Computationally Efficient Classification of Audio Events Using Binary Masked Cochleagrams

Tomasz Maka[0000−0001−5898−2201]

West Pomeranian University of Technology, Szczecin
Zolnierska 49, 71-210 Szczecin, Poland
tmaka@wi.zut.edu.pl

Abstract. In this work, a computationally efficient technique for acoustic events classification is presented. The approach is based on cochleagram structure by identification of dominant time-frequency units. The input signal is splitting into frames, then cochleagram is calculated and masked by the set of masks to determine the most probable audio class. The mask for the given class is calculated using a training set of time aligned events by selecting dominant energy parts in the time–frequency plane. The process of binary mask estimation exploits the thresholding of consecutive cochleagrams, computing the sum, and then final thresholding is applied to the result giving the representation for a particular class. All available masks for all classes are checked in sequence to determine the highest probability of the considered audio event. The proposed technique was verified on a small database of acoustic events specific to the surveillance systems. The results show that such an approach can be used in systems with limited computational resources giving satisfying classification results.

Keywords: audio events · cochleagram mask · time–frequency units.

1 Introduction

As a part of auditory scene analysis, acoustic events detection plays an essential role in the machine listening systems. The number of events and its occurrence in time creates a specific structure of the acoustic environment. Acoustic events detection is a well-studied problem. Recently, due to the popularity of deep learning paradigm, many more robust solutions have been proposed [8], [12], [17]. However, these systems require a lot of data to create robust models. The requirement of memory and computational resources, in this case, can be significant and cause difficulties in using them into low–power systems. Moreover, such solutions are sensitive to the varying real acoustic conditions which cause performance deterioration. The event detection system in many practical applications has to be run continuously to detect the specific events and perform actions according to the type of event. Additionally, such a system is often organised as a set of separate and cooperated modules powered by a battery which become more and more popular in IoT (Internet of Things) systems [13], [2]. Therefore,

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_56

the overall computational cost needs to be minimised primarily if the system is dedicated to acoustic surveillance. Moreover, such distributed system gathering the information from many modules placed in various locations may require a mechanism to fuse information of detected events [16] which can be used to track moving sound sources in the monitored area. Using many IoT devices, it is possible to balance the calculations and to map a collection of events to several IoT modules according to the characteristics of acoustic scenes. A set of such devices with acoustic sensors can also be used for tracking sound sources by a dedicated spatial configuration of modules. Additionally, the detection of acoustic events in many cases needs a quick reaction which requires reliable and secure communication [1]. Event in the acoustic scene is often characterized by an abrupt change in energy and frequency properties in various bands of an audio stream. The process of audio event identification involves a comparison of a current signal frame with a time-frequency template. The situation of overlapping events makes it harder to detect due to the shared data in time and frequency domains. The model of events can be represented in various forms, and the detection stage may exploit a matched filtering [9], supervised learning [15] or deep learning [4] approaches. The audio event detection process depends on the many factors and a lot of techniques is applied in the analysis chain. Furthermore, such systems are rarely considered in the context of low memory requirements and computational expenditures.

The selected representation of audio signal based on the time and frequency domains plays an important role in the detection accuracy. Thus, various sets of features, its dependencies and different configurations are used in the analysis. For example, authors in [7] proposed a hierarchical structure with different feature sets with SVM classifier and found for 7 event classes that only MFCC features and their derivatives are more useful for the event classification. A method of using various audio features with a bag-of-features concept for sound events detection with low computational cost has been presented in [3]. The proposed system uses soft quantisation, supervised cookbook learning, and temporal modelling. The feature set includes MFCC, GFCC, loudness and temporal index attributes and the detection stage exploits the SVM classifier with a sliding window approach. The joint properties in time and frequency domains have been used in the work [19]. For overlapping sound event detection, a nonnegative matrix factor 2-D deconvolution and RUSBoost techniques were used. The method exploits spectral and temporal transition characteristics of the audio signal using features calculated from activations obtained from Mel spectrogram. In [10], an analysis of robust sound event recognition in adverse conditions was presented. The proposed technique uses missing feature cepstral coefficients, and ESTI Advanced Front End feature to detect the events in four different types of additive noise.

In this study, a generation and analysis of acoustic events models in time–frequency (TF) plane are presented. We proposed a simple scheme to determine a set of TF units for a given number of acoustic events by preserving its parts with the highest energy. The obtained binary masks for a specific event are then used in the process of classification. The paper is organised as follows. In the next section a process of time–frequency representation calculation, binary mask estimation an classification of acoustic events is introduced. Subsequently, in section 3 an experimental evaluation is described including the thresholding analysis used in binary mask generation and an event detection evaluation process for an example database of audio events. Finally, a short discussion of the proposed technique and obtained results is presented.

2 Audio Event Classification

The process of audio event classification has to identify which time–frequency units explicitly belongs to an acoustic event. For this reason, various audio representations are exploited in existing systems. The distribution of energy in the signal representation depends on the type of sound source and the acoustic conditions like background noise, reverberation and others. The selection in such circumstances require a lot of computational power due to requirement of adaptative mechanisms.

2.1 Peripheral Auditory Representation

As a basic description of audio events we have selected cochleagram due to its importance in machine hearing [11]. The cochleagram is a model that reflects basilar membrane mechanics in the inner ear and is calculated by using gammatone filters which cover the cochlea frequencies range. Also, such representation is more robust to noise in comparison to the spectrogram representation [14]. The audio signal is converted into cochleagram in the following steps [18]:

- Bandpass filtering by a set of gammatone filters in the selected frequency range (e.g. from 50Hz to 8kHz).
- Calculation of the time–domain envelopes using half–wave rectification of signals at the outputs of the filter bank.
- Applying a static nonlinearity function (e.g. square root).

The obtained time–frequency representation has different frequency resolution compared to the spectrogram. The impulse response of gammatone bandpass filters can be expressed in the following form [6]:

$$g_t(t) = t^{n-1} \cdot e^{-2\pi \cdot b(f_0) \cdot t} \cdot \cos(2\pi \cdot f_0 \cdot t), \qquad t \geq 0,$$

where n is the order of the filter, f_0 denotes the filter centre frequency [Hz] and $b(f_0)$ is the bandwidth for a given f_0 frequency. The bands and the centre frequencies of the filters used in gammatone filter bank are estimated according to the equivalent rectangular bandwidth (ERB) of human auditory filters. In our experiments, we have used 4th order ($n = 4$) bandpass filters, and their bandwidth can be approximated with the formula:

$$b(f) \approx 1.019 \cdot (24.7 + 0.108 \cdot f).$$

DOI: 10.1007/978-3-030-22744-9_56

In the filter bank, the centre frequencies of the filters f_0 are located across frequency proportionally to their bandwidths $b(f_0)$. An example set of the gammatone bandpass filters is depicted in Figure 1.

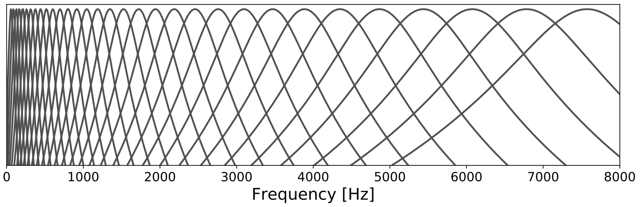


Fig. 1. A part of the gammatone filter bank (32 out of 128 band-pass filters are shown for clarity) used in the cochleagram calculation.

2.2 Binary Mask Computation

To extract TF units from cochleagram specific for a defined group of events, we decided to use a masking scheme. The binary mask is computed in the training phase where for a set of events the energy of TF units are amplified by increasing the shared values of consecutive events. All input training data is aligned and optionally interpolated in the time domain for the unification of its size. Then every cochleagram is thresholded using L_1 value and added up to binary mask template. The resulting temporary mask is eventually thresholded using level L_2, and then after binarisation, the final representation is obtained. This process converts every TF unit in the mask to value 1 when source value is greater than zero else, it replaces with value 0. The whole described scheme is illustrated in Figure 2.

The thresholding operation is described by the following formula for both modules where $k = 1, 2$:

$$H_k(x) = x \cdot \left\lfloor \frac{\operatorname{sgn}(x - L_k) + 1}{2} \right\rfloor . \tag{1}$$

The motivation behind this scheme is the selection of the TF units with the dominant energy in the events. The TF units selected below the threshold are removed from the mask template. An essential assumption in this scheme is that all events used in the binary estimation process have to be time aligned according to their onsets. The final mask depends on the number of input audio items in the training set. For example, in Figure 3, an evolution of the mask structure depending on the number of input events is shown. The obtained masks can be efficiently coded and stored with a small memory footprint due to their sparsity and binary representation.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_56

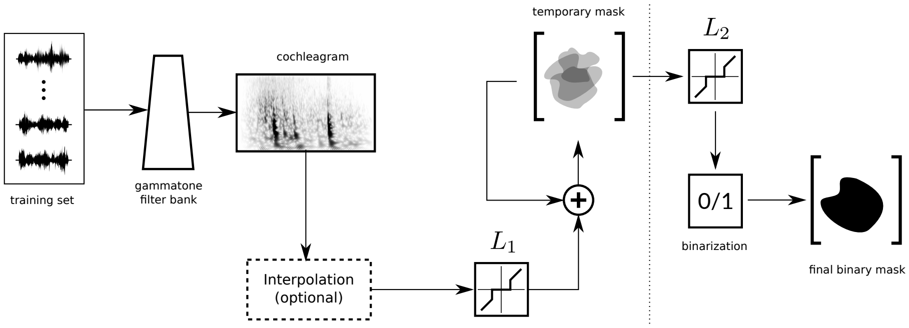


Fig. 2. The computational scheme for binary mask generation.

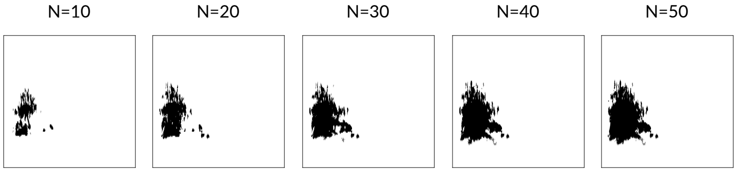


Fig. 3. An evolution of binary mask for different number of training signals.

Because the thresholding operation removes TF units, the analysis window may be reduced after mask estimation. Initially, the window has the size equal to the longest event in the set. In Figure. 4 the final width of the binary mask is presented for 5 example events. The duration of the event's mask is dependent on the type of sound source and acoustic conditions of recorded audio templates.

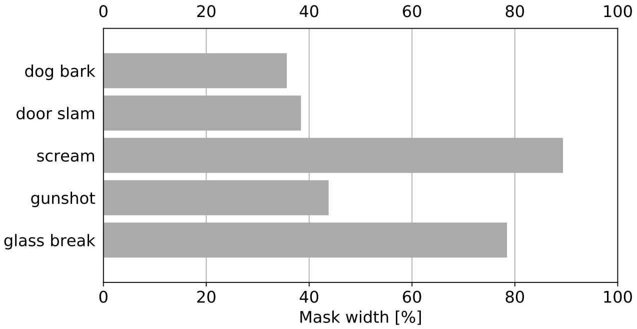


Fig. 4. The width of the final mask in the analysis window.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_56

2.3 Events Classification

Let's consider a set of event classes $\Theta = \{C_1, C_2, C_3 \ldots\}$. The event classification in our approach is performed by multiply the cochleagram with a binary mask for successive classes. The class presence probability can be expressed as follows:

$$P(\theta) = \frac{\mathbf{e}^{\mathrm{T}} \cdot (\mathbf{A} \circ \mathbf{B}_\theta) \cdot \mathbf{e}}{\mathbf{e}^{\mathrm{T}} \cdot \mathbf{B}_\theta \cdot \mathbf{e}}, \tag{2}$$

where $\mathbf{A}$ denotes input cochleagram, $\mathbf{B}_\theta$ is the final mask for class θ, $\mathbf{e}$ is the all-ones vector, and $\circ$ is the Hadamard product operator.

The final result is determined by the selection of the class with the highest probability:

$$\tilde{\theta} = \underset{\theta \in \Theta}{\arg\max} \, [P(\theta)] \,. \tag{3}$$

After calculating the probabilities for all classes, the additional rules can be applied to improve the final classification accuracy. However, in this study, we have just selected the event with the highest probability value.

3 Experimental Evaluation

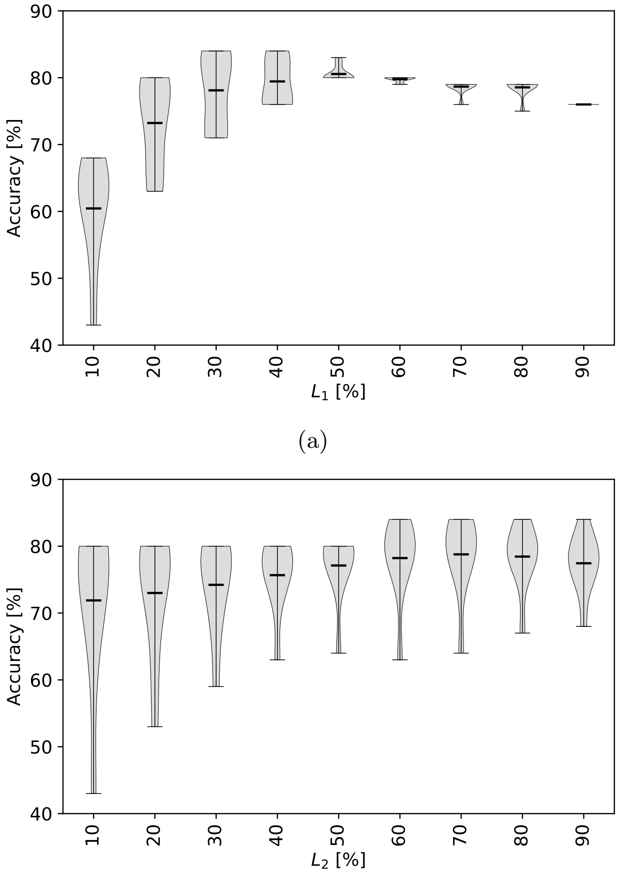
The performance of the proposed technique was evaluated by using a set of acoustic events recorded in clean conditions with one channel and 44.1kHz sampling rate. The dataset we used in the experiments contains five different acoustic events occurring in acoustic surveillance situations. The events include *'screaming'*, *'dog bark'*, *'gunshot'*, *'door slam'* and *'glass break'* sounds. Every event is in isolated form and is aligned in the class to the time onsets. The total number of items in the set contains 250 individual recordings with 70/30 data split to use as training and testing sets.

3.1 Thresholding Analysis

In the proposed approach two parameters have a direct influence on the mask generation. At each iteration of mask creation the input TF plane is thresholded using L_1 level, then the final mask is additionally thresholded using level L_2. In this way, the performance of the system can be tuned to the acoustic environment. The values are determined as a percentage value of the whole dynamic range of the cochleagram. To verify how the thresholds affect the effectiveness of classification, we have generated binary masks for all the combinations of both L_1 and L_2 values with step in subsequent attempts equal to 10%. The classification results for five classes are depicted in Figure. 5.

To illustrate the results we have decided to use violin plots [5] as it additionally shows local density estimates. For the analysed dataset the best accuracy has been achieved with $L_1 = 30\%$ and $L_2 = 70\%$. Selection of these parameters to obtain the best results should be performed whenever a new dataset will be used.

DOI: 10.1007/978-3-030-22744-9_56



(a)

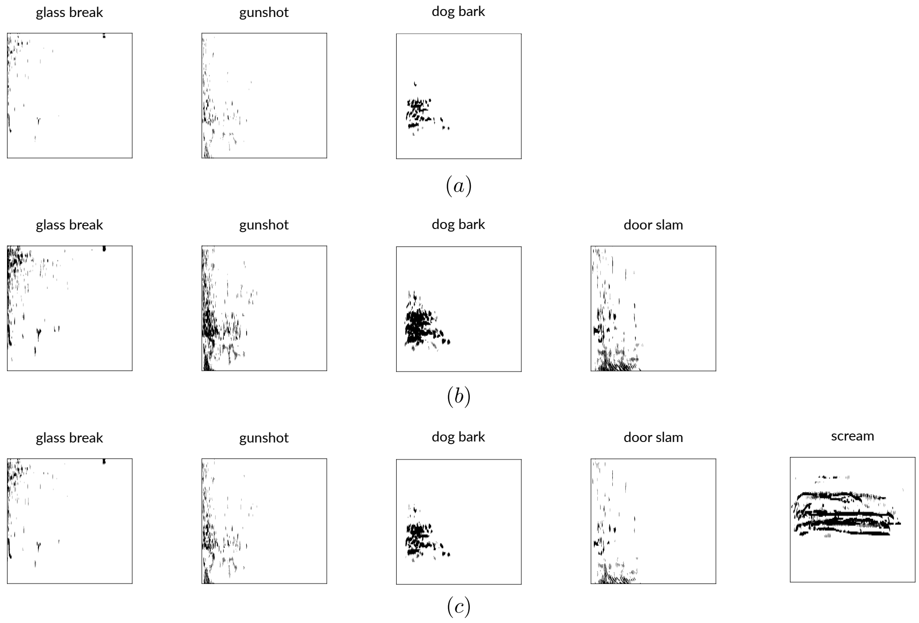
(b)

Fig. 5. The influence of thresholding on the classification accuracy for 5 classes with gradual changes of level L_1 (a), and L_2 (b).

3.2 Classification

We evaluated the performance of our technique for three different sets of randomly selected events contained 3 (*'dog bark'*, *'gunshot'*, *'glass break'*), 4 (*'dog bark'*, *'gunshot'*, *'door slam'*, *'glass break'*) and 5 (*'screaming'*, *'dog bark'*, *'gunshot'*, *'door slam'*, *'glass break'*) classes. Then for each case a binary mask was estimated as is shown in Figure. 6.

For the first set (Figure. 6a), the best thresholds were equal to $L_1 = 60\%$ and $L_2 = 80\%$. In situation of four classes (Figure. 6b) the best result was achieved with $L_1 = 40\%$ and $L_2 = 60\%$. Finally, the last case (Figure. 6c) with five



(a)

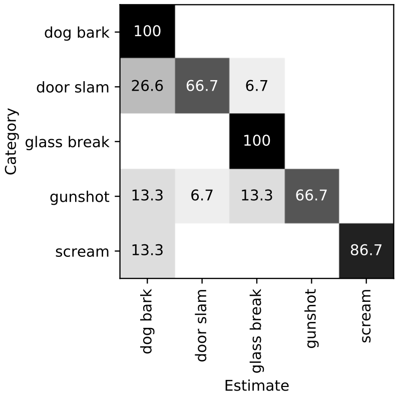
(b)

(c)

Fig. 6. The final representations of binary masks for 3, 4 and 5 classes obtained for the best thresholds L_1 and L_2 in each case.

classes was obtained with thresholds $L_1 = 30\%$ and $L_2 = 70\%$. The classification accuracies for each case were equal to 95.6%, 83.3% and 84% respectively. It's interesting to observe that in every case the second threshold L_2 is bigger than L_1 which suggest that more TF units selected from source cochleagrams are omitted than in the final thresholding before binarisation.

For the last case, a confusion matrix is presented in Figure. 7. It follows that events *'dog bark'* and *'glass break'* were recognized perfectly, while the most mistakes occur for *'gunshot'* and *'door slam'* classes. The occurring mistakes are related to similarities in the shared frequency band and the similarities in the duration. The main reason for misclassification is the variability of physical properties of sound sources. The changes are rather small, but they have a direct impact on the computed mask. Moreover, the range of frequencies in cochleagrams calculated in our study was limited to 50-8000 Hz range what could have been influenced the final representation of the mask. Finally, it is difficult to indicate unambiguously the way to adjust the parameters of the proposed system to maximise the classification accuracy. As always it is a kind of the trade–off between the efficiency and the computational cost. Despite the low computational expenditures, the proposed approach has to be adapted to the application taking into account the events recorded in the target acoustic conditions.



Category \ Estimate	dog bark	door slam	glass break	gunshot	scream
dog bark	100				
door slam	26.6	66.7	6.7		
glass break			100		
gunshot	13.3	6.7	13.3	66.7	
scream	13.3				86.7

Fig. 7. The confusion matrix for 5 classes.

4 Conclusion

A computationally effective and straightforward approach to acoustic event classification is presented. The proposed method uses binary masks to determine the discriminative TF units for the joint structure of the same type of acoustic events. The fast parametrisation and classification stages along with a very low memory requirement make the proposed approach attractive to low–resource and low–power applications. Such a solution may be implemented as an auxiliary module with low computational resources to introduce additional information in multimodal systems used in smart environments. As an example application, we have selected a simple surveillance system with five specific acoustic events. The performed experiments show how to configure the mechanism of binary mask estimation. The achieved classification accuracy is acceptable in situations where the number of events is limited to a few. The presented scheme can be easily adapted to real-time analysis using the sliding window approach. In future work, the robustness analysis to background noise and the influence of the overlapping level between events will be investigated.

References

1. Ali, M.I., Ono, N., Kaysar, M., Ush-Shamszaman, Z., Pham, T., Gao, F., Griffin, K., Mileo, A.: Real-time data analytics and event detection for IoT-enabled communication systems. Journal of Web Semantics **42**, 19–37 (2017)
2. Antonini, M., Vecchio, M., Antonelli, F., Ducange, P., Perera, C.: Smart audio sensors in the internet of things edge for anomaly detection. IEEE Access **6**, 67594–67610 (2018)
3. Grzeszick, R., Plinge, A., Fink, G.A.: Bag-of-features methods for acoustic event detection and classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(6), 1242–1252 (June 2017)

4. Hertel, L., Phan, H., Mertins, A.: Comparing time and frequency domain for audio event recognition using deep learning. In: International Joint Conference on Neural Networks – IJCNN'2016. Vancouver, Canada (July 24–29 2016)
5. Hintze, J.L., Nelson, R.D.: Violin plots: a box plot-density trace synergism. The American Statistician **52**(2), 181–184 (1998)
6. Holdsworth, J., Nimmo-Smith, I., Patterson, R., Rice, P.: Implementing a gammatone filter bank. Annex C of the SVOS final report (Part A: The auditory filter bank) APU (Applied Psychology Unit) Report 2341, Cambridge, UK (February 1988)
7. Huang, W., Lau, S., Tan, T., Li, L., Wyse, L.: Audio events classification using hierarchical structure. In: Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, and Fourth Pacific Rim Conference on Multimedia. vol. 3, pp. 1299–1303. Singapore (December 15-18 2003)
8. Jansen, A., Gemmeke, J.F., Ellis, D.P.W., Liu, X., Lawrence, W., Freedman, D.: Large-scale audio event discovery in one million youtube videos. In: 42th IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP'2017. New Orleans, USA (March 5–9 2017)
9. Kintzley, K., Jansen, A., Hermansky, H.: Event selection from phone posteriorgrams using matched filters. In: 12th Annual Conference of the International Speech Communication Association – Interspeech'2011. pp. 1905–1908. Florence, Italy (August 27–31 2011)
10. Leng, Y.R., Tran, H.D.: Using blob detection in missing feature linear-frequency cepstral coefficients for robust sound event recognition. In: 13th Annual Conference of the International Speech Communication Association – Interspeech'2012 (2012)
11. Lyon, R.F.: Human and Machine Hearing. Cambridge University Press (May 2017)
12. McFee, B., Salamon, J., Bello, J.P.: Adaptive pooling operators for weakly labeled sound event detection. IEEE Transactions on Audio Speech and Language Processing **26**(11), 2180–2193 (2018)
13. Navarro, J., Vidaa-Vila, E., Alsina-Pags, R.M., Hervs, M.: Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios. Sensors **18**(8), 2492 (2018)
14. Sharan, R.V., Moir, T.J.: Cochleagram image feature for improved robustness in sound recognition. In: IEEE International Conference on Digital Signal Processing – DSP'2015. pp. 441–444. IEEE, Singapore (July 21–24 2015)
15. Sharma, A., Kaul, S.: Two-stage supervised learning-based method to detect screams and cries in urban environments. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(2), 290–299 (February 2016)
16. Siantikos, G., Sgouropoulos, D., Giannakopoulos, T., Spyrou, E.: Fusing multiple audio sensors for acoustic event detection. In: 9th International Symposium on Image and Signal Processing and Analysis – ISPA'2015. pp. 265–269 (Sep 2015)
17. Takahashi, N., Gygli, M., Pfister, B., Gool, L.V.: Deep convolutional neural networks and data augmentation for acoustic event detection. In: 17th Annual Conference of the International Speech Communication Association – INTERSPEECH'2016. San Francisco, USA (September 8–12 2016)
18. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis – Principles, Algorithms, and Applications. IEEE Press / Wiley–Interscience (2006)
19. Yang, W., Krishnan, S.: Sound event detection in real-life audio using joint spectral and temporal features. Signal, Image and Video Processing **12**(7), 1345 (Oct 2018)