Detection of Alarm Sounds in Noisy Environments

Dean Carmel, Ariel Yeshurun, Yair Moshe
Signal and Image Processing Laboratory (SIPL)
Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion – Israel Institute of Technology
Technion City, Haifa, Israel, http://sipl.technion.ac.il/

***Abstract*—Sirens and alarms play an important role in everyday life since they warn people of hazardous situations, even when these are out of sight. Automatic detection of this class of sounds can help hearing impaired or distracted people, e.g., on the road, and contribute to their independence and safety. In this paper, we present a technique for the detection of alarm sounds in noisy environments. The technique is not limited to particular alarms and can detect most electronically generated alerting sounds within 200 ms. We consider a set of acoustic features and use the ReliefF algorithm to select only the ones that best differentiate between alarms and other sounds. We use an SVM classifier as the detector. On the tested dataset, consisting of several dozen alarm sounds and several dozen background noises, the proposed technique shows an accuracy of 98% per audio frame. With a larger training dataset, this result is expected to substantially improve.**

Keywords- alarm detection; siren detection; acoustic event detection; sound recognition; assistance for hearing impaired

I. INTRODUCTION

The diversity of environmental sounds is vast and includes the sounds generated in indoor and in outdoor environments. These sounds convey information about human and social activities. Examples include cheering of the audience in a sports event, a gunshot in the street and hasty steps in a nursing home. Such information is helpful in applications that analyze audio and video content. Alerting sounds such as emergency vehicles, smoke alarms and medical monitoring alarms are of a special importance, as they are usually designed to warn people of hazardous situations, even when these are out of sight. Unfortunately, alerting sounds may be missed due to hearing impairment or just simple distraction, leading to hazardous and life-threatening situations. Automatic detection of such sounds may have many applications, both for hearing aids, and for intelligent systems that need to respond to their acoustic environments. A strong need exists for hearing impaired people driving a car. In such a situation, background noise is often very loud, causing many people with a hearing loss to drive without amplifications. Therefore, they are unable to hear the siren of an approaching emergency vehicle. To support hearing impaired people, different assistive devices are commercially available, but no supporting devices for environmental sound awareness are available for mobile use [1].

Humans can identify a particular sound as an alarm even when they have never heard it before, and even with a significant background noise. Since by their nature alarm sounds are intended to be easily detected, we could expect alarm sound detection to be a relatively simple task. However, the distinctive characteristics of alarm sounds are not formally defined. Moreover, such sounds are varied, and it is not obvious that they do indeed share common characteristics, rather than being learned by listeners as the conjunction of a set of more special purpose sound types [2]. The International Organization for Standardization has defined a recommendation for "auditory danger signals" (ISO 7731) [3]. However, this recommendation only gives rough guidelines for alarm sounds and is not widely used worldwide. Instead, most countries have their unique siren standardization. Moreover, many alarm sounds are not standardized at all, e.g., alarm clocks. In an attempt to define common characteristics shared by alarm sounds, three types of siren are defined in [4]:

- Pulsed alarms – Consist of a repetition of the same sound with a silent gap between each instance.
- Sirens – Sounds in which the frequency continuously changes.
- Alternating alarms – Consist of two different alternating tones with no silence gap between them.

Spectrograms of examples of these three types of alarm sounds are shown in Fig. 1.

In the last few decades, automatic processing of audio signals drew great interest in both academia and industry. However, most effort has been invested in speech and music processing and significantly less in the processing of environmental sounds, albeit hearing impairment is one of the most widespread physical disabilities worldwide. The majority of works dealing with environmental sounds are concentrated on audio classification, e.g., explosions vs. door slams vs. dog barks, or detection of abnormal audio events. The main research problem these works deal with is choosing a set of suitable audio features. Common features in use for these tasks are the Mel-frequency cepstral coefficients (MFCC), wavelet-based features and individual temporal and frequency features. As the set of sound classes in these works is very limited, and the features selected manually, these works are usually hard to generalize. Thus, to create a reliable alarm sound detector, a more specialized technique should be designed.

With this insight in mind, several previous works have tried to deal specifically with the task of alarm sound detection. Since building a general model of alarm sounds is difficult, most works try to detect only particular alarms, usually sirens of emergency vehicles of a specific country. Many of these works do not perform well out of laboratory conditions since they do not model well enough ambient background sounds and usually do not consider shifts in frequency due to the Doppler effect. For example, in the work presented in [5], the authors try to detect a small set of pre-selected warning sounds in a simulated environment by cross-correlation. In [6], an artificial neural network was used to detect police vehicles in Italy. MFCC

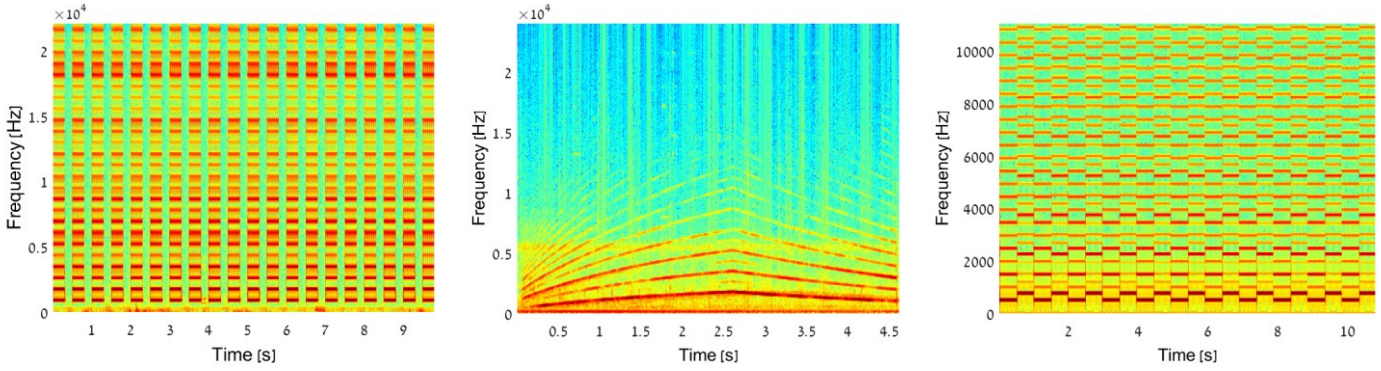

Fig. 1. Spectrograms of three types of alarm sounds. Red represents high values and blue represents low values. (a) Alarm clock: Pulsed alarm - Repetition of the same sound with a silent gap between each instance. (b) Ambulance driving away: Siren - The frequency continuously changes. (c) Fire alarm: Alternating alarm - - Two different alternating tones with no silence gap between them.

coefficients were extracted as features for classification. Another work for detecting sirens of emergency vehicles in Italy is presented in [7]. The system runs in real-time by estimating the pitch frequency and comparing it to pre-defined siren frequencies. Another simple system for siren detection that runs in real-time is described in [8]. Based on the periodicity of alarm sounds, the results of autocorrelation are analyzed by a classifier tailored to acoustic emergency signals in Germany. Sirens of German police cars are also being detected in [9], where part-based models originally proposed in computer vision are trained and used for classification. In [10], a technique for detecting an ambulance siren sound in Taiwan in presented. It uses frequency matching by finding the longest common subsequence, a technique commonly applied for pattern recognition.

One of the first works [2] to attempt to detect alarms in general compared two different approaches – an artificial neural network and a sinusoidal modeling. All tests were done at 0 dB signal-to-noise ratio (SNR), and both approaches were stated to perform poorly. A much simpler approach for automated detection of alarm sounds is described in [11]. Four classes of alarm sounds (bells/chimes, buzzers/beepers, horns/whistles, and sirens) are detected in four noise backgrounds (cafeteria, park, traffic, and music) based on the normalized autocorrelation function and the overall sound level. The authors report true positive rate of 80% with false positive rate above 25%. The autocorrelation function is used for detection of alarms also in [4]. For improved robustness, autocorrelation is performed on the envelope of the signal after band-pass filtering around the fundamental frequencies of the alarms to be detected. The system was tested on alarms compliant to the ISO 7731 recommendation [3] with SNRs ranging from 0 to 15 dB. An average true positive rate of 95% with an average latency of 1.27 sec are reported for pulsed alarms. For other types of alarms, the reported true positive rate is about 50% and for alarms not compliant to the ISO 7731 recommendation, even poorer performance is reported. Recently, a smartphone application for recognizing warning sounds was described in [1]. The application can adapt to sounds defined by the user. Warning sounds occurring in road traffic were chosen as a demonstrator. For classification, the application uses 13 MFCC coefficients and the zero crossing rate (ZCR) in a feed forward neural network. The authors report an average recall of 83% with an average precision of 99%, but it is not clear what background noises were used in the experiments.

In this paper, we present a technique for detecting alarm sounds in general in noisy environments that is able to generalize away from the specific examples used in development to cover most real-world alarm sounds. The technique is primarily intended for a future assistive device in road traffic for hearing impaired persons, but can also be used in other application domains, and was tested accordingly. We use a machine learning approach where an SVM classifier is trained beforehand on features extracted from labeled sound samples and is used for classification. As the classification results heavily depend on the representation power of the audio features, we consider a large set of audio features, both in the time and frequency domain, and use only the ones that best differentiate between alarms and no-alarm sounds. Feature selection is performed using the ReliefF algorithm [12].

II. Alarm Sound Detection

We perform alarm sound detection using a supervised machine learning approach, as shown in Fig. 2. In the offline training phase, audio frames labeled as alarm or no-alarm are used to train an SVM classifier. This trained classifier is used in the online classification phase to classify new unlabeled audio frames as alarms or no-alarms. In both the training and classification phases, audio samples are first band-pass filtered and their time envelope is extracted. Then, features characterizing the signal are extracted and used for training or classification. To make classification results more robust, a median filter of size 3 is applied to the binary SVM classification results. We use audio frames of size 100 ms with a 50% overlap. These short time frames contain enough information for accurate alarm sound detection and they allow us to do so with a short delay of 200 ms. Moreover, using short time frames enables detection of short or intermittent alarm sounds.

Following [4], to remove noise and irrelevant information, pre-processing is performed before feature extraction. First, the signal is band-pass filtered. As alarm sounds tend to be periodic with fundamental frequencies usually in the range of

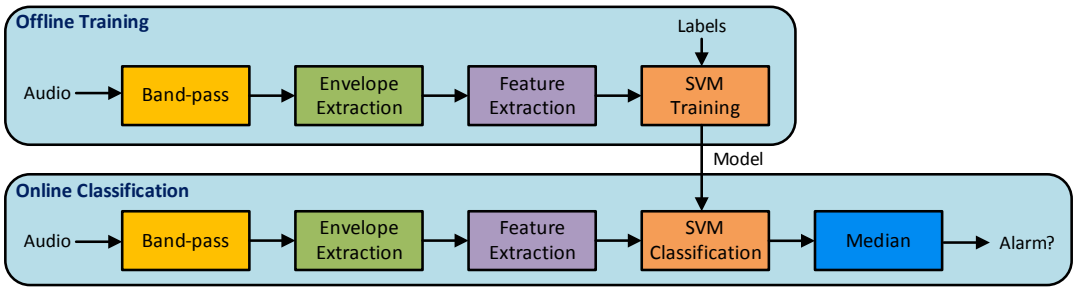

Fig. 2. Proposed alarm sound detection technique. In the offline training phase, audio frames labeled as alarm or no-alarm are used to train an SVM classifier. This trained classifier is used in the online classification phase to classify new unlabeled audio frames as alarms or no-alarms.

500-1500 Hz, a band-pass filter in this frequency range is used. Then, we extract the time envelope of the filtered signal using the magnitude of the analytic signal computed using the Hilbert transform [13]. The extracted envelope is proportional to the amplitude of the signal and captures the slowly varying features of the signal. Therefore, it allows detection of an alarm's energy variation in time.

Careful selection of features is a key issue when designing an effective audio signal detector. For each time frame, we seek features that comprise the essence of information relevant to classification, are robust to variations in the signal and to noise, and are efficient in terms of computational and space complexity. To find such features, we have considered a large set of features that were previously used for different tasks of audio processing, especially the features described in [14]. We will now describe these features briefly and then describe a technique we have used to leave only a subset of these features that contribute to the accuracy of alarm sound detection. We divide the features into three categories:

1) Time domain features

- Pitch - The fundamental frequency computed using the YIN algorithm [15].
- Short time energy – Alarm sounds tend to have high energy in the given spectral band compared to the energy attributed to noise.
- Zero crossing rate (ZCR) - The number of times the sign of a time series changes within a frame. It roughly indicates the frequency that dominates during that frame.

For the short-time energy and the zero crossing rate, we divide each frame into sub-frames of size 20 ms, with a 50% overlap. For each feature, we compute 18 statistics of the sub-frames:

- maximum (max)
- minimum (min)
- mean
- median
- standard deviation (std)
- max/mean
- max/median
- $std^2/mean^2$
- 2nd smallest value
- 3rd smallest value
- 4th smallest value
- 5th smallest value
- mean of 4 smallest values
- mean of values > 10^{-6}
- $\frac{\#values>mean}{\#values}$
- $\frac{\#values>0.1mean}{\#values}$
- $\frac{\#values>5mean}{\#values}$
- $\frac{\#values>0.4}{\#values}$

This results in a vector of size 1+2*18=37.

2) Frequency domain features

- MFCC – We use 13 Mel-frequency cepstral coefficients to describe the spectral shape of the signal. These coefficients are extracted by applying the discrete cosine transform (DCT) to the log-energy outputs of the nonlinear mel-scale filter-bank.
- Spectral flux, spectral roll-off, spectral centroid, and spectral flatness – Spectral flux is a measure of how quickly the power spectrum of a signal is changing, defined as the Euclidean distance between the spectra of two adjacent frames. The spectral flux can be used to determine the timbre of an audio signal, or for onset detection. Spectral roll-off is defined as the Nth percentile of the power spectral distribution, where N is usually 85% or 95%. The roll-off point is the frequency below which the N% of the magnitude distribution is concentrated. This measure is useful in distinguishing voiced speech from unvoiced. Spectral centroid is a measure indicating where the "center of mass" of the spectrum is. Perceptually, it has a robust connection with the impression of "brightness" of a sound. Spectral flatness measures the variance of the frequency spectrum of the signal. High spectral flatness indicates similar amount of energy at all frequencies, e.g., white noise. Low spectral flatness indicates that the signal energy is concentrated mainly around a few frequency bands, e.g., alarm sound. As with the short-time energy and the zero crossing rate, these features are computed by dividing each frame into sub-frames and calculating 18 statistics of these sub-frame features. This results in a feature vector of size 4*18=72.

3) Wavelet-based features

The wavelet coefficients capture time and frequency localized information about the audio waveform. The wavelet transform can be viewed as a multi-level process of filtering the signal using a low-pass (scaling) filter and a high-pass (wavelet) filter. The first layer of the transform decomposition of a signal splits it into two bands giving low-pass approximation coefficients and high-pass detail coefficients.

- Discrete Wavelet transform (DWT) – In the DWT, each level is calculated by passing only the previous wavelet approximation coefficients through discrete-time low and high pass filters. Thus, DWT results in a binary tree like structure which is left recursive (where the left child represents the lower frequency band). We decompose each frame using the DWT with 10 decomposition levels. For the

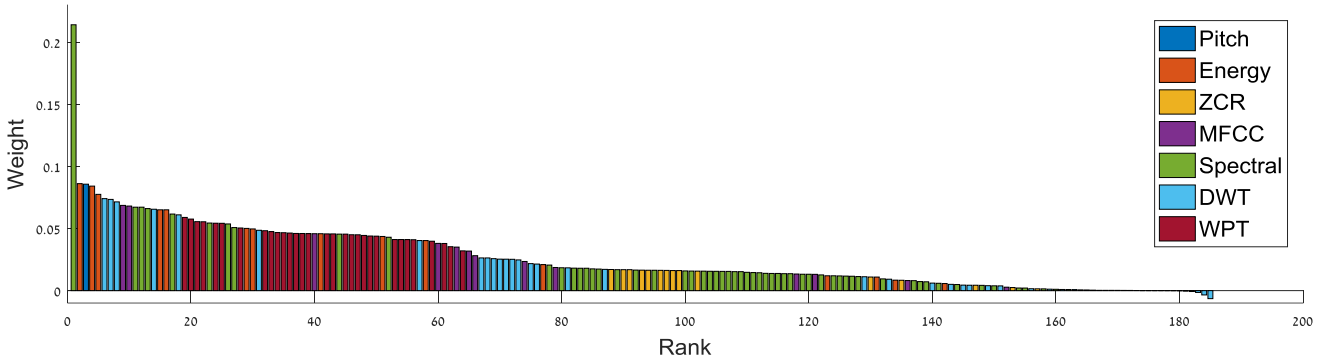

Fig. 3. Feature ranks and weights computed by the ReliefF algorithm. All feature types contribute to classification accuracy. The highest weights are assigned to features based on the spectral flux (leftmost green bars), features based on the short time energy and the pitch. Wavelet-based features also have a significant contribution to classification accuracy. Weight ≤ 0 indicates a feature that does not contribute to a higher classification accuracy.

resulting coefficients, we compute 4 features - energy, variance, standard deviation, and waveform length [16]. This results in a vector of size 15*4=60.

- Wavelet packet transform (WPT) - In the WPT, both the detail and approximation coefficients are decomposed to form a full binary tree. The WPT results in equal-width sub-band filtering of signals as opposed to the coarser octave band filtering found in the DWT. This makes wavelet packets an attractive alternative to the DWT in certain applications such as audio classification. We decompose each frame using the WPT with 4 decomposition levels. This results in a vector of size 31.

Feature Selection

The set of features we have described for representing the audio data consists of many features but only some of them may be useful for alarm sound detection. Feature selection is the process of selecting a subset of the features by removing redundant and irrelevant ones. This process reduces the dimensionality of the dataset, which in turn reduces the computational and storage complexity of the classification. We select a subset of the features by using the ReliefF algorithm [12]. ReliefF receives as input a dataset with n instances of p features, belonging to two known classes. The key idea of ReliefF is to estimate features importance according to how well their values distinguish among neighboring instances. An instance X is denoted by a p-dimensional vector $(x_1, x_2, \dots, x_p)$ where x_i denotes the value of feature f_i of X. Given x_i, ReliefF searches for its nearest two neighbors, one from the same class (called the nearest hit) and the other from a different class (called the nearest miss). The weight vector W_i of the instance x_i is then updated by:

$$W_i = W_i - |x_i - nearHit_i| + |x_i - nearMiss_i| \qquad (1)$$

Thus the weight of any feature decreases if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, and increases in the reverse case. After n iterations, the algorithm averages the contribution of the nearest hits and misses to the weight vector to obtain the relevance vector. Features are selected if their relevance is greater than a threshold value. In practice, to increase the algorithm robustness to noise, ReliefF does not search only for nearest two neighbors but for k nearest hits and misses and averages their contribution to the weights of each feature. We use k=35. The feature ranks and weights computed by ReliefF on our training dataset are depicted in Fig. 3. We can see in the figure that all feature types contribute to classification accuracy. The highest weights are assigned to features based on the spectral flux (leftmost green bars), to features based on the short time energy, and to the pitch. Wavelet-based features also have a significant contribution to classification accuracy. We have decided to put a threshold of 0 on the weight, removing about 12% of the features that do not contribute to a higher classification accuracy.

III. Results

To simulate real-life conditions, we assembled a dataset of diverse alarm sounds and ambient noises. The dataset contains 70 audio signals 30 seconds long, 35 alarm sounds and 35 noises. Sounds were collected by searching the web and by making recordings around the home and office. Of the 35 alarm sounds, 20 are clean and 15 are real-life recordings with significant ambient noise. Of the 35 noises, some are typical everyday noises and some were carefully selected according to their similarity to alarm sounds, e.g., busy roads noises, a truck motor, and a helicopter.

To evaluate the performance of the proposed technique, we used k-fold cross-validation with $k = 10$. We obtained 98% accuracy per audio frame (for both alarms and noises). In a real-life application that constantly monitors the environment, the number of no-alarm audio frames is orders of magnitude higher than the number of alarm audio frames. To make such an application feasible, the false positive rate (percentage of no-alarm frames miss classified as alarm frames) should be negligible. We have found the false positive rate of the proposed technique to be 0.4%. This result is encouraging compared with previous works but is not small enough to make an audio monitoring application feasible. Further analysis revealed that false positives resulted from two noises in the dataset – a loud sound of birds chirping and the sound of a waterfall. The miss classification of frames of those two noises is due to the fact that those noises differ greatly from other sounds in the dataset. Therefore, in the k-fold cross-validation procedure, when one of those sounds is used for classification, no similar sounds are available for training and it is difficult for

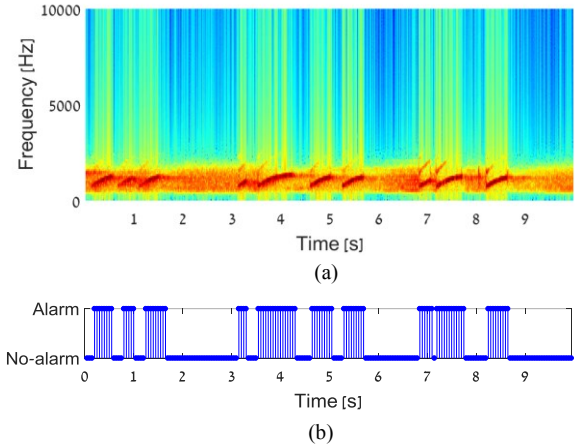

Fig. 4. (a) A spectrogram of a police car operating its siren in intermittent bursts and (b) alarm/no-alarm classification results of its audio frames. Classification results are accurate (accuracy of 100%) although the siren operates in short bursts and is interleaved with noise.

the classifier to generalize to this sound. This problem can be solved by an expanding the dataset with more diverse ambient noises. Note that the false positive rate can be further reduced by allowing a longer delay and incorporating the results from more consecutive frames, e.g., using a median filter of a larger size.

We have also performed several tests with real-life unlabeled data to verify the robustness of our alarm detection technique to noise, low amplitude, Doppler effect (e.g., moving emergency vehicles) and short alarms with intermittent bursts. For all those tests, satisfactory results were obtained. For example, Fig. 4 demonstrates the ability of the proposed technique to detect short alarms. It shows a spectrogram of a police car operating its siren intermittently and alarm/no-alarm classification results of its audio frames. Although the siren operates in short bursts and is interleaved with noise, the proposed alarm detection technique is able to detect the siren accurately.

IV. Conclusions

In this paper, we present a detection technique for alarm sounds in noisy environments. The technique is not restricted to the recognition of a specific set of pre-defined alarm sounds and can generalize to most electronically generated alerting sounds. Acoustic features, which were verified to contribute to this task, are used with an SVM classifier. Alarm sounds can be detected with a short delay of 200 ms. On a dataset of several dozen audio samples, we achieve an accuracy of 98% per 100 ms audio frame. The false positive rate of the proposed technique is 0.4% and can be further improved by future expansion of the dataset. The proposed technique is robust to noise and to other real-life conditions such as low alarm amplitude, Doppler effect and short alarms with intermittent bursts.

Acknowledgment

The authors would like to thank Prof. David Malah, head of SIPL, for his advice and helpful comments.

References

[1] M. Mielke and R. Brueck, "Design and evaluation of a smartphone application for non-speech sound awareness for people with hearing loss," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015, pp. 5008-5011.

[2] D. Ellis, "Detecting alarm sounds," in *Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-2000*, 2001.

[3] "ISO 7731: Ergonomics - Danger signals for public and work areas - Auditory danger signals," International Organization for Standardization, 2013.

[4] M.-A. Carbonneau, N. Lezzoum, J. Voix, and G. Gagnon, "Detection of alarms and warning signals on an digital in-ear device," *International Journal of Industrial Ergonomics,* vol. 43, pp. 503-511, 2013.

[5] E. R. Bernstein, A. J. Brammer, and G. Yu, "Augmented warning sound detection for hearing protectors," *The Journal of the Acoustical Society of America,* vol. 135, pp. EL29-EL34, 2014.

[6] F. Beritelli, S. Casale, A. Russo, and S. Serrano, "An automatic emergency signal recognition system for the hearing impaired," in *Digital Signal Processing Workshop, 12th-Signal Processing Education Workshop, 4th*, 2006, pp. 179-182.

[7] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci, and P. Desii, "A real-time siren detector to improve safety of guide in traffic environment," in *Signal Processing Conference, 2008 16th European*, 2008, pp. 1-5.

[8] M. Mielke, A. Schäfer, and R. Brück, "Integrated circuit for detection of acoustic emergency signals in road traffic," in *Mixed Design of Integrated Circuits and Systems (MIXDES), 2010 Proceedings of the 17th International Conference*, 2010, pp. 562-565.

[9] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models," in *ICASSP*, 2013, pp. 493-497.

[10] J.-J. Liaw, W.-S. Wang, H.-C. Chu, M.-S. Huang, and C.-P. Lu, "Recognition of the ambulance siren sound in Taiwan by the Longest Common Subsequence," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 2013, pp. 3825-3828.

[11] R. A. Lutfi and I. Heo, "Automated detection of alarm sounds," *The Journal of the Acoustical Society of America,* vol. 132, pp. EL125-EL128, 2012.

[12] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence,* vol. 7, pp. 39-55, 1997.

[13] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal - Part I: Fundamentals," *Proceedings of the IEEE,* vol. 80, pp. 520-538, 1992.

[14] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Transactions on information forensics and security,* vol. 3, pp. 763-775, 2008.

[15] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America,* vol. 111, pp. 1917-1930, 2002.

[16] R. Khushaba, "Feature Extraction Using Multisignal Wavelet Transform Decomposition, https://www.mathworks.com/matlabcentral/fileexchange/37950-feature-extraction-using-multisignal-wavelet-transform-decomposition," 2012.