UDC 004.934 **doi:** https://doi.org/10.20998/2522-9052.2024.4.08

Anton Poroshenko, Andriy Kovalenko

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

AUDIO EVENT ANALYSIS METHOD IN NETWORK-BASED AUDIO ANALYTICS SYSTEMS

Abstract. **Relevance**. In the rapidly evolving field of network-based audio analytics systems, the detection and analysis of audio events play a crucial role across various applications, including security, healthcare, and entertainment. **Subject.** This paper examines a method for recognizing audio events in network-based audio analytics systems, including preprocessing, sound separation, and the creation of machine learning models for analyzing audio signals. **Objective.** The objective is to develop and improve integrated methods for analyzing audio signals in network-based audio analytics systems to enhance the accuracy, speed, and reliability of data analysis. **Methods.** The proposed approach uses a modified ResNet architecture for multi-event classification and a convolutional neural network for separating sound sources in multi-channel recordings. **Results.** The method achieves competitive results, comparable to baseline results in contemporary challenges like DCASE and demonstrates robust performance in noisy environments. **Conclusions.** The proposed method shows potential for improving the accuracy and reliability of audio event recognition in real-world scenarios, particularly in complex acoustic environments.

Keywords: audio signal analysis; preprocessing; audio event detection; sound separation; machine learning; error rate.

Introduction

Relevance. In recent years, the demand for intelligent audio analytics has surged due to the increasing volume of audio data being transmitted and processed across various networks. A network-based audio analytics system is a technology that utilizes signal processing algorithms and machine learning to analyze audio data. This technology can be applied in various fields, including security, healthcare, marketing, and entertainment. For example, it can be used to detect specific sounds or speech patterns in the environment, analyze emotions in voice, recognize speech, and so on. The main components of such a system typically include gathering audio data from the surrounding environment, preprocessing the collected data, transmitting audio data to a server for processing, extracting features and processing the obtained audio data, as well as providing feedback mechanisms.

In the modern world, the increasing amount of information and audio data flowing through networks necessitates the development of effective methods for their analysis and processing. One important area of this analysis is the recognition of audio events in network-based audio analytics systems. This technology enables the detection and classification of various sound phenomena in real-time or based on recordings, using various signal processing algorithms and methods.

Audio event recognition in network-based audio analytics systems is the process of identifying and classifying various audio events occurring in audio recordings by understanding their content and context. This process can be used for automatically detecting events such as voice commands, noises, speech, transportation sounds, explosions, weather conditions, and more.

An overview of scientific works. Before conducting the classification of audio events, it is necessary to detect, preprocess them and transmit the data to the server. Methods for audio event detection are outlined in [1, 2] Preprocessing audio data is a critical step in the audio analysis process, as it prepares raw audio data for further processing and analysis. Preprocessing may include noise filtering, alignment, segmentation, and echo removal. The specifics of preprocessing are outlined in [3]. The method of transmitting audio signals plays a crucial role in network-based audio analytics systems as it determines how collected audio data is transmitted from the recording source to the analysis and processing systems. The method of transmitting audio signals in network-based audio analytics systems is proposed in [4, 5].

The quality of the developed machine learning or deep learning model significantly depends on the quality and relevance of the dataset on which it was trained. The dataset should be representative of the real world and free from errors, missing values, or incorrect labels. Machine learning methods are discussed in [6], and their application to audio in [7. 8]. The main datasets for training sound separation and sound event detection models are listed in [9–11]. Sound separation as a preprocessing stage in the method of audio event recognition in network-based audio analytics systems is important for increasing the accuracy and efficiency of analysis. In the real world, audio signals often contain a mixture of different sounds - speech, music, ambient noise, etc. Sound separation allows isolating these components for more accurate analysis. The specifics of modern sound separation methods are detailed in [12, 13].

Thanks to the annual DCASE Challenge [14], audio event recognition methods are constantly evolving. The DCASE Challenge provides unified criteria for evaluating and comparing different methods of audio event recognition. This helps identify the strengths and weaknesses of various approaches and contributes to their further improvement. Each new edition sets more challenging tasks for researchers, such as event recognition in multi-sound conditions or in real-time, stimulating the development of more advanced technologies. Modern audio event recognition methods are outlined in [14, 15].

Definition of key aspects of the general problem

Analysis of audio signals in network-based audio analytics systems consists of several important components that require careful consideration and development. The first part involves preprocessing and detection of audio events. This includes the application of various techniques such as filtering, alignment, and

segmentation for preprocessing audio data before further analysis. The second part involves developing algorithms for separating different audio signals, which typically contain a mixture of various sounds. This requires the development of sound separation methods and their optimization for real-time operation. The third part is the creation of machine learning models for audio event analysis. This part involves developing machine learning algorithms for classification and recognition of audio events. Challenges in this part include selecting appropriate models, training on relevant data, and managing the reliability and accuracy of the model.

Setting objectives. Therefore, there is a need to develop and enhance comprehensive methods for analyzing audio signals in network-based audio analytics systems. Based on this, **the goal of this paper** is to develop and improve integrated methods for analyzing audio signals in network-based audio analytics systems to enhance the accuracy, speed, and reliability of audio data analysis. To achieve this, it is necessary to develop an audio signal analysis method that takes into account preprocessing, sound separation, and the stage of creating machine learning models for audio event analysis.

Justification of the quality metric for the transmission and analysis of audio signals

Justification of the quality metric for the transmission and analysis of audio signals is a critically important task in the development of audio analytics systems. These metrics may encompass various aspects such as sound quality, speech recognition accuracy, compression efficiency, and others. It is essential to consider the specific context and objectives of the system when justifying the selection of particular quality metrics.

The quality metric should account for the accuracy of the audio signal recognition system. This may involve correctly identifying sound events, speech recognition, or other audio signal characteristics according to the requirements of the specific context. High accuracy in speech or sound recognition is crucial for systems used in tasks such as voice command understanding, audio transcription, and more. Improving recognition accuracy contributes to better user interaction and ensures proper handling of audio data. However, the quality of analysis should be justified in terms of the resources expended by the system. For example, optimizing algorithms to use minimal computational and memory resources.

The analysis quality should be robust against noise and other interferences that may affect the audio signal. This is important to ensure the adequacy of results in different conditions. Maintaining system stability in noisy environments helps maintain high-quality audio data processing in real-world conditions. In real-time applications such as voice chat systems or video conferencing, minimizing delays in audio signal transmission and processing is crucial. Low latency helps avoid a sense of delay in interacting with the system and ensures smooth interaction. Additionally, in cases of audio data transmission over the network, it is important to use efficient compression methods to reduce data volume. Reducing data volume contributes to faster transmission and preserves network bandwidth. The quality metric may consider the system's scalability to handle large amounts of audio data or to expand functionality in the future. The system should be able to adapt to different conditions and be configurable depending on specific user requirements or use cases. If the system integrates with others, the quality metric may consider the efficiency of this interaction and compatibility with other technologies. To assess the accuracy of the model, so-called Event-based metrics such as F1-score and Error Rate (ER) are used. A key advantage of event-based metrics is their ability to provide detailed analysis of event detection quality, allowing developers to better understand and optimize their systems for specific requirements and usage scenarios. F1-score is considered as the harmonic mean between precision and recall. To calculate these values, true positives (TP), false positives (FP), and false negatives (FN) are used. Precision is a measure of how many selected items are correct. The precision value is calculated as:

$$P = \frac{TP}{TP + FP}.$$

Recall is a measure of how many relevant items were selected. The formula for recall is:

$$R = \frac{TP}{TP + FN}.$$

The F1-score value is calculated as:

$$P = \frac{2PR}{P + R}.$$

The Error Rate, as a metric in audio event recognition tasks, utilizes the values of substitutions (S), insertions (I), deletions (D), and the number of reference events (N) to determine the effectiveness of the model. This metric is particularly crucial in scenarios where accurate recognition and classification of audio events are critical.

Substitutions (S) represent the number of cases where the model incorrectly identifies one event as another. Insertions (I) indicate false alarms, where the model detects an event that did not actually occur. Deletions (D) signify the instances where the model misses a real event. Reference events (N) correspond to the total number of events that should have been detected according to the reference data. Based on these values, the overall event error rate (EER) is calculated, typically expressed using the following formula:

$$ERR = \frac{S + D + I}{N}.$$

This metric allows for the assessment of the overall effectiveness of the system in detecting and recognizing audio events, taking into account all possible types of errors. A low value of EER indicates high accuracy and efficiency of the system, while a high value of EER indicates issues in event detection or classification.

Dataset

FSD50K is a large dataset used for tasks related to sound recognition. This open dataset contains over 51,000 audio clips totaling more than 100 hours, manually labeled using 200 classes from the AudioSet ontology. The licensing of FSD50K audio clips under Creative Commons makes them freely distributable and accessible.

This open accessibility is a significant advantage compared to some other datasets, making it an invaluable resource for researchers and developers worldwide. Sounds in FSD50K cover a wide range of categories, including natural sounds, music, human voices, urban sounds, and many others. Each sound in the dataset is accompanied by metadata, including information about the source, license, tags, descriptions, and more.

Preparing data from FSD50K for use in the model is essential. This includes selecting relevant audio recordings from the dataset that correspond to the event classes to be recognized. For training the sound event recognition method, 25 sound classes with loud audio events such as Alarm, Gunshot, Gong, and similar ones were used. The total dataset comprises 7,763 recordings. All recordings are appropriately labeled and have timestamp information.

Moreover, due to the diversity of audio samples, the FSD50K dataset provides a challenging platform for training machine learning models, as it includes various background noises, overlapping sound events, and differing recording conditions.

Data partitioning into training, validation, and test sets is an important part of preparing data for machine learning tasks. The training set is the primary dataset used for model training. It should be representative and include a wide range of possible cases that the model may encounter in the real world. The validation set is used for fine-tuning model parameters and checking its performance during training. The validation set helps determine when the model starts overfitting to the training data. The test set is an independent dataset used to evaluate the final performance of the model after training and tuning. This dataset allows assessing how the model will perform on new, unseen data, simulating real-world usage. The test set should be entirely independent of the training and validation sets. The overall dataset was divided into training, validation, and test sets in a 70-20-10 ratio. Additionally, each of the samples was combined into a single audio file while preserving the labels, to simulate real-time operation.

Audio event analysis method

The task of sound separation involves obtaining audio tracks of individual sound sources from a single audio recording containing multiple sources. In general, it can be represented as follows:

$$a_{mixed} = \sum_{i=0}^{N} a_i \ .$$

The task of sound separation involves identifying the components a_i of the primary audio file a_{mixed}. To solve this task, audio data is represented in the form of a spectrogram. After detecting anomalies in the input audio stream using a noise threshold and filtering, the audio data is buffered and fed into a special event classifier, which determines the number of simultaneous audio events. The main idea of the proposed method is illustrated in Fig. 1.

If the number of audio events N equals 1, it is directly passed to the audio event classifier; otherwise, simultaneous audio events are determined through the sound separator, and their respective characteristics are obtained. Subsequently, the obtained audio data undergo sound recognition, resulting in N audio events and one from the undistributed record. A separate model called the channel classifier receives all audio data and selects the main audio event.

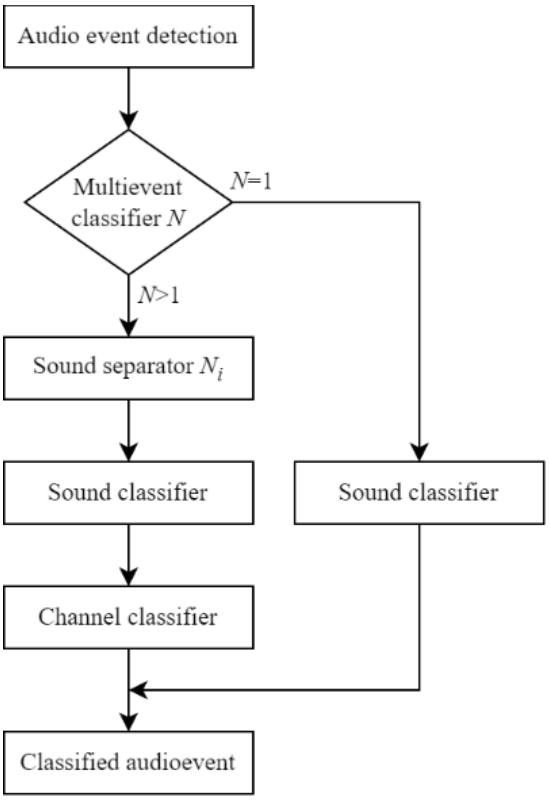

Fig. 1. Audio event analysis method

During the multievent classification stage, it is crucial to quickly determine the algorithm's path. In the case of a single audio event, this stage can significantly increase the overall audio signal processing time. Therefore, the model needs to be relatively small. A modified ResNet on spectrograms with one input and three output classes, corresponding to the possible number of simultaneous audio events, was used as the multievent classifier. Cases where N>3 are rare and require complex solutions for processing.

During the sound separation stage, a method of sound separation using a convolutional neural network (CNN) and binary masks is employed. CNN training is based on spectrograms. Instead of directly reconstructing sounds, CNN is used to create binary masks. A binary mask determines the presence or absence of a specific sound source in each time frame and frequency range of the spectrogram. The obtained binary mask is applied to the input spectrogram to highlight the intensity of those frequencies corresponding to the target sound source, separating the target sound from the rest of the mix.

The CNN consists of two blocks, each containing two convolutional layers. The convolutional layers are used to detect local features in the input data, such as edges, textures, or specific patterns in the spectrogram. Using multiple convolutional layers allows the network to learn more complex and abstract features of the data.

Each block has a max-pooling layer, reducing the data dimensionality by selecting the most significant elements in the pooling window. This helps reduce

computational complexity and control overfitting. After each max-pooling block, dropout is applied. This makes the network less sensitive to specific characteristics of the training data, thereby preventing overfitting.

After passing through the convolutional and max-pooling layers, the data is transformed into a one-dimensional format and fed into a fully connected layer. This layer integrates information from all previously extracted features, allowing the model to make complex conclusions based on the overall information. Using another dropout layer with a value of 0.2 before the last fully connected layer further reduces the risk of overfitting. The last fully connected layer generates a vector of values corresponding to the number of frequency bins in a typical STFT spectrogram. The vector represents binary masks for each frequency component over time.

For sound classification tasks, a basic combination model based on Residual Networks (ResNet) and Recurrent Neural Networks (RNN) is used. ResNet is adapted from the ResNet model of solving image recognition tasks. Combining ResNet and RNN creates a model capable of working with audio data, performing both spatial sound analysis (using ResNet) and temporal dependency analysis (using RNN). Initially, ResNet is used to analyze sound spectrum, and then the obtained features are passed to RNN to analyze time sequences and solve the specific classification task.

The channel classifier performs a classification task over all received mixes. At this stage, all audio signal features, binary masks from the sound separator, and classification results from the sound classifier are available. These inputs are fed into the classifier, and the classification's target variables are established based on the probability associated with each hypothesis.

Results

After constructing, training, and testing the model, F1-score results for the detection and multievent classification of audio events were obtained, which amounted to 56.3%. This corresponds to average values compared to other contemporary methods for solving the audio event detection task. However, for the task of processing audio signals in a networked audio analytics system, which includes tasks of detecting and recognizing audio events, F1-score and ER results were obtained, amounting to 49.6% and 0.44, respectively (Table 1). These results are quite respectable compared, for example, to the baseline in the DCASE 2023 challenge task 3, where 48.7% and 0.54 were obtained, respectively. By achieving comparable or superior F1-scores and lower error rates, the proposed method showcases its potential for robust and accurate audio signal analysis in real-world scenarios.

Additionally, the quality of analysis should be robust to noise and other interferences, facilitated by the presence of the sound separator block. Ensuring the system's stable operation in noisy environments helps maintain high-quality audio data processing in real-world conditions. A particular challenge for the system is the presence of a large number of simultaneous audio events and the classification of loud sounds with high tonality. Conversely, the system demonstrates good results in classifying loud bass sounds.

Table 1 – **F1-score and ER**

Method	F1	ER
The nerc-slip system for sound event localization and detection	62.7%	0.33
Attention mechanism network and data augmentation	58.5%	0.35
The distillation system based on ResNet-Conformer model guided by a ResNet-GRU	51.4%	0.40
Proposed	49,6%	0.44
Baseline	48.7%	0.54
One audio augmentation chain proposed for sound event localization and detection	45.0%	0.48
A framework for SELD using conformer and multi-ACCDOA strategies	33.1%	0.56
Based on omi-dimensional dynamic convolution and feature pyramid attention module	22.1%	0.64

The research results indicate the potential of such an approach in solving the task of audio signal processing in a networked audio analytics system. Future studies may consider expanding and optimizing the neural networks and classifiers used. In the current work, a relatively simple model with a limited number of layers and parameters was used for compactness and training efficiency.

Conclusion

The effectiveness of audio signal processing methods in networked audio analytics systems is an important aspect, as this field has significant practical potential for applications in various domains, including security, medicine, automatic speech recognition, video analysis, and many others.

In this article, the following results were obtained:

1. The quality metric for transmitting and analyzing audio signals has been justified.

2. Datasets for training, validation, and testing of methods for audio event recognition in networked audio analytics systems have been investigated.

3. A method for audio event recognition in networked audio analytics systems based on multievent classifier has been proposed.

Future research in the field of audio signal processing methods in networked audio analytics systems may focus on improving the accuracy and robustness of models for audio event recognition, using more powerful neural networks, new architectures, or more efficient training methods.

REFERENCE

1. Mesaros, A., Heittola, T., Virtanen, T. and Plumbley, M. D. (2021), “Sound Event Detection: A tutorial”, *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, doi: https://doi.org/10.1109/MSP.2021.3090678

2. Xin, Y., Yang, D. and Zou, Y. (2023), "Background-aware Modeling for Weakly Supervised Sound Event Detection", *Proc. INTERSPEECH 2023*, pp. 1199–1203, doi: https://doi.org/10.21437/Interspeech.2023-330
3. Barkovska, O. and Havrashenko, A. (2023), "Analysis of the influence of selected audio pre-processing stages on accuracy of speaker language recognition", *Innovative technologies and scientific solutions for industries*, vol. 4 (26), pp. 16–23, doi: https://doi.org/10.30837/ITSSI.2023.26.016
4. Poroshenko, A. (2022), "Mathematical model of the passage of audio signals in network-based audio analytics systems", *Advanced Information Systems*, vol. 6, no. 4, pp. 25–29, doi: https://doi.org/10.20998/2522-9052.2022.4.04
5. Poroshenko, A. and Kovalenko, A. (2023), "Audio signal transmission method in network-based audio analytics system", *Innovative technologies and scientific solutions for industries*, vol. 4 (26), pp. 58–67, doi: https://doi.org/10.30837/ITSSI.2023.26.058
6. Sharifani, K. and Amini, M. (2023), "Machine Learning and Deep Learning: A Review of Methods and Applications", *World Information Technology and Engineering Journal*, vol. 10, is. 07, pp. 3897–3904, available at: https://ssrn.com/abstract=4458723
7. Grumiaux, P.-A., Kitić, S., Girin, L. and Guérin, A. (2022), "A survey of sound source localization with deep learning methods", J. Acoust. Soc. Am., vol. 152 (1), pp. 107–151, doi: https://doi.org/10.1121/10.0011809
8. Zaman, K., Sah, M., Direkoglu, C. and Unoki, M. (2023), "A Survey of Audio Classification Using Deep Learning", *IEEE Access*, vol. 11, pp. 106.620–106.649, doi: https://doi.org/10.1109/ACCESS.2023.3318015
9. Fonseca, E., Favory, X., Pons, J., Font, F. and Serra, X. (2021), "FSD50K: An Open Dataset of Human-Labeled Sound Events", *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 30, pp. 829–852, doi: https://doi.org/10.1109/TASLP.2021.3133208
10. Piczak, K. J. (2015), "ESC: Dataset for Environmental Sound Classification", *Proceedings of the 23rd ACM International Conference on Multimedia*, MM'15, Association for Computing Machinery, New York, NY, USA, pp. 1015–1018, doi: https://doi.org/10.1145/2733373.2806390
11. Foster, P., Sigtia, S., Krstulovic, S., Barker, J. and Plumbley, M. D. (2015), "Chime-home: A dataset for sound source recognition in a domestic environment", *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, WASPAA, New Paltz, NY, USA, 2015, pp. 1–5, doi: https://doi.org/10.1109/WASPAA.2015.7336899
12. Kavalerov, I., Wisdom, S., Erdogan, H., Patton, B., Wilson, K., Le Roux J. and Hershey J. R. (2019), "Universal Sound Separation", *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, WASPAA, New Paltz, NY, USA, pp. 175–179, doi: https://doi.org/10.1109/WASPAA.2019.8937253
13. Tzinis, E., Wisdom, S., Hershey, J. R., Jansen, A. and Ellis, D. P. W. (2020), "Improving Universal Sound Separation Using Sound Classification", *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP, Barcelona, Spain, pp. 96–100, doi: https://doi.org/10.1109/ICASSP40776.2020.9053921
14. (2023), "Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes", *DCASE Challenge 2023*, available at: https://dcase.community/challenge2023/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes
15. Kovalenko, A. and Poroshenko, A. (2022), "Analysis of the sound event detection methods and systems", *Advanced Information Systems*, vol. 6, no. 1, pp. 65–69. https://doi.org/10.20998/2522-9052.2022.1.11

Received (Надійшла) 11.07.2024
Accepted for publication (Прийнята до друку) 23.10.2024

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Порошенко Антон Ігорович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Anton Poroshenko – Ph.D 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;
https://www.scopus.com/authid/detail.uri?authorId=57250025600.

Коваленко Андрій Анатолійович – 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@nure.ua; ORCID ID: https://orcid.org/0000-0002-2817-9036;
https://www.scopus.com/authid/detail.uri?authorId=56423229200.

Метод аналізу аудіоподій у мережних системах аудіоаналітики

А. І. Порошенко, А. А. Коваленко

Анотація. **Актуальність.** У стрімко розвиваючійся сфері мережевих систем аудіоаналітики виявлення та аналіз аудіоподій відіграють вирішальну роль у різних галузях, включаючи безпеку, охорону здоров'я та розваги. **Предмет.** У цій статті розглядається метод розпізнавання аудіоподій у мережевих системах аудіоаналітики, включаючи попередню обробку, поділ звуків і створення моделей машинного навчання для аналізу аудіосигналів. **Мета.** Метою є розробка та вдосконалення інтегрованих методів аналізу аудіосигналів у мережевих системах аудіоаналітики для підвищення точності, швидкості та надійності аналізу даних. **Методи**. Запропонований підхід використовує модифіковану архітектуру ResNet для класифікації кількох аудіоподій і згорткову нейронну мережу для розділення звукових джерел у багатоканальних записах. **Результати.** Метод демонструє конкурентоспроможні результати, які порівнюються з базовими показниками в сучасних методах, представлених на DCASE, та показує стійку роботу в умовах шуму. **Висновки.** Запропонований метод має потенціал для підвищення точності та надійності розпізнавання аудіоподій у реальних сценаріях, особливо в умовах складних акустичних середовищ..

Ключові слова: аналіз аудіосигналу; передобробка; виявлення аудіоподій; розділення звуків; машинне навчання; коефіцієнт помилок.