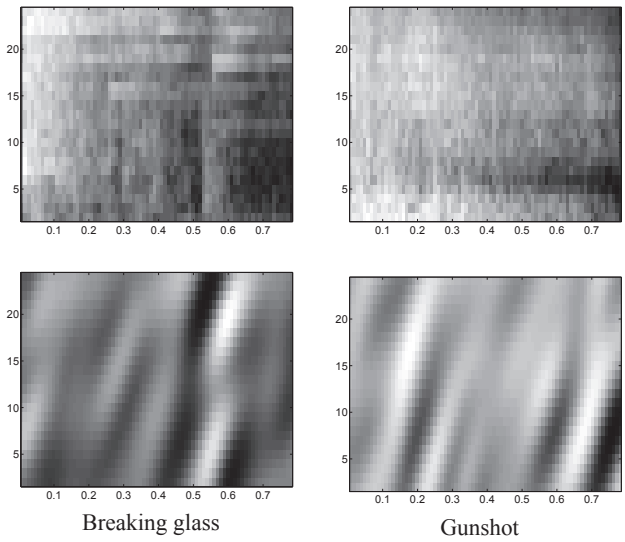

Fig. 1. log Mel-spectrogram (top) and output for one Gabor filter (bottom) for two recordings. The x-axis represents the time in s and the y-axis the Mel-frequency channels.

4. EXPERIMENTS

4.1. Database

There are no standard publicly available databases for acoustic surveillance scenarios. As in most of the previous studies on event detection for audio surveillance, we created our own evaluation database. Different classes of target sounds were mixed into realistic background recordings at various signal-to-noise ratios (SNRs). As target sounds, we considered the four classes *breaking glass*, *explosion*, *gunshot*, and *scream*. For each of these classes, we collected 100 samples from the public repository `www.findsounds.com`. Background sounds were chosen from the database of acoustic scenes from the D-CASE challenge [10]. The classes *busystreet*, *openairmarket*, *park*, and *tubestation* were selected as potential scenarios for audio surveillance. Although the background recordings in the D-CASE database are available as binaural recordings, only the left channel was used, to simulate a simple, realistic single-microphone setup. Target sounds were mixed into background recordings at different SNR values from 20 dB to 0 dB, in steps of 5 dB. The detection models are trained with matched noise settings, i. e., for each background noise, a separate model is trained.

In order to provide negative samples for the detection experiments, consisting only of background sounds, extracts are cut from the background recordings. It was found that the length of the target sounds follows a Gamma distribution. Background samples were extracted with a length randomly

Table 1. Event detection equal error rate, comparing MFCC and GBFB features

	SNR (dB)						
	0	5	10	15	20	∞	mean
MFCC	23.1	17.2	8.6	5.6	4.1	1.7	10.1
GBFB	22.5	13.8	7.5	4.5	3.1	1.3	8.8

drawn from a Gamma distribution with shape and scale parameters adjusted to the length distribution of the target sounds.

In total, the created database consists of 8 800 samples: four target classes with 100 samples each, mixed with four different backgrounds at five SNR values, together with the clean recordings; furthermore, four background classes with 100 samples each. The database is divided into a training set (60 of each of the 100 samples) and a test set (the other 40 samples). For training, only clean recordings of the target events as well as background recordings are used, while tests are performed for all SNR values in addition to the clean recordings.

4.2. Event Detection

Firstly, event detection is evaluated separately for each class of acoustic event. The task is to detect target events in a background recording. Therefore, evaluation can be carried out in terms of false detections and false rejections. For each recording, an event detector yields a detection score, which can be used, together with a detection threshold, to trade off false detections and false rejections. Results for different detection threshold can be plotted in a detection error trade-off (DET) curve. The equal error rate (EER) is used as a universal performance measure in this work. It is defined as the operating point with equal false detection and false rejection rates.

Detection experiments are performed separately for each of the backgrounds, in matched conditions. This means that the detector is trained and tested with the same background class. The results can be averaged over all backgrounds and over all target sounds, to obtain one averaged EER per SNR value. Table 1 shows these results, for MFCC and GBFB features. GBFB features achieve better results than MFCCs for all SNR values. On average, using GBFB instead of MFCC leads to a relative performance improvement of 13 %.

Table 2 reports results (for GBFB features) separately for each of the target classes. As could be expected, *breaking glass* and *scream* are easier to detect, since for the other two classes, confusions with background sounds are more likely because of the low-frequency noise-like structure.

Results for the separate background classes are given in Table 3. The worst results are obtained with the background class *openairmarket*. This class contains a wide range of diverse sounds, which could lead to the relatively high error rates. For the background class *park*, the error rates are also relatively high. On average, the *park* recordings are relatively quiet, which means that they have to be amplified unnaturally strong in order to arrive at certain SNR values. Realistically, the SNR values would be much higher in a *park* environment compared to, for example, *busystreet*.

Table 2. Event detection results (EER) separately for each target event, using GBFB features

	SNR (dB)						
	0	5	10	15	20	∞	mean
Glass	22.5	14.4	9.4	4.4	1.9	0.0	8.8
Expl.	22.5	15.6	8.8	6.9	5.0	3.8	10.4
Gun	26.9	16.9	9.4	6.3	5.0	1.3	10.9
Scream	18.1	8.1	2.5	0.6	0.6	0.0	5.0

Table 3. Event detection results (EER) separately for each background class, using GBFB features

	SNR (dB)						
	0	5	10	15	20	∞	mean
Street	16.3	8.8	2.5	0.6	0.0	0.0	4.7
Market	33.1	19.4	10.6	6.3	3.1	1.9	12.4
Park	21.9	15.0	10.0	6.3	5.6	2.5	10.2
Tube	18.8	11.9	6.9	5.0	3.8	0.6	7.8

To illustrate the detection result of one exemplary experiment, Figure 2 shows the DET curves for the detection of the class *gunshot* in *tubestation* noise. For an SNR value of 20 dB, the EER is 7.5 %, and for 0 dB, the EER goes up to 25 %.

4.3. Event Classification

In order to perform multi-event detection, the same framework as for single-event detection is used. Results for event classification are shown in Table 4. The comparison shows again that GBFB features achieve a better performance than MFCCs. Only in the case of clean sounds, MFCCs are slightly better, while for all other SNR values, GBFB features perform consistently better. On average, the relative performance improvement from MFCCs to GBFBs is 6 %. The practical advantage of the multiclass detection system is that it uses the same models as the single-class detectors, which are trained with clean data only. Further improvements in classification accuracy are expected with the introduction of concepts such as multi-condition training.

5. CONCLUSIONS

We proposed an event detection system for audio surveillance scenarios. Acoustic events are modelled with GMMs and single-class detectors are trained for different realistic

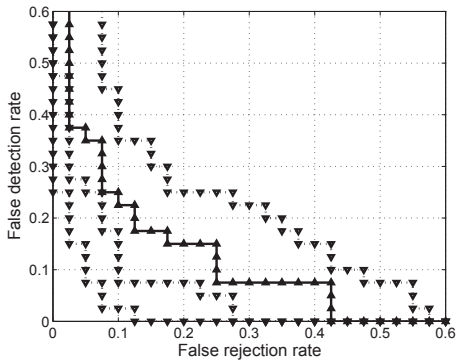

Fig. 2. DET curves for the class gunshot in tubestation noise, for the SNR values of 20, 10, 5, and 0 dB.

Table 4. Event classification accuracy

	SNR (dB)						
	0	5	10	15	20	∞	mean
MFCC	38.1	51.4	66.4	74.2	77.2	93.8	66.9
GBFB	42.3	55.9	68.9	78.9	84.7	93.4	70.7

background noise conditions. As an alternative to the classical MFCC features, we evaluated Gabor filterbank features, which extract spectral and temporal modulation frequencies from the signal. In an evaluation with realistic background recordings in noisy conditions, the proposed Gabor features achieved a better detection and classification performance than the MFCCs. In particular, in the classification experiments, where MFCCs performed slightly better in clean conditions, GBFB features showed a better noise robustness. In this work, matched models with known background sound are assumed. In order to run the system automatically in different background scenes, a combination with a system for acoustic scene recognition makes sense, such as the one presented in [19]. The evaluated GMMs are well suited to model stationary sounds such as *scream*, while for other non-stationary sounds, better models need to be found in future work.

REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *arXiv preprint arXiv:1409.7787*, 2014.

[2] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, "Automatic sound detection and recognition for noisy environment," in *Proc. EUSIPCO*, 2000.

[3] 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*. IEEE, 2005, pp. 1306–1309.

[4] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proc. EUSIPCO*, 2007.

[5] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. ICASSP*, 2009, pp. 165–168.

[6] A. Chacon-Rodriguez, P. Julián, L. Castro, P. Alvarado, and N. Hernández, "Evaluation of gunshot detection algorithms," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 58, no. 2, pp. 363–373, 2011.

[7] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *Proc. AVSS*. IEEE, 2013, pp. 81–86.

[8] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *Information Forensics and Security, IEEE Transactions on*, vol. 3, no. 4, pp. 763–775, 2008.

[9] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*, pp. 311–322. Springer, 2007.

[10] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Proc. WASPAA*, 2013.

[11] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. EUSIPCO*, 2010, pp. 1267–1271.

[12] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Multimedia*, 2014.

[13] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 130–133, 2011.

[14] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[15] A. Qiu, C. E. Schreiner, and M. A. Escabí, "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition," *Journal of Neurophysiology*, vol. 90, no. 1, pp. 456–476, 2003.

[16] J. Schröder, N. Moritz, M. R. Schädler, B. Cauchi, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "On the use of spectro-temporal features for the IEEE AASP challenge detection and classification of acoustic scenes and events," in *Proc. WASPAA*, 2013.

[17] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[18] A. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, Acoustical Society of America, 1989.

[19] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *Proc. WASPAA*, 2013.