



Audio Surveillance: Detection of Audio-Based Emergency Situations

Zhandos Dosbayev¹, Rustam Abdrakhmanov^{2(✉)}, Oxana Akhmetova³,
Marat Nurtas^{4,5}, Zhalgasbek Iztayev⁶, Lyazzat Zhaidakbaeva⁶,
and Lazzat Shaimerdenova³

¹ Satbayev Kazakh National Technical University, Almaty, Kazakhstan

² Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
rustam.abdrakhmanov@ayu.edu.kz

³ Al-Farabi Kazakh National University, Almaty, Kazakhstan

⁴ International Information Technology University, Almaty, Kazakhstan

⁵ Kazakh-British Technical University, Almaty, Kazakhstan

⁶ M.Auezov South Kazakhstan University, Shymkent, Kazakhstan

Abstract. The subject of the study was the recognition of sounds of critical situations in the audio signal. The term “critical situation” is understood as an event, the characteristic sound signs of which can speak about acoustic artifacts as a shot, a scream, a glass crash, an explosion, a siren, etc.. The paper considers the scope of audio analytics, its advantages, the history of spectral analysis, as well as analyzes and selects tools for further development of system components. In the paper, we propose our dataset that consists of 14 classes that contains 1000 sounds of each, and a model to detect emergency situations using audio processing and analytics.

Keywords: Acoustic signals · Event detection · Audioanalysis · Audio surveillance

1 Introduction

In recent years, we have increasingly encountered various audio content that is distributed for both commercial and non-commercial purposes. Due to the growing availability of audio materials and the growth of computing power, automated signal-based audio processing is currently at the center of various studies [1].

Depending on the storage format, user requirements, data volume, and many other parameters, a variety of applications and trends have emerged to solve various audio analysis tasks. The following popular tasks of audio analysis can be distinguished [2]: speech recognition, speaker identification, music information search (MIR), event detection, emotion recognition, and film content analysis.

Audio information can be presented in different ways and in different formats. For example, a composer can record a work in the form of a musical score (sheet music). A note has several properties, including pitch, timbre, volume, and duration [3].

MusicXML has become a universal format for storing music files for use in various music notation applications. This is a music file format, the main task of which is to correctly display the graphics, that is, to demonstrate how a piece of music will look. For electronic instruments and computers, music can be transmitted using standard protocols, such as the widely used Digital Musical Instrument Interface (MIDI) protocol, where event messages determine pitch, speed, and other parameters to generate the intended sounds [4]. Unlike symbolic representations, audio representations, such as WAV or MP3 files, do not explicitly define musical events [5]. These files encode the acoustic waves that are generated when a source (such as an instrument) makes a sound.

Another commonly used form of representation of an audio signal is the representation of a signal in the time domain. The change in the characteristics of the audio signal over time is represented as a graph. Computers can receive audio signal characteristics at specific points in time. The speed at which the computer analyzes the audio data is called the sampling rate [6].

2 Related Works

Recently, automatic systems that control a person's daily activities are becoming more common. Their main purpose is to ensure public safety, which is achieved through surveillance in public places and the recognition of potentially dangerous situations. Research in the field of automatic surveillance systems is mainly focused on the detection of events using video analytics. In turn, acoustic monitoring can be used as an additional source of information, and, being integrated with video surveillance systems, can increase the efficiency of event detection. This makes it necessary to study the problem of automated recognition of sounds of critical situations in order to further develop a system that searches for them in the audio signal in real time [7, 8].

Progress in the study of acoustic characteristics of sounds is associated with the name of the German scientist Hermann von Helmholtz. He developed the theory of resonance, on the basis of which, in the middle of the XIX century, a resonator was invented, called the Helmholtz resonator. The resonator repeatedly amplifies the amplitude of the spectral components of periodic and aperiodic signals, the frequency of which is close to its natural frequency. With a set of resonators with different natural frequencies, the researcher can perform spectral analysis of audio signals. Initially, this was done as follows: in a resonator on the opposite side of the neck, a process was created, which the researcher inserted into the ear; by listening to the sound under study using a set of such resonators, the scientist could determine which tones and with what volume are present in this sound [9–11].

The next step in the development of the technique of spectral analysis was made a few years later by Rudolf Koenig. Using a set of tunable Helmholtz resonators, he was able to provide visualization of spectral analysis using the manometric capsule he invented in 1862. The principle of operation of the capsule was as follows: in one half of the capsule, separated by an elastic membrane, the lamp gas was supplied, in the other half sound was supplied, and, thus, fluctuations in sound pressure modulated the height of the flame in the capsule: the greater the amplitude of the vibration, the higher the flame [12].

A milestone invention in the field of spectral analysis and visualization of sounds was made by American scientists. They created a new type of spectrograph called a sonograph. It made it possible to visualize a dynamic spectrogram obtained by burning an electrosensitive paper with a pen. In fact, the sonograph has completed a century of analog spectral analysis techniques [13].

The acoustic monitoring system allows you to solve the following tasks in automatic mode and in real time [14–16]:

1. selection of acoustic artifacts in the sound stream (characteristic sound signs of a particular event);
2. perform classification of acoustic artifacts (shot, scream, glass fight, explosion, siren, etc.);
3. selection of speech and its emotional component in the audio stream (for the Russian language) with automatic recognition of keywords and phrases (“police!”, “call an ambulance!”, etc.);
4. determination of the approximate direction to the source of the acoustic artifact relative to the terminal device (if the terminal device is equipped with a stereo microphone);
5. determination of the coordinates of the alarm event (if the terminal device is equipped with a GPS/GLONASS signal receiver);
6. transmitting information about the recorded alarm event to the processing center with the indication of the event attributes (device ID, event time, event class, audio recording of the event, relative direction to the sound source, etc.);
7. saving information about disturbing events in the archive;
8. notification of external video surveillance systems about the registration of an alarm event.

A well-known acoustic monitoring system is the Shot Spotter system, which has been used in the United States since 2006, and although it is limited to only one class of disturbing events – shots, it is not able to recognize speech, it has proven its effectiveness. Over the years, the system has localized 39,000 firearm shots, and police have been able to respond quickly on a case-by-case basis.

3 Materials and Methods

3.1 Extracting Features from an Audio Signal

Feature extraction is an important step in both audio analysis and image recognition and machine learning in general. The goal is to extract a set of characteristics that informatively reflect the properties of the source data from the data set of interest. This allows you to reduce the dimension of the data [17].

To achieve this goal, it is important to have a good understanding of the subject area in order to decide which of the features are important and which are not. After extracting the necessary signal features for their further use, the features are normalized. In this case, well-known and theoretically studied methods of reducing the dimension of the feature vector (LDA, PCA, etc.) are used [18].

3.2 Short-Term Audio Signal Analysis

In most applications, the audio signal is analyzed using the so-called short-term (or short-term) processing technology, according to which the audio signal is divided into windows (frames) and the analysis is performed based on these frames [19] (Fig. 1).

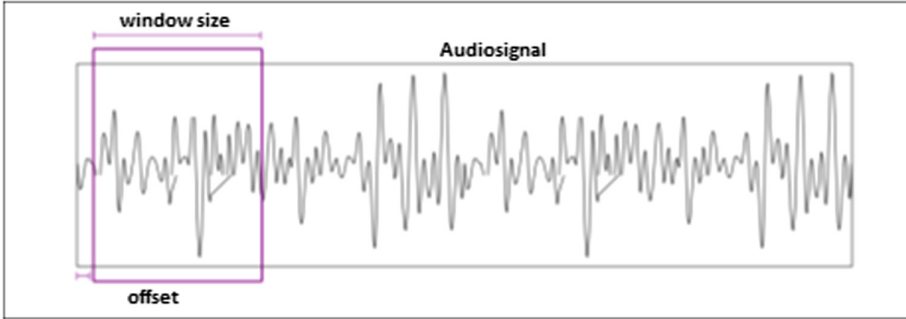


Fig. 1. Example of short-term processing using a window

During short-term processing, we focus each time on a small part (frame) of the signal, i.e., at each stage of processing, we multiply the audio signal by the shifted window function of the finite duration $w(n)$ [20]. The resulting signal $x_i(n)$ at the i th stage of processing is given by the formula 1.

$$x_i(n) = x(n) * w(n - m_i), \quad i = 0, \dots, K - 1 \tag{1}$$

where K – number of frames, m_i – shift delay (the number of samples by which the window is shifted to get the i -th frame).

Formula 1 implies that $w(n)$ is zero everywhere except in the region of samples with indexes $m_i, \dots, m_i + WL - 1$, where WL - length of the moving window. The value of m_i depends on the step WS of the window. Usually WL varies from 10 ms to 50 ms. On the other hand, the window pitch (S) controls the degree of overlap between consecutive frames. If, for example, 75% overlap is required and the window length is 40 ms, then the window pitch should be 10 ms. Hence, the total number of short-term windows K can be obtained using the formula 2.

$$K = \begin{cases} \frac{N - WL}{WS} + 1 \\ 0, \text{ else} \end{cases} \tag{2}$$

As for the window types, you can use a rectangular window, in which the signal is simply truncated outside the window and remains unchanged inside the window. This logic can be described by formula 3.

$$w(n) = \begin{cases} 1, & 0 \leq n \leq WL - 1 \\ 0, & \text{else} \end{cases} \tag{3}$$

In addition to rectangular windows, you can use such windows as the Hamming window, the Bartlett window, and others.

3.3 System Architecture

The microphone’s audio is constantly recorded by the device. A time limit has been set for the recording. When the user subsystem hits this time limit, it saves the freshly documented file for review, deletes the previously stored file, and starts writing the next file.

Figure 2 illustrates the proposed device architecture. Preprocessor, first stage frame classifier, and second stage frame classifier are among the components of the device.

From the initial audio signal, the preprocessor produces a series of overlapped audio frames and removes a collection of features from each frame. Each function vector is given a collection of labels by the first stage frame classifiers (frame). It’s worth remembering that each frame may have several labels allocated to it. It finally group frames into intervals. Second Stage Interval Classifier conducts an interval-level classification, assigning a final prediction to each interval based on a Weighted Majority Voting (WMV) strategy within the frames that make up the interval.

The process of identifying different troubling audio incidents is broken down into two parts [21]:

- the identification (selection) of sharp pulse signals in the audio data stream from background noise;
- the assignment (recognition) of the identified signal as one of the categories of audio events.

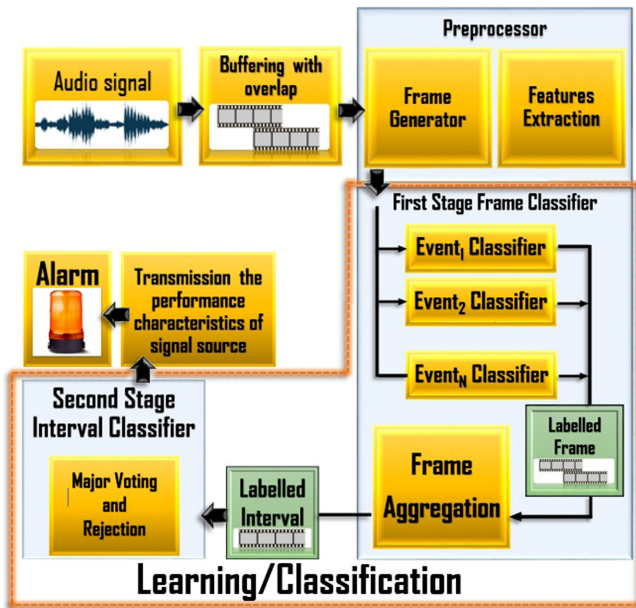


Fig. 2. System architecture

Figure 3 depicts an acoustic measurement device that detects pulsed noises and activities automatically and in real time. The machine receives a signal as well as the sound event’s characteristics. Machine learning is used to classify audio patterns and identify pulsed tones. “Police,” “Ambulance,” and “Blast” are described as keywords in the incident. Identifying the site of possibly hazardous incidents. The next move is to submit reports regarding the warning case to the relevant authority, archive the event, and apply the evidence to the database.

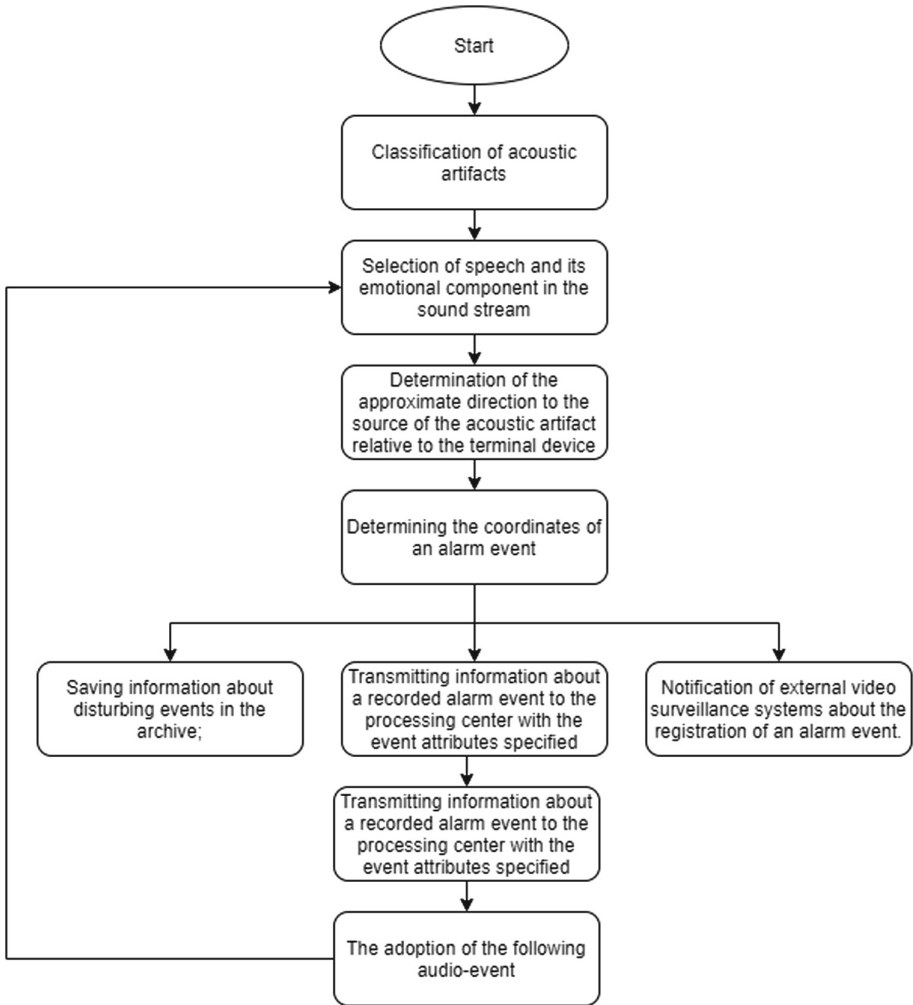


Fig. 3. Audio event detection and classification flowchart

3.4 Data Collection

The proposed system's output was assessed for an automatic monitoring program that needed to detect the following incidents (considered "abnormal" in the observed environment): shots, cries, and broken windows.

We created a data collection for this reason by combining multiple audio samples collected in different railway station scenarios.

Background noise signals such as select, shot, and broken glass make up the data collection. To accommodate for the characteristics of different device scenarios, background noise was recorded both indoors and outdoors.

The signals were divided into one-second intervals (the average time period of each occurrence of interest) for our studies, and then each interval was divided into frames of 200 MS, overlapping by 50%: each interval consists of nine frames.

Table 1 summarizes the data set's signal, frame, and interval structure in terms of signals, frames, and intervals.

4 Evaluation and Experiment Results

We selected test F as a measure of reliability for each Classifier because it is a reasonable balance between precision and recall:

$$precision = \frac{tp}{tp + fp} \quad (4)$$

True positive classified samples are referred to as tp, and false negative classified samples are referred to as fp.

$$Recall = \frac{tp}{tp + fn} \quad (5)$$






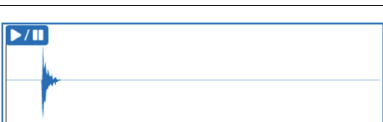
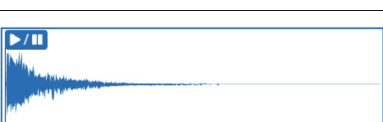
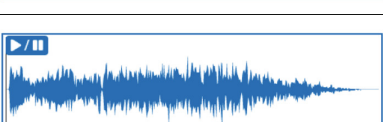

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

Table 2 illustrates experiment results of audioevent classification.

5 Discussion

One of the most critical factors for the proper running of every community is the war against violence. While video surveillance is important in this field, it only provides a visual aspect. Environmental noises that may raise situational sensitivity should be used in a more comprehensive approach.

Table 1. Samples of different impulsive sounds

Table Column Head		
<i>Types of impulsive sounds</i>	<i>Time (sec)</i>	<i>Image</i>
Automobile glass shattering	3.84	
Dog barking	22.15	
Police siren	24.19	
Ambulance siren	15.41	
Constant Wail from Police Siren	56.87	
Single gun shot	3.84	
Explosion	7.78	
Artillery shell explosion	4	
Baby crying	6.66	

(continued)

Table 1. (continued)






Table Column Head		
<i>Types of impulsive sounds</i>	<i>Time (sec)</i>	<i>Image</i>
Burglar alarm	11.13	
Fire alarm beeping	1.41	
Fire alarm bell	1.59	
Smoke alarm	0.99	
Fire alarm yelp	2.3	

Table 2. Results of impulsive audio event type detection

Event type	Accuracy	Precision	Recall	F1 score
Gunshot	0.9178	0.9245	0.9427	0.8945
Broken glass	0.9372	0.9765	0.9215	0.9154
Fire	0.9435	0.9346	0.9215	0.9345
Siren	0.9537	0.9462	0.9876	0.9642
Explosion	0.8132	0.8254	0.8352	0.8124
Cry	0.8635	0.8524	0.8864	0.8754
Dog barking	0.8456	0.8325	0.8571	0.8254
Fire alarm bell	0.8654	0.8452	0.8576	0.8457

Both stored archives and web streams may be processed using audio analytics. Microphones are often cheaper than cameras and do not need specific requirements for positioning and repair, so they are often used as an alternative to video monitoring [21]. The system detects noises in total darkness, and microphones are much cheaper than cameras and do not need special criteria for placement and maintenance. Sound recognition systems can be used to recognize specific noises in an audio stream (screams, explosions, footsteps, sounds of breaking glass, crying), simple audio recordings of noise, identify individuals by their gestures, enhance the accuracy of the speaker's speech, and identify flaws in the function of structures [22].

To combat violence, several communities depend on video surveillance services. However, according to the article, video monitoring alone is not an adequate solution for identifying and stopping crimes [23].

“Today, the most critical components of urban defense applications are hazard identification and data processing tools, such as motion sensors, thermal imaging devices, and license plate recognition apps. They must, though, concentrate solely on visual influences. Sound sensor technology should be used in a fully robust urban defense approach, according to experts [24].

Operators may hear whether an individual is in danger, send them orders, or scare suspects away by alerting them over a loudspeaker using a surveillance solution of audio transmission.

“According to current studies, physical violence is accompanied by verbal aggression in 90% of situations [24]. The violence sound detection device is useful because it helps security staff to sense agitation in voices and other noises consistent with rage, anxiety, and verbal aggression “According to the paper. Safety and law enforcement officers would be able to use audio monitoring to decide which noises are of concern and which are not. The program that senses violence uses sophisticated algorithms to interpret the sounds and adapt them to trends. If the sound is marked as noteworthy, the app sends a warning to the monitoring team right away.

The sounds of violence (for example, physical abuse) and weapons are the two types of sounds that need to be studied the most in order to maintain city protection. Law enforcement authorities may be able to properly cope with violence by utilizing devices that identify violent noises and the use of weapons.

6 Conclusion and Future Work

The study proved the feasibility and potential of a method for automatically detecting impulsive sounds in audio files that combines the usage of amplitude-time and spectral signal parameters. Additional experiments would focus on a more detailed collection and mathematical study of low-level signal characteristics, as well as exploring the possibility of utilizing deep machine learning models to identify impulsive sounds.

References

1. Tharwat, A., Mahdi, H., Elhoseny, M., Hassanien, A.E.: Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm. *Exp. Syst. Appl.* **107**, 32–44 (2018)

2. Vanus, J., et al.: Monitoring of the daily living activities in smart home care. *Hum. Centr. Comput. Inf. Sci.* **7**(1), 30 (2017)
3. Bux, A., Angelov, P., Habib, Z.: Vision based human activity recognition: a review. In: Angelov, P., Gegov, A., Jayne, C., Shen, Q. (eds.) *Advances in Computational Intelligence Systems*. AISC, vol. 513, pp. 341–371. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-46562-3_23
4. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.M.: Computer vision for assistive technologies. *Comput. Vis. Image Understand.* **154**, 1–15 (2017)
5. Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., Baik, S.W.: Efficient deep CNN-based fire detection and localization in video surveillance applications. *IEEE Trans. Syst. Man Cybernet. Syst.* **49**(7), 1419–1434 (2018)
6. Goldenberg, A., et al.: Use of ShotSpotter detection technology decreases prehospital time for patients sustaining gunshot wounds. *J. Trauma Acute Care Surg.* **87**(6), 1253–1259 (2019)
7. Weiss, A., Halevi, O., Manus, H., Springer, D.: U.S. Patent No. 10,021,457. U.S. Patent and Trademark Office, Washington, DC (2018)
8. <http://www.audioanalytic.com/>
9. Virtanen, T., Plumbley, M.D., Ellis, D. (eds.): *Computational analysis of sound scenes and events*, pp. 3–12. Springer, Berlin (2018)
10. Gabriel, D., Kojima, R., Hoshiba, K., Itoyama, K., Nishida, K., Nakadai, K.: 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system. *Adv. Robot.* **33**(7–8), 403–414 (2019)
11. Morehead, A., Ogden, L., Magee, G., Hosler, R., White, B., Mohler, G.: Low cost gunshot detection using deep learning on the raspberry pi. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 3038–3044. IEEE (2019)
12. Alsina-Pagès, R.M., Navarro, J., Alfás, F., Hervás, M.: homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. *Sensors* **17**(4), 854 (2017)
13. Wang, K., Yang, L., Yang, B.: Audio event detection and classification using extended R-FCN approach. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 128–132 (2017)
14. Choi, I., Bae, S.H., Kim, N.S.: Deep convolutional neural network with structured prediction for weakly supervised audio event detection. *Appl. Sci.* **9**(11), 2302 (2019)
15. Romanov, S.A., Kharkovchuk, N.A., Sinelnikov, M.R., Abrash, M.R., Filinkov, V.: Development of a non-speech audio event detection system. In: 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 1421–1423. IEEE (2020)
16. Bello, J.P., Mydlarz, C., Salamon, J.: Sound analysis in smart cities. In: Virtanen, T., Plumbley, M.D., Ellis, D. (eds.) *Computational Analysis of Sound Scenes and Events*, pp. 373–397. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-63450-0_13
17. Tseng, S.Y., Li, J., Wang, Y., Szurley, J., Metze, F., Das, S.: Multiple instance deep learning for weakly supervised small-footprint audio event detection (2017). <https://arxiv.org/abs/1712.09673>
18. Cao, Y., Iqbal, T., Kong, Q., Galindo, M., Wang, W., Plumbley, M.: Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. DCASE2019 Challenge, Tech. Rep. (2019)
19. Cerutti, G., Prasad, R., Brutti, A., Farella, E.: Neural network distillation on IoT platforms for sound event detection. *Proc. Interspeech* **2019**, 3609–3613 (2019)
20. Zinemanas, P., Cancela, P., Rocamora, M.: MAVD: A Dataset for Sound Event Detection in Urban Environments (2019)

21. Wu, D.: An audio classification approach based on machine learning. In: 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 626–629. IEEE (2019)
22. Alfas, F., Alsina-Pagès, R.M.: Review of wireless acoustic sensor networks for environmental noise monitoring in smart cities. *J. Sens.* **2019**, 1–13 (2019)
23. McFee, B., Salamon, J., Bello, J.P.: Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(11), 2180–2193 (2018)
24. Sammarco, M., Detyniecki, M.: Car accident detection and reconstruction through sound analysis with Crashzam. In: Donnellan, B., Klein, C., Helfert, M., Gusikhin, O. (eds.) SMART-GREENS/VEHITS -2018. CCIS, vol. 992, pp. 159–180. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26633-2_8