UDC 004.934:141.1 **doi:** https://doi.org/10.20998/2522-9052.2022.1.11

Andriy Kovalenko, Anton Poroshenko

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

ANALYSIS OF THE SOUND EVENT DETECTION METHODS AND SYSTEMS

Abstract. Detection and recognition of loud sounds and characteristic noises can significantly increase the level of safety and ensure timely response to various emergency situations. Audio event detection is the first step in recognizing audio signals in a continuous audio input stream. This article presents a number of problems that are associated with the development of sound event detection systems, such as the deviation for each environment and each sound category, overlapping audio events, unreliable training data, etc. Both methods for detecting monophonic impulsive audio event and polyphonic sound event detection methods which are used in the state-of-the-art sound event detection systems are presented. Such systems are presented in Detection and Classification of Acoustic Scenes and Events (DCASE) challenges and workshops, which take place every year. Beside a majority of works focusing on the improving overall performance in terms of accuracy many other aspects have also been studied. Several systems presented at DCASE 2021 task 4 were considered, and based on their analysis, there was a conclusion about possible future for sound event detection systems. Also the actual directions in the development of modern audio analytics systems are presented, including the study and use of various architectures of neural networks, the use of several data augmentation techniques, such as universal sound separation, etc.

Keywords: sound event detection; sound event recognition; monophonic sounds; polyphonic sounds; standard deviation; median filter; dynamic threshold; sound separation.

Introduction

Various illegal actions of people, man-made accidents or natural disasters are most often preceded or accompanied by loud sounds and characteristic noises. Detection and recognition of such sounds can significantly improve safety and ensure timely response to such emergencies.

Sound event detection has become an active area of research in recent years. The main reason for this is the holding of DCASE Workshop and DCASE Challenge during the last years. So far, beside a majority of works focusing on the improving overall performance in terms of accuracy many other aspects have also been studied.

The process of recognizing an audio signal can be generally described by the sequential execution of several steps, namely, detecting the audio event in the incoming data stream, obtaining information about the distinctive features of the initial audio signal (feature extraction) and classifying the resulting features [1]. The audio signal recognition process is shown in Fig. 1.

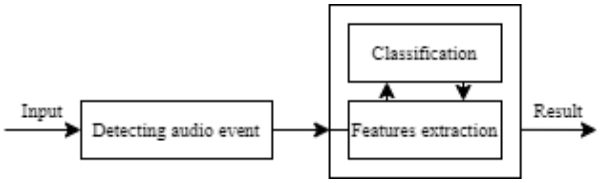

Fig. 1. The process of recognizing an audio signal

The detection task is usually much more difficult than the classification task [2]. This is primarily due to the fact that the detection task must distinguish not only the categories of events, but also the target categories of events from rich background sounds. Also, during classification, there is access to the global context of events, while in the detection task this context must first be determined, usually using unreliable local features of the audio signal.

In turn, in complex systems for detecting audio events can be used a verification of detected audio events [2]. The verification system is used to reduce the risk of an unknown audio event occurring during the classification step. After detecting an audio event, the system has access to the estimated limits of the detected event, has access to its global context, and can calculate its global features and perform verification. As a result of this approach, the inconsistency between the training and testing data is reduced, which leads to an improvement in the performance of the classifier. When using verification, the number of false positives is also reduced, since the correspondence between detection labels and classification labels is checked. The audio event detection process is shown in Fig. 2.

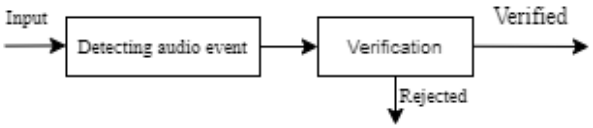

Fig. 2. The process of detection an audio signal with the verification step

Most often, systems for automatic detection and recognition of audio events are developed for specific tasks and environments [3]. There are a number of problems when using a multi-environment system and a large set of event categories. The deviation for each environment and category makes it difficult to automatically recognize audio event. Additionally, this task is complicated by the presence of overlapping audio events. Such audio events are called polyphonic. An example of polyphonic audio pods is shown in Fig. 3. Despite the fact that in real life predominantly occur polyphonic audio events, methods for detecting monophonic audio events have the right to exist, for example, in security systems for detecting gunshots.

One of the main problems of methods for detecting and classifying sound events is the problem of training data [4].

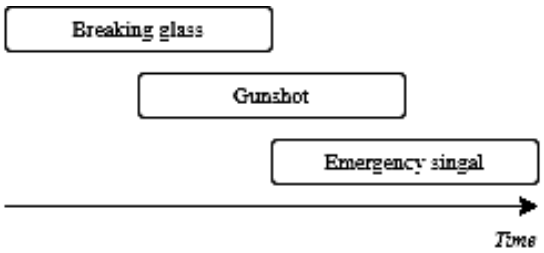

Fig. 3. Polyphonic audio events

To ensure high detection and classification accuracy, it is necessary to use a data set consisting of strictly labeled soundscapes containing time stamps for the beginning and the end of an audio event. However, such strict labeling of a sufficiently large data set is challenging, and annotations with such labels are highly likely to contain human errors and inconsistencies, especially given the ambiguity in the perception of the start and the end of some sound events. One of the solutions to this problem can be the artificial creation of soundscapes, and their designation with strict labeling [4,5]. But this can lead to a mismatch between the artificial data for training and the recorded live data that is used during the solution, which, in turn, can lead to classifier failure. Another solution to this problem can be the use of weakly labeled data that does not contain timestamps, but has information only about the presence of an event in the record [6]. However, this solution has a number of disadvantages related to the difference in the length of the audio events and the presence of background noise. This article discusses methods for sound event detection.

Section II presents the existing methodologies for detecting monophonic impulsive sounds with energy calculation for consecutive non-overlapping blocks,

Section III presents the state-of-the-art polyphonic sound event detection systems. In Section IV potential directions for the development of such systems are given.

Monophonic impulsive sound detection

Most methods for detecting monophonic impulsive audio events are based on determining the energy for a set of consecutive non-overlapping blocks [1, 7]. The various methods differ in the way in which the block corresponding to the sharp impulsive sound is automatically detected: based on the standard deviation of the normalized block energy values; based on the use of a median filter for block energy values; based on dynamic threshold for block energy values. All these methods are based on the goal of minimizing computational complexity and ensuring the highest possible detection accuracy.

The key aspect of the method based on the standard deviation of the normalized values is the normalization of the input set of energy block values to the range from zero to one. Next, the standard deviation of the resulting set of values is calculated. In the case of background noise, the energy values of the blocks will be approximately evenly distributed in the range from zero to one. Since when a new energy value for an audio signal block arrives, the values are renormalized to the specified range, when a block with a significantly higher energy level arrives, the standard deviation value will significantly decrease compared to the same value for a set of previous background blocks. By reducing the standard deviation below the threshold value, it is possible to automatically detect a block with an impulsive signal.

The advantage of the method based on the standard deviation is its resistance to noise and the ability to detect a slowly varying signal by analyzing the average normalized value of block energies.

The next method for detecting audio events is a method based on the use of a median filter. Median filters are often used in practice for preprocessing digital data. Their properties make it possible to apply median filtering to eliminate anomalous values in data arrays, reduce outliers and impulse noise.

To detect a block with an impulse event, a conditional median filter is applied, which leaves the initial signal value if the difference between the initial sample and the median value is less than the threshold value and the median value otherwise. The result of the filtering is the filtered signal and the filtering residue, which indicates that audio event has occurred. By calculating the difference between the signal after applying the conditional median filter and the biased output signal, we can automatically select a block with an impulsive event.

The advantages of using the median filter include only its simple structure, which makes easy to release its hardware and software implementations with relatively low computational complexity. It has a number of disadvantages associated with the fact that the filter does not work in conditions of single impulsive noise, and as the filter window size increases, abrupt signal changes are blurred.

The dynamic threshold method proposes to detect a pulsed signal using the average of a set of block energies and the standard deviation as the dynamic threshold. Automatic detection occurs when the energy of the next block exceeds the threshold value.

The dynamic threshold method has the advantages of both of the past methods and avoids several of the disadvantages. But despite this, it is quite difficult to use it. This is due to the requirement to set the sensitivity parameter of the algorithm, which complicates the hardware implementation. The need to calculate the sensitivity parameter of the algorithm, taking into account the location and direction of the equipment, makes this method difficult to implement, but effective when applied correctly. The accuracy of audio event detection with decreasing signal to noise ratio is shown

in Table 1. As a data set, a set of 476 impulsive audio events is used, each with an appropriate labeling with a timestamps. All signals were digitized and sampled at 44.1 kHz.

After an audio event is detected, additional features are extracted and the detected audio event is classified by one of the known classification methods [8-10].

Table 1 – **The accuracy of audio event detection with decreasing signal to noise ratio**

SNR, dB	Accuracy, %		
	Standard deviation method	Median filter method	Dynamic threshold method
5	100	100	100
0	99,79	99,37	99,58
-5	97,69	97,06	97,48
-10	81,52	80,05	81,1
-15	41,81	40,34	40,55
-20	0,42	0,21	0,63

Polyphonic sound event detection

In state-of-the-art sound event detection systems, detection, verification and classification are closely linked, and they are often combined into one to ensure maximum classification accuracy. These systems are presented in Detection and Classification of Acoustic Scenes and Events (DCASE) challenges and workshops, which take place every year.

For example, in DCASE2021 challenge, there is a task 4 called "Sound Event Detection and Separation in Domestic Environments". The purpose of the test is to evaluate sound event detection systems using real weakly labeled data and strictly labeled simulated data.

The task evaluates sound event detection systems that are trained on weakly labeled data that does not contain timestamps. The purpose of systems is to provide not only the class of the event, but also the localization of the event in time, given that there may be several events in an audio recording. participants are provided with isolated sound events, background sound files, and scripts for developing a training set with strictly labeled synthetic data.

The data for the DCASE 2021 task 4 consist of several datasets designed for sound event detection, such as DESED [11] (dataset with 6 subsets, 4 with recorded data and 2 with synthetic, with different annotations), SINS and TUT Acoustic scenes 2017 development dataset (background sound without annotations), FUSS and FSD50K datasets (isolated events and recorded soundscapes with weak annotations), YFCC100M dataset (recorded soundscapes without annotations).

There is a baseline solution [4], which uses a mean-teacher model based on neural networks. This model is a combination of two models with same architecture: a student model and a teacher model.

The teacher model aims at helping the student model during training, while the student model is the final model.

During training, the teacher and student models are given the same input data, but the teacher model input has additive Gaussian noise, which allows the student model to be trained using consistency loss for both strong and weak predictions for all the clips in the batch.

There is also an attempt to improve sound event detection using sound separation [12] as pre-processing to a sound event detection system.

The task of audio separation is to restore or reconstruct one or more original signals that are mixed with other signals as a result of a linear or convolutional process. This area of research has many practical applications, including sound quality improvement and noise elimination, music remixing, sound spatialization, remastering, etc. However, recent works has demonstrated that the universal sound separation can be used to separate sounds of arbitrary classes [13,14]. As it turns out, combination of sound separation and sound event detection showed the potential to improve the performance of sound event detection systems, however the benefits are still limited due to a mismatch between the sound separation training conditions and the sound event detection test conditions [12].

There are many systems for sound event detection in the DCASE 2021 task 4, which are based on the baseline system [15-20].

In 2021 challenge, system with the best ranking score proposed three major change over the baseline solution [15], namely the use of the selective kernel unit, the use of soft detection output by setting proper temperature parameter in sigmoid and the use of several data augmentation techniques. This changes allows neuron to adaptively adjust for both short- and long- duration events, and overall improves in stability and robustness of the system performance.

Second place in systems rankings [16] proposes the sound event detection model, which is based on self-training with a noisy student model. It is proposed to use an RCRNN-based mean-teacher model to predict the target label of each audio clip. Data augmentation-based feature noise, dropout-based model noise, and semi-supervised loss function based label noise were used to realize self-training,

The third place in systems rankings [17] suggests using both recurrent structure and transformer structure to model the complicated dynamics in real life domestic audio data. This was done in order to provide an overall performance boost over the baseline solution, since different models exhibit differently under the different sce-

narios. Additionally, semi-supervised mean-teacher learning and different data augmentations are used.

Discussion

The task of analyzing sound signals is inextricably linked with the entire history of human as a living organism, since it is critically important to obtain information about what is happening around us. Audio analytics systems aim to automatically extract important information from audio signals. They include such disciplines as sound scene classification or sound event detection and classification, and many others.

For further development of such systems, many problems must be overcome, for example, the deviation for each environment and each sound category, overlapping audio events, unreliable training data, etc. However, the potential inherent in audio analytics systems is incredibly high, which is confirmed by modern discoveries and state-of-the-art sound event detection systems. The current directions in the development of modern audio analytics systems are the study and use of various architectures of neural networks, the use of several data augmentation techniques, such as universal sound separation, etc.

Conclusions

Thus, this article presents both methods for detecting monophonic impulsive audio event and polyphonic sound event detection methods which are used in the state-of-the-art sound event detection systems. Also the actual problems associated with the development of such systems are presented.

Three methods for detecting monophonic pulsed audio events have been implemented, namely the method based on the standard deviation of normalized block energies, the method based on applying a median filter for block energies, and the dynamic threshold method for block energies.

The results of their work and comparative analysis are also presented.

Several systems presented at DCASE 2021 task 4 were considered, and based on their analysis, there was a conclusion about the potential directions for the further development of audio analytics systems.

REFERENCES

1. Alain, Dufaux, Laurent, Besacier, Michael, Ansorge, and Fausto, Pellandini (2000), “Automatic sound detection and recognition for noisy environment”, *2000 10th European Signal Processing Conference*, IEEE, pp. 1-4.
2. Phan, Huy, Koch, Philipp, Katzberg, Fabrice, Maass, Marco, Mazur, Radoslaw, McLoughlin, Ian and Mertins, Alfred (2017), “What makes audio event detection harder than classification?”, *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2739-2743, doi: https://doi.org/10.23919/EUSIPCO.2017.8081709.
3. Sami Ur, Rahman, Adnan, Khan, Sohail, Abbas, Fakhre, Alam and Nasir Rashid (2021), “Hybrid system for automatic detection of gunshots in indoor environment”, *Multimedia Tools and Applications*, Vol. 80, No. 3, pp. 4143-4153, doi: https://doi.org/10.1007/s11042-020-09936-w.
4. Nicolas, Turpault and Romain, Serizel (2020), Training Sound Event Detection On A Heterogeneous Dataset. DCASE Workshop, Nov, Tokyo, Japan, available at: https://arxiv.org/abs/2007.03931.
5. J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello (2017), “Scaper: A library for soundscape synthesis and augmentation”, *Proc. WASPAA*, pp. 344–348, doi: https://doi.org/10.1109/WASPAA.2017.8170052.
6. A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, “A closer look at weak label learning for audio events,” in arXiv:1804.09288. [Online]. available at: http://arxiv.org/abs/1804.09288.
7. Порошенко А.І. Методи та підходи до детектування аудіоподій різних типів [Текст] / А.І. Порошенко, А.А. Коваленко // Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління. Матеріали одинадцятої міжнародної НТК. – Баку: ВА ЗС АР; Харків: НТУ «ХПІ»; Київ: НАУ; Харків: ДП «ПДПРОНДІАВІАПРОМ»; Жиліна: УмЖ, 2021. – 8-9 квітня 2021. – Т.2. – С. 114.
8. Порошенко, А.І. Методи класифікації ознак аудіосигналів [Текст] / А.І. Порошенко, А.А. Коваленко // Проблеми інформатизації : тези доп. 9-ї міжнар. наук.-техн. конф., 18-19 листопада 2021 р., м. Черкаси, м. Харків, м. Баку, м. Бельсько-Бяла : [у 3 т.]. Т. 1 / Черк. держ. технолог. ун-т [та ін.]. – Харків : Петров В. В., 2021. – С. 90.
9. K. Kumar and K. Chaturvedi, "An Audio Classification Approach using Feature extraction neural network classification Approach," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: https://doi.org/10.1109/IDEA49133.2020.9170702.
10. K. Hirata, T. Kato and R. Oshima, "Classification of Environmental Sounds Using Convolutional Neural Network with Bispectral Analysis," 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2019, pp. 1-2, doi: https://doi.org/10.1109/ISPACS48206.2019.8986304.
11. Romain Serizel, Nicolas Turpault, Ankit Shah, Justin Salamon. Sound event detection in synthetic domestic environments. ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing, May 2020, Barcelona, Spain.
12. Nicolas Turpault, Scott Wisdom, Hakan Erdogan, John Hershey, Romain Serizel, et al.. Improving Sound Event Detection In Domestic Environments Using Sound Separation. DCASE Workshop 2020 - Detection and Classification of Acoustic Scenes and Events, Nov 2020, Tokyo / Virtual, Japan.
13. E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen and D. P. W. Ellis, "Improving Universal Sound Separation Using Sound Classification," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 96-100, doi: https://doi.org/10.1109/ICASSP40776.2020.9053921.
14. S. Sose, S. Mali and S. P. Mahajan, "Sound Source Separation Using Neural Network," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-5, doi: https://doi.org/10.1109/ICCCNT45670.2019.8944614.
15. Zheng X., Chen H., Song Y. Zheng ustc teams submission for dcase2021 task4 semi-supervised sound event detection. – DCASE2021 Challenge, Tech. Rep, 2021.
16. Kim, N. K. and Kim, H. K. (2021), “Self-training with noisy student model and semi-supervised loss function for dcase 2021 challenge task 4”, available at: http://arXiv:2107.02569.

17. Rui, Lu, Wenzheng, Hu, Zhiyao Duan and Ji, Liu (2021), “Integrating advantages of recurrent and transformer structures for sound event detection in multiple scenarios”, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop* (DCASE2021), Tech. Rep., Challenge.
18. Ebbers, J. and Haeb-Umbach, R. (2021), “Self-Trained Audio Tagging and Sound Event Detection in Domestic Environments”, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop* (DCASE2021), Online, pp. 15-19.
19. Hyeonuk, Nam, Byeong-Yun, Ko, Gyeong-Tae, Lee, Seong-Hu, Kim, Won-Ho, Jung, Sang-Min, Choi and Yong-Hwa, Park (2021), “Heavily augmented sound event detection utilizing weak predictions”, *Detection and Classification of Acoustic Scenes and Events 2021* (DCASE2021), arXiv preprint arXiv:2107.03649.
20. Gangyi ,Tian , Yuxin, Huang, Zhirong, Ye, Shuo, Ma, Xiangdong, Wang, Hong, Liu, Yueliang, Qian, Rui, Tao, Long, Yan, Kazushige, Ouchi, Janek, Ebbers and Reinhold, Haeb-Umbach (2021), “Sound event detection using metric learning and focal loss for dcase 2021 task 4”, Tech. Rep., *Detection and Classification of Acoustic Scenes and Events 2021* (DCASE2021), Challengeb.

Received (Надійшла) 19.11.2021
Accepted for publication (Прийнята до друку) 12.01.2022

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Коваленко Андрій Анатоліойвич – доктор технічних наук, професор, завідувач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Andriy Kovalenko – Doctor of Technical Sciences, Professor, Head of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: andriy_kovalenko@yahoo.com; ORCID ID: https://orcid.org/0000-0002-2817-9036.

Порошенко Антон Ігорович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Anton Poroshenko – postgraduate student at Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: anton.poroshenko@nure.ua; ORCID ID: https://orcid.org/0000-0001-7266-4269.

Аналіз методів та систем детектування аудіоподій

А. А. Коваленко, А .І. Порошенко

Анотація. Виявлення та розпізнавання гучних звуків і характерних шумів дозволяє значно підвищити рівень безпеки та забезпечити своєчасне реагування на різні аварійні ситуації. Детектування аудіоподій – це перший крок у розпізнаванні аудіосигналів з безперервним вхідним аудіопотоком. У даній статті представлено ряд проблем, пов'язаних з розробкою систем виявлення аудіоподій, таких як відхилення для кожного середовища і кожної звукової категорії, звукові події, що перекриваються, недостовірні навчальні дані та ін. Представлені як методи виявлення монофонічних імпульсних звукових подій, так і методи виявлення поліфонічних аудіоподій, які використовуються в сучасних системах виявлення звукових подій. Такі системи представлені у завданнях та семінарах Detection and Classification of Acoustic Scenes and Events (DCASE), які відбуваються щороку. Більшість робіт спрямовані на покращення загальної продуктивності з точки зору точності, хоча також були вивчені багато інших аспектів. Було розглянуто кілька систем, представлених на DCASE 2021 в задачі 4, і на основі їх аналізу був зроблений висновок про можливе майбутнє систем виявлення звукових подій. Також представлені актуальні напрямки розвитку сучасних систем аудіоаналітики, в тому числі вивчення та використання різних архітектур нейронних мереж, використання декількох методів попередньої обробки даних, таких як універсальний розділ звуку та ін.

Ключові слова: виявлення звукових подій; розпізнавання звукових подій; монофонічні звуки; поліфонічні звуки; середньоквадратичне відхилення; медіанний фільтр; динамічний поріг; звуковий поділ.

Анализ методов и систем обнаружения аудиособытий

А. А. Коваленко, А. І. Порошенко

Аннотация. Обнаружение и распознавание громких звуков и характерных шумов позволяет значительно повысить уровень безопасности и обеспечить своевременное реагирование на различные аварийные ситуации. Обнаружение аудиособытий — это первый шаг в распознавании аудиосигналов с непрерывным входным аудиопотоком. В данной статье представлен ряд проблем, связанных с разработкой систем обнаружения звуковых событий, таких как отклонения для каждой среды и каждой звуковой категории, перекрывающиеся звуковые события, недостоверные обучающие данные и т. д. Представлены как методы обнаружения монофонических импульсных звуковых событий, так и методы обнаружения полифонических звуковых событий, которые используются в современных системах обнаружения звуковых событий. Такие системы представлены в задачах и семинарах Detection and Classification of Acoustic Scenes and Events (DCASE), которые проходят каждый год. Большинства работ направленны на улучшение общей производительности с точки зрения точности, хотя также были изучены и многие другие аспекты. Было рассмотрено несколько систем, представленных на DCASE 2021 в задаче 4, и на основе их анализа был сделан вывод о возможном будущем систем обнаружения звуковых событий. Также представлены актуальные направления развития современных систем аудиоаналитики, в том числе изучение и использование различных архитектур нейронных сетей, использование нескольких методов предварительной обработки данных, таких как универсальное разделение звука и др.

Ключевые слова: обнаружение звуковых событий; распознавание звуковых событий; монофонические звуки; полифонические звуки; среднеквадратичное отклонение; медианный фильтр; динамический порог; звуковое разделение.