Time-frequency analysis for audio event detection in real scenarios

Saggese, Alessia; Strisciuglio, Nicola; Vento, Mario; Petkov, Nicolai

Published in:
2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016

DOI:
10.1109/AVSS.2016.7738082

Time-frequency analysis for audio event detection in real scenarios

Alessia Saggese[1], Nicola Strisciuglio[1,2], Mario Vento[1], Nicolai Petkov[2]
[1] University of Salerno - DIEM, Italy
[2] University of Groningen - JBI, The Netherlands
{asaggese, nstrisciuglio, mvento}@unisa.it, n.petkov@rug.nl

Abstract

We propose a sound analysis system for the detection of audio events in surveillance applications. The method that we propose combines short- and long-time analysis in order to increase the reliability of the detection. The basic idea is that a sound is composed of small, atomic audio units and some of them are distinctive of a particular class of sounds. Similarly to the words in a text, we count the occurrence of audio units for the construction of a feature vector that describes a given time interval. A classifier is then used to learn which audio units are distinctive for the different classes of sound. We compare the performance of different sets of short-time features by carrying out experiments on the MIVIA audio event data set. We study the performance and the stability of the proposed system when it is employed in live scenarios, so as to characterize its expected behavior when used in real applications.

1. Introduction

Automatic interpretation and detection of events of interest have lately attracted several scientists in the field of pattern recognition. Existing approaches traditionally focused on the analysis of video signals acquired by surveillance cameras (245 millions of installed cameras in 2014) [10].

However, in the last years a growing interest towards the analysis of audio signals for surveillance purposes emerged. This is mainly due to the fact that several IP cameras are already equipped with a microphone, thus making inexpensive the employing of audio-based event detection (even combined with video-based event detection [5]) together with existing surveillance infrastructures. Furthermore, there are several kinds of event (e.g. gun shots, screams or glass breaking) that could be very difficult to be recognized by video inspection since their visual appearance does not provide enough useful information. They could be, instead,effectively detected by using audio sensors.

Audio analysis is a very challenging problem. One of the most relevant issues is related to the fact that the sounds of interest are typically mixed to various types of background sound. It makes practically very difficult to separate the sound to be recognized from the background. Furthermore, the properties of the events of interest could be evident at different time scales: consider, for instance, that a gun shot is an impulsive sound, while a scream is a sustained one. In order to face these issues, several methods were proposed in the last years, as reviewed in [6]. Although it is not possible to make a sharp organization of existing methods, two groups can be identified, depending on the complexity of their classification architecture.

The first group contains methods that process small audio frames, extract a set of characteristic features (Mel-Frequency Cepstral Coefficients, Wavelet-based coefficients, etc.) and finally take a decision according to a previously trained classifier. For instance, in [23] and [3] the authors employ Gaussian Mixture Model (GMM) based classifiers for the detection of screams and gun shots. In [24, 15], GMMs were used to model the background sound and perform anomaly detection. A pool of One Class Support Vector Machines (OC-SVM) and novel dissimilarity measures were proposed in [19] and [13].

Methods in the second group are based on more complex architectures: for instance, the decisions GMMs and Support Vector Machines (SVM) are combined in [20]. A two-stage GMM based is adopted in [14], where the first stage separates a sound of interest from the background while the second stage classifies the particular event of interest. In [4] two Learning Vector Quantization (LVQ) classifiers with reject options were trained by feature vectors extracted at different time scales, so as to detect both impulsive and sustained sounds. Architectures for the temporal evaluation of short-time decisions were proposed in [11] and in [2], where a feature augmentation and a classifier based on Genetic Motif Discovery were presented, respectively. Audio phrases composed of sequences of basic audio units are also employed in [18]. In [9], the authors apply object detection techniques to perform the audio event detection task in spectrogram-like images of sounds.

One of the drawbacks of complex architectures is that

978-1-5090-3811-4/16/$31.00 c 2016 IEEE

IEEE AMMDS Workshop @ IEEE AVSS 2016, August 2016

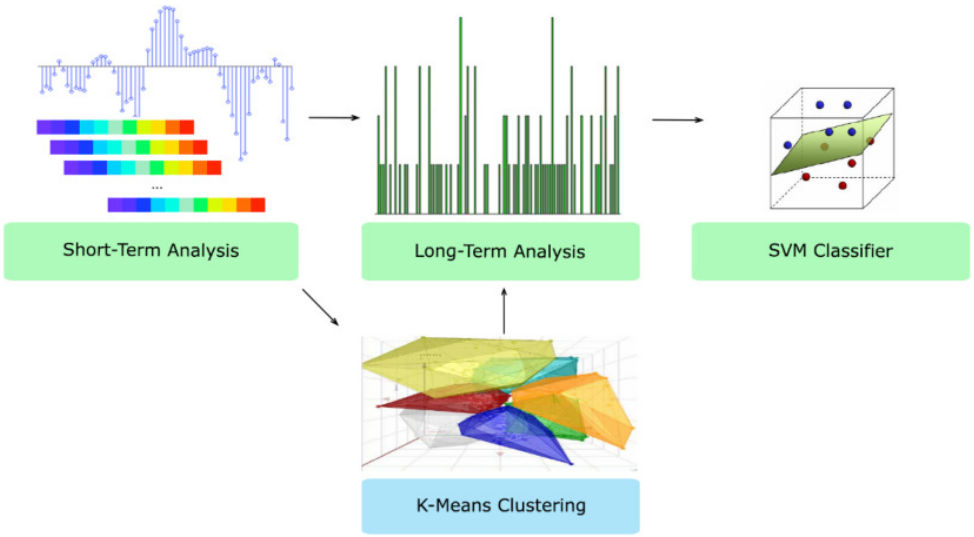

Figure 1: Architectural overview of the proposed system. A short-time analysis for extraction of instantaneous properties of the audio signal is combined with long-time analysis, which aims at increasing the robustness of the system with respect to noise. A multi-class SVM classifier is employed for the classification of the input sound. A K-Means clustering process is used in the training phase to learn the set of basic audio units from training data.

they typically require the definition of a ground truth both at short- and long-time level, thus increasing the human labor time needed to label the data set. Moreover, larger computational resources are typically required to process the sound. Starting from these considerations, an intermediate solution is to create a multi layer representation of the data. For instance the bag of words representation has been demonstrated to be a feasible approach [7, 8].

The method that we propose employs a short-time analysis based on a description of the time-frequency distribution of the energy of the sound. The performance of two different feature extraction methods are compared, namely the Gammatone filterbank and the Undecimated Wavelet Transform. Differently from state of the art methodologies, we evaluated the performance of the proposed approach in two different ways: (i) by using a public dataset, namely the MIVIA dataset [7] and (ii) in real environments, by acquiring the audio stream with a microphone and by processing it in real time. In this way, we studied the effectiveness of the proposed approach in real surveillance scenarios.

The rest of the paper is organized as follows: in Section 2 we provide details on the method, in Section 3 we present the experimental results and provide a discussion. Finally, we draw conclusions in Section 4.

2. The method

The method presented in this paper detects audio events of interest by combining short- and long-term analysis and is able to work with both impulsive and sustained sounds. The basic idea is that a sound is composed of atomic units of sound, whose occurrence in a certain time interval is distinctive for the presence of the event of interest. We hypothesize that some atomic units are discriminant for specific classes of sound and, like words in a text, can be used for the classification of the input sound. The proposed method is based on the bag of words architecture proposed for text classification. In Fig. 1 we show an architectural scheme of the proposed method.

2.1. Short-time analysis

Differently from video signals, audio signals show high variability within few milliseconds. Thus, short-time analysis becomes important in order to take into account and describe such variability. In this work, we perform two types of short-time analysis using the Gammatone filterbank and the Undecimated Wavelet Transform (UWT). They provide different descriptions of the time-frequency distribution of the energy of the sound. The two considered transformations take into account that time and frequency are complementary variables, thus cannot be precisely determined at the same time, and perform a multi-resolution analysis of the input signal.

We divide the input audio stream, sampled at a rate F_s, in groups of N partially overlapped frames, with L PCM samples per frame. Two consecutive frames are overlapped by 75% of their length, which guarantees continuity of the short-time analysis and avoids border effects.

2.1.1 Gammatone filterbank

Gammatone filters are inspired by how the *cochlea membrane* in the human auditory system responds to incoming sound waves. It vibrates during time accordingly to the base frequency and the energy of the input sound. Such vibrations determine a firing activity of the so called inner-hair cells, which are posed behind the cochlea and transfer the electrical stimuli to the auditory nerve. Gammatone filters have been demonstrated to be a linear approximation of the impulse response of the cochlea membrane [16, 17].

In signal processing, a classical way of analyzing the time-frequency distribution of the energy of a signal is through the spectrogram. In the spectrogram, the input signal is processed by a bank of band-pass filters, which all have the same bandwidth. This differs from the way the human auditory system processes the sound. In the Gammatone filterbank the bandwidth of the band-pass filters increases with their central frequency. The resolution in the perception of frequency differences is not constant: the minimum frequency difference that the ear is able to resolve is proportional to the base frequency of the sounds. Thus a given frequency difference is perceived as much stronger at low frequencies than at high frequencies.

Formally, the impulse response of a Gammatone filter is defined as the product of a gamma function with a sinusoidal tone:

$$g(t) = at^{n-1}e^{-2\pi bt}cos(2\pi f_c t + \phi) \tag{1}$$

where f_c is the central frequency of the filter, and ϕ is the phase which is usually equal to 0. The constant a controls the gain, while n is the order of the filter. Finally b is the decay factor, which determines the bandwidth of the filter and the duration of the impulse response. The bandwidth of the band-pass filters increases linearly with their central frequency, following the Equivalent Rectangular Band scale:

$$ERB = 24.7 + 0.108 f_c \tag{2}$$

In this work, we employ a bank of $N_G = 64$ Gammatone filters and use their responses to form a feature vector that describes the properties of each audio frame.

2.1.2 Undecimated Wavelet Transform

The Undecimated Wavelet Transform (UWT) is a time-frequency transformation of the input signal. It is designed to overcome the problems of the discrete wavelet transform (DWT) due to dependence to translations of the input signal. Unlike the DWT, the UWT does not employ down-samplers and up-samplers, but rather involves an up-sampling procedure of the filter coefficients by a factor of $2^{(j-1)}$ in the j-th level of the algorithm [22]. This algorithm is also known as "algorithme à trou", which refers to the insertion of zeros between the coefficients of the filters [12].

The UWT is, thus, a redundant transformation as the output coefficients of each level contains the same number of samples as the input signal. The redundancy and translation-invariance provide a better approximation of the continuous wavelet transform, but with an increased requirement of computational resources with respect to the DWT.

2.2. Dictionary

The space of short-time feature vectors is continuous and theoretically infinite. In order to determine a finite set of basic units of sound, we perform a quantization of the feature space using the K-Means clustering algorithm. The output of the K-Means algorithm is a set of cluster centroids w_i, which are points in the feature space. We consider such points as discrete basic audio units, which can be compared to the words in a text. We call such discrete units *audio words*. The so computed K audio words form the codebook, or dictionary, of the system: $D = (w_1, ..., w_K)$ Each vector w_i in the dictionary is representative of a recurrent atomic audio unit, whose occurrence increases the statistical evidence of being in presence of a given sound.

2.3. Long-time analysis

In the same way a text cannot be classified by a single word, a single audio unit is not representative of an event of interest. Thus, we perform a long-time analysis that considera $N = 375$ frames, which correspond to $S = 3$ seconds of audio. The long-time analysis creates a description of a time interval of S seconds by means of the type and number of audio words that it contains. For each frame, a short-time feature vector v_i is computed. Then, for each feature vector v_i, the dictionary of audio words D is searched for the closest word w_j. Finally a long-time feature vector $H = (h_1, ..., h_K)$ is calculated as the histogram of the occurrence of the audio words within the considered interval. Formally, the j-th bin of the histogram is computed as follows:

$$h_j = \sum_{i=1}^{N} a_{ij}, \quad \forall j = 1, ...K \tag{3}$$

where a_{ij} is an indicator function whose value is 1 if the closest word to v_i is w_j:

$$a_{ij} = \begin{cases} 1, \text{if } i = \arg\min_j d(v_i, w_i), \forall j = 1, ...K \\ 0, \text{otherwise} \end{cases} \tag{4}$$

where $d(v_i, w_j)$ is a distance measure between the i_{th} vector in the time window and the j_{th} word of the dictionary (for uniformity with the distance metric employed in the K-Means algorithm, $d(\cdot, \cdot)$ is the Euclidean distance).

2.4. Classification

The histogram of the occurrence of audio words within a time interval is a descriptor of the sound that is contained in such time interval. We use the feature vectors H to train a multi-class Support Vector Machine (SVM) classifier. SVM is originally a binary classifier. Thus, we employ a pool of SVM classifiers, each of them trained using a 1-vs-all strategy. Let $C_i, i = 0, \ldots, M-1$ be the M classes of interest for the system. The i-th classifier is trained using as positive samples the samples from the class C_i and negative samples the ones from the remaining classes. In the application at hand, we trained a SVM also to recognize the background sound. It increases the robustness to background noise and reduces the detection of false events when only background sounds are present in the time window of analysis. In the operating phase, the output of the i-th classifier is a classification score s_i. The final class C is assigned to the input pattern H through the following combination rule:

$$C = \begin{cases} C_0 & \text{if } s_i < 0,\ \forall i = 1, ...M \\ \arg\max_i s_i & \text{else} \end{cases} \quad (5)$$

The class that corresponds to the SVM with the highest score is assigned to the input pattern. In case all the scores are negative, the considered time window is classified as containing only background sound.

3. Experiments

3.1. Data set

For our experiments we used a large public data set for audio surveillance applications, namely the MIVIA audio event data set [7]. The data set contains glass breaking, gun shot and scream events, which are superimposed to various combinations of background sounds. This aims at simulating the occurrence of such events in various scenarios, with different values of signal to noise ratio (SNR).

The data set is composed of PCM audio clips sampled at $32KHz$ and with a resolution of 16 bits per sample. A total of 6000 events of interest are present in the data set, divided in 4200 events for training and 1800 events for testing. The data set includes 20 hours of audio data for the training set and about 9 hours for the testing set. Given a given sequence of foreground event, six versions of the audio clip are present, so that each event occurs with different values of SNR in the set $\{5dB, 10dB, 15dB, 20dB, 25dB, 30dB\}$. The events in the audio clips are distanced each other by 5 to 8 seconds in which only background sound is present. In the following we refer at glass breaking with GB, gun shots with GS and screams with S. We indicate the background noise with BN. In Table 1 we report details about the composition of the data set.

MIVIA data set

	Training set		**Test set**	
	#Events	**Duration (s)**	**#Events**	**Duration (s)**
BN	-	58371.6	-	25036.8
GB	4200	6024.8	1800	2561.7
GS	4200	1883.6	1800	743.5
S	4200	5488.8	1800	2445.4

Table 1: Summary of the composition of the data set. The total duration of the sounds in the four classes is expressed in seconds.

3.2. Results

In the application phase of the proposed method, we used a time-window of duration 3 seconds that slides on the audio signal by 1 second steps. The audio signal contained in such time windows is represented with the histogram of the occurrences of audio words, which is used as input for the classification module. We count an event of interest as correctly detected if at least one of the time windows that overlap with it is correctly classified.

In surveillance applications, it is important to correctly recognize events of interest and take actions consequently, but it is also relevant to not detect such events when only background sound is present in the environment. In the evaluation process, we take into account such considerations and evaluate the performance of the system by computing different metrics:

- recognition rate (RR) is the rate of correctly detected events
- false positive rate (FPR) is the rate of events that are detected when only background sound is present
- miss detection rate (MDR) is the rate of events of interest that are not detected
- error rate (ER) is the rate of abnormal events that are detected but that are classified as belonging to the wrong class

In Table 2, we report the classification matrices achieved by the proposed system when Gammatone filterbank and UWT short-time features are employed as short-time descriptors. We compare the performance results with the ones achieved by the original system proposed in [7]. In the case of the use of Gammatone features, the proposed system achieved an average recognition rate $RR = 88.6\%$ with a $FPR = 1.4\%$. In the case of UWT features, instead, the system achieved an average recognition rate equal to 77.9% and a $FPR = 1.5\%$.

In this work, we studied the stability of the system when it is employed in a real environment. To this concern, we carried out a number of experiments by using a version of

Gammatone - classification matrix

		Guessed class			
		GB	**GS**	**S**	**Miss**
True class	**GB**	97.6%	0.3%	0.3%	1.8%
	GS	1.2%	88.69%	0.6%	9.3%
	S	1%	0.9%	79.3%	18.8%

UWT - classification matrix

		Guessed class			
		GB	**GS**	**S**	**Miss**
True class	**GB**	90.7%	0.6%	0.2%	8.6%
	GS	0.9%	57.3%	0.7%	4.1%
	S	0.8%	0.8%	85.7%	12.7%

[7] - classification matrix

		Guessed class			
		GB	**GS**	**S**	**Miss**
True class	**GB**	93.6%	0.2%	0.2%	6%
	GS	3.3%	81.6%	0.5%	14.6%
	S	2.8%	0.9%	79.3%	17%

Table 2: Classifications matrices achieved by the proposed method with Gammatone and UWT short-time analysis. The classification results are compared with the ones reported in [7].

the system that acquires the audio stream by a microphone and processes it in real-time. We reproduced the whole test set by loudspeakers and analyzed the performance of the proposed method when different short-time features are used. Tests with loudspeakers allows to evaluate the effectes caused by the noise that is introduced due to propagation of the sound in the test environment. In general, we observe a decrease of the performance in the case of use of the system in a live scenario. In Table 3, we report the results achieved on the live test set and compare them to the ones achieved on the test set in the original MIVIA audio event data set. The short-time features computed by the gammatone filterbank are the ones that showed the strongest stability of performance in live environments and achieved highest results than the feature set proposed in [7].

3.3. Discussion

The performance results achieved by the proposed method when Gammatone filterbank are used as short-time feature extractors demonstrate that it is suitable for use in real applications. The variability of the input audio signal is a central problem in audio analysis, since it can determine strong differences of the sound events with respect to the ones that are used to build the classification models. In this work, we studied the effects that different short-time features have on the performance and stability of an audio

Live test set - Results

Features	**Set**	**RR**	**MDR**	**ER**	**FPR**
Gammatone	Orig.	88.6%	9.9%	1.4%	1.4%
	Live	81.1%	16.6%	2.3%	8.1%
UDWT	Orig.	77.9%	20.8%	1.3%	1.5%
	Live	53.6%	39.8%	6.7%	5.3%
Foggia et al. [7]	Orig.	84.8%	12.5%	2.7%	2.1%
	Live	78.5%	16%	5.5%	10.3%

Table 3: Comparison of performance results on the original test set and on the live test set, for different short-time features.

recognition system based on the bag of words approach. In the proposed system, the characteristics of the short-time features influence also the quality of the learned higher-level representation. We observed that Gammatone filters contributed to effectively catch important and distinctive instantaneous properties of the events of interest. The use of mathematical models and tools that are inspired by neurophysiological evidences of the functions of the human auditory system allows for taking into account those characteristics of the signals that are not usually studied by classical pattern recognition approaches. In our case, the Gammatone filterbank computes an approximated response of the cochlea membrane that is sufficient for our brain to detect any kind of event of interest even in noisy environments. Then, the successive stages of the presented classification architecture are devoted to exploit such basic information and learn a high-level representation for the classification of both impulsive and sustained sounds.

The use of Gammatone filters for the extraction of short-time properties of the input audio signal contributed to a consistent improvement of performance and stability with respect to the system presented in [7]. In the latter work, a set of features was engineered in order to overcome specific problems and challenges of the problem at hand. In this work, instead, we employ a more general representation of the audio signal that has been shown effective also for the processing of music [1] and speech signals [21].

The Undecimated Wavelet Transform (UWT) aims at overcoming the problems caused by temporal shifting of the input signal. Indeed, the algorithm for the computation of the Discrete Wavelet Transform involves a decimation step, which keeps, step by step, lower frequency components for further computation of the wavelet coefficients. It is clear that a shift of the input signal causes strong variations in the computation of lower frequency wavelet coefficients. The UWT aims at reducing the influence of temporal shifts of the input signal by performing an expansion of the signal at each level of analysis. Although the UWT deals with the problem of temporal shifting of the input signal, in our experiments it was not able to extract effective features for the

description of the events of interest. For a frame of $32ms$, which corresponds to 1024 PCM samples, we could use a maximum of 6 wavelet levels. This caused an excessive compression of the audio signal which did not contribute to the learning of an effective dictionary of audio words.

It is worth noting that our implementation of the proposed system achieves real-time responses. When the system is already trained, in the application phase, the algorithm takes less than 1 second to process an audio signal of 3 seconds at a 32 KHz sampling rate and requires about 3% of the time of a single CPU core (2GHz). Moreover, the proposed system also runs in real-time on an embedded STM32F4 board. It makes its deployment very easy also for large installations, due to reduced hardware cost.

4. Conclusions

The approach that we present combines short- and long-time analysis of the audio stream, so as to take into account both instantaneous and long-term properties of the input signal. We studied the performance of the system when several short-time feature extractors are used to describe audio signal. The experiments that we performed on a public benchmark data set, together with the tests that we carried out in a live environment, suggest the applicability of the proposed method in real scenarios. Moreover, our implementation achieves real-time responses on cheap embedded devices (e.g. STM32F4 card), which facilitates large-scale deployment due to reduced cost.

References

[1] P. antoine Manzagol, T. Bertin-mahieux, and D. Eck. On the use of sparse time-relative auditory codes for music. In *ISMIR 2008*, 2008.

[2] M. Chin and J. Burred. Audio event detection based on layered symbolic sequence representations. In *IEEE ICASSP*, pages 1953–1956, 2012.

[3] C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *ICME*, pages 1306 –1309, 2005.

[4] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento. An ensemble of rejecting classifiers for anomaly detection of audio events. In *IEEE AVSS*, pages 76–81, 2012.

[5] M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Trans. Multimedia*, 9(2):257–267, 2007.

[6] M. Crocco, M. Cristani, A. Trucco, and V. Murino. Audio surveillance: A systematic review. *ACM Comput. Surv.*, 48(4):52:1–52:46, Feb. 2016.

[7] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22 – 28, 2015.

[8] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Trans. Intell. Transp. Syst.*, 17(1):279–288, 2016.

[9] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento. Cascade classifiers trained on gammatonegrams for reliably detecting audio events. In *IEEE AVSS*, pages 50–55, 2014.

[10] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento. Exploiting the deep learning paradigm for recognizing human actions. In *IEEE AVSS*, pages 93–98, Aug 2014.

[11] R. Grzeszick, A. Plinge, and G. Fink. Temporal acoustic words for online acoustic event detection. In *Pattern Recognition*, volume 9358 of *LNCS*, pages 142–153. 2015.

[12] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. *A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform*, chapter Wavelets: Time-Frequency Methods and Phase Space. 1990.

[13] 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. In *IEEE AVSS*, pages 124 –129, 2011.

[14] S. Ntalampiras, I. Potamitis, and N. Fakotakis. An adaptive framework for acoustic monitoring of potential hazards. *EURASIP J. Audio Speech Music Process.*, 2009:13:1–13:15, Jan. 2009.

[15] S. Ntalampiras, I. Potamitis, and N. Fakotakis. Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Trans. Multimedia*, 13(4):713 –719, 2011.

[16] R. D. Patterson and B. C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency selectivity in hearing*, pages 123–177, 1986.

[17] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand. Complex Sounds and auditory images. In *Auditory Physiology and Perception*, pages 429–443. 1992.

[18] H. Phan, L. Hertel, M. Maass, R. Mazur, and A. Mertins. Audio phrases for audio event recognition. In *EUSIPCO*, 2015.

[19] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze. Using one-class svms and wavelets for audio surveillance. *IEEE Trans. Inf. Forensics Security*, 3(4):763–775, 2008.

[20] J.-L. Rouas, J. Louradour, and S. Ambellouis. Audio events detection in public transport vehicle. In *IEEE ITSC*, pages 733–738, 2006.

[21] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney. Gammatone features and feature combination for large vocabulary speech recognition. In *IEEE ICASSP*, volume 4, pages IV–649–IV–652, 2007.

[22] M. J. Shensa. The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on Signal Processing*, 40(10):2464–2482, Oct 1992.

[23] M. Vacher, D. Istrate, L. Besacier, J. F. Serignat, and E. Castelli. Sound Detection and Classification for Medical Telesurvey. In *ICBME*, pages 395–398, 2004.

[24] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *IEEE AVSS*, pages 21–26, 2007.