CNN-based Segmentation and Classification of Sound Streams under realistic conditions

Eleni Tsalera
Department of Informatics and Computer Engineering
University of West Attica, Greece
etsalera@uniwa.gr

Andreas Papadakis*
Department of Electrical and Electronic Engineering
Educators, School of Pedagogical and Technological Education (ASPETE), Athens, Greece
apapadakis@aspete.gr

Maria Samarakou
Department of Informatics and Computer Engineering
University of West Attica, Greece
marsam@uniwa.gr

Ioannis Voyiatzis†
Department of Informatics and Computer Engineering
University of West Attica, Greece
voyageri@uniwa.gr

ABSTRACT

Audio datasets support the training and validation of Machine Learning algorithms in audio classification problems. Such datasets include different, arbitrarily chosen audio classes. We initially investigate a unifying approach, based on the mapping of audio classes according to the Audioset ontology. Using the ESC-10 audio dataset, a tree-like representation of its classes is created. In addition, we employ an audio similarity calculation tool based on the values of extracted features (spectrum centroid, the spectrum flux and the spectral roll-off). This way the audio classes are connected both semantically and in feature-based manner. Employing the same dataset, ESC-10, we perform sound classification using CNN-based algorithms, after transforming the sound excerpts into images (based on their Mel spectrograms). The YAMNet and VGGish networks are used for audio classification and the accuracy reaches 90%. We extend the classification algorithm with segmentation logic, so that it can be applied into more complex sound excerpts, where multiple sound types are included in a sequential and/or overlapping manner. Quantitative metrics are defined on the behavior of the combined segmentation and segmentation functionality, including two key parameters for the merging operation, the minimum duration of the identified sounds and the intervals. The qualitative metrics are related to the number of sound identification events for a concatenated sound excerpt of the dataset and per each sound class. This way the segmentation logic can operate in a fine- and coarse-grained manner while the dataset and the individual sound classes are characterized in terms of clearness and distinguishability.

*Corresponding Author
†Corresponding Author

PCI 2022, November 25–27, 2022, Athens, Greece

ACM ISBN 978-1-4503-9854-1/22/11…$15.00
https://doi.org/10.1145/3575879.3576020

CCS CONCEPTS

• **Machine learning approaches**; • **Convolutional Neural Networks**; • **Classification and regression trees**;

KEYWORDS

Sound classification, Segmentation, Mel spectrogram, YAMNet, VGGish

ACM Reference Format:
Eleni Tsalera, Andreas Papadakis, Maria Samarakou, and Ioannis Voyiatzis. 2022. CNN-based Segmentation and Classification of Sound Streams under realistic conditions. In *26th Pan-Hellenic Conference on Informatics (PCI 2022), November 25–27, 2022, Athens, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3575879.3576020

1 INTRODUCTION

Machine learning (ML) techniques are increasingly applied in sound classification problems as an alternative to the *traditional* sound classification techniques, where sound features are being extracted and compared to separate different classes of sounds. Deep learning (DL) techniques are achieving high classification accuracy, including specific Convolutional Neural Networks (CNNs) which have been designed especially for sound classification, such as YAMNet, VGGish [1], OpenL3 [2] and Crepe [3]. In parallel, audio datasets are being created, consisting of sounds of different types (from mainstream or more specialized contexts such the operation of machines), used to train and validate the ML algorithms (e.g. Environmental Sound Classification (ESC) [4]). A challenging point regarding the usage of the audio datasets is due to the lack of an association mechanism among sound classes included in the same or different datasets. The association is related to the contextual (e.g. origin) and technological (e.g. acoustic features) similarity of the classes and the labelling of the sound, which typically takes place manually. Labelling can also have different level of fine-graining e.g. from the more generic *human-sounds* to the more specific *sneezing*. Another difference among audio datasets is related to the quality of the sounds (in some cases a sound type can be clear while in others sound types may be mixed). One further aspect is that under realistic conditions of sound classification, sounds appear in a sequential (or even intermixed) fashion calling for robust segmentation techniques. These factors create a segmented approach in the audio dataset landscape.

1.1 Objectives

In this work, we investigate two types of systematic associations among sound classes: a) the 'semantic' (considering the origin, e.g. sounds coming from humans), and b) the comparison based on audio features. Regarding (a), sound classes can be *semantically* connected, considering a unifying graph ontology. One of the first and most systematic approach has been performed by [5], with the Audioset ontology. The Audioset ontology includes 632 classes connected in a tree – like structure (ontology), with the association among sound types being based on the origin of the sound. This can allow for high-level mapping of classes belonging to different datasets, especially for identical and similar classes. Regarding (b), sound class association can be based on the extraction of audio features and calculation of the (Euclidean or other) distance among them. Features can be extracted and weighted according to existing methodologies [6]. Such similarity metrics (distances) may not necessarily coincide with the semantic affinity of the sounds.

Sound coming from realistic settings may include multiple types, combined in a sequential and/or overlapping manner. In such cases, sound classification and identification should also consider the segmentation of the sounds and management of the identified sounds. This involves decision on a) the separation or merging of the same sounds excerpts depending on their temporary adjacency and b) to threshold-related decisions, such as the minimum duration so that an identified sound excerpt is considered. This area is effectively researched in dynamic analysis and segmentation techniques of signal timeseries but it presents challenges in generic (i.e. without predetermined dialogues) audio 'streaming' scenarios.

Considering the above, this research has the following objectives:

- Investigate association among sound classes, considering both the semantic and the technological perspectives.
- Evaluate sound classification using state-of-the-art CNN tools and contribute to the mapping of sound types, belonging to the same or different audio datasets, in a quantitative manner, considering the semantic affinity and the feature-based similarity.
- Extend the sound classification scenarios from independent, discrete sound classes towards complex sound streams of larger duration, which include multiple sound types and design a sound segmentation and classification algorithm in such settings.

1.2 Similar work

Sound classification is a challenging research field given the complexity and the dynamic nature of the signal and the simultaneous presence of sounds from different sources either outdoor (environmental or urban) [7, 8], or indoor (business, residential, educational) [9]. Sound classification techniques are increasingly oriented towards Machine Learning mechanisms [10] and Deep Learning [11]. In the first case, features of the sound (from the time, frequency or perceptual domain) are extracted, evaluated in terms of their descriptive power (using methodologies such as Relief-F [12] or Principal Component Analysis (PCA) [13, 14]) and feed classification algorithms. In DL the process takes place internally in the CNNs which are gradually trained from layer to layer with the last one to perform the classification [15]. Transfer learning is a widely used classification method where CNNs that have been already trained on a large set of either images (i.e. ImageNet[1]) or sounds (i.e. AudioSet) are retrained on the dataset under consideration. Retraining of CNNs on specific datasets typically has benefits in computational resources and classification accuracy [16].

Audio signal segmentation is a subset of the signal segmentation for non-deterministic signals. Signal segmentation is based on the identifying and processing changes in signal frequency and amplitudes, while other techniques have been applied, e.g. [17] has employed discrete wavelet transform (DWT) to decompose signals into orthonormal time series with different frequency bands (e.g., in EEG signals). In the case of sound, the Bayesian Information Criterion has been used for speech segmentation [18]. In our work we perform sound classification based on deep learning using the retrained YAMNet and VGGish networks, to achieve high classification accuracy upon the ESC-10 dataset. In addition, we investigate segmentation of complex sound streams (consisting not only of speech but also of arbitrary sound types), leveraging the classification performed upon the fine-grained sound frames.

The structure of the document is the following: Section 2 describes the methodology, the selection of the audio dataset and the ML algorithms for sound classification scenarios. Section 3 discusses the results of a set of sound segmentation and classification scenarios using different parameter values. Section 4 includes the conclusions of the work.

2 METHODOLOGY

2.1 Sound type mapping and similarity

Audio datasets include different types of sounds, arbitrarily chosen. Semantic association among these types can be achieved considering the tree-like hierarchy based on the audio set ontology. This is verified with the ESC-10 audio dataset which includes 10 classes [dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster and fire crackling]. Mapping is straightforward as these classes can be associated with [bark, rain, waves-surf, baby cry-infant cry, tick, sneeze, helicopter, chainsaw, chicken-rooster, crackle], which are depicted in Figure 1. For example, tick, helicopter and chainsaw belong to the sound of things category (2^{nd} level), while rain and surf belong to water category (3^{rd} level category).

Another criterion to associate sound classes is the calculation of the feature-based similarity. Audio is characterized by temporal, spectral and perceptual features. Extracting a subset of (representative) features and calculating the Euclidean distance among their values can provide for a quantitative association of the 'technical' similarity. We select the spectrum centroid, the spectrum flux and the spectral roll-off based on the work performed by [14]. For each ESC-10 class, a representative file consisting of 20 concatenated excerpts is created, and pre-processed (converted to monophonic in case the original samples are stereo, with a single sampling rate of 44100 Hz and quantization scheme). These files are processed to extract the audio features using a Hanning window of 3 ms and overlapping length of 2 ms.

[1] https://www.image-net.org/

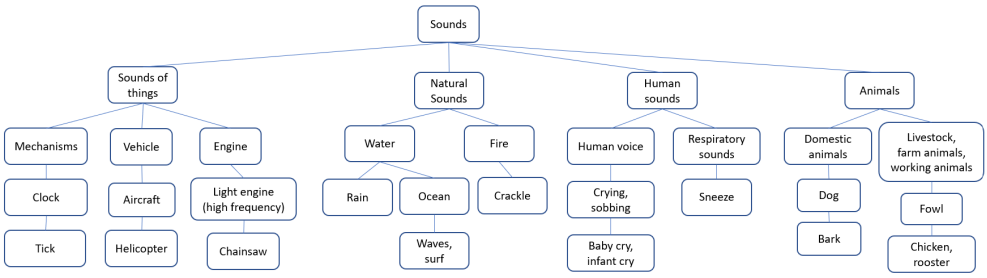

Figure 1: Subset of audio set ontology mapping the ESC-10 classes

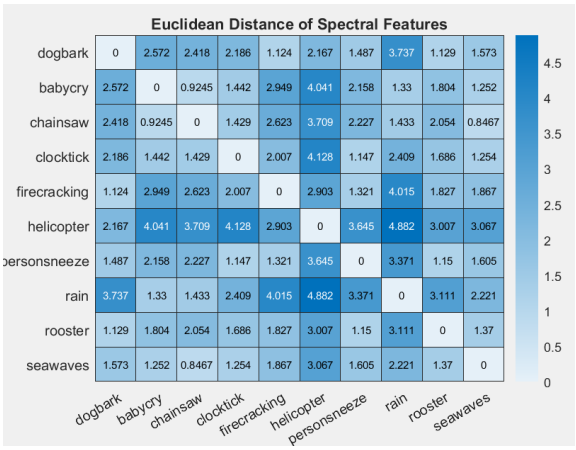

Figure 2: Sound type similarities

After calculating the mean of the moving median for each feature and normalizing, the Euclidean distance of the feature values is calculated. The results are depicted in Figure 2 (as heatmap) and provide a quantitative indication of the technical similarity among classes. The distance between dog bark and rooster is closer than rain and helicopter.

2.2 Sound classification and segmentation

Sound type identification under realistic conditions involves audio streams consisting of multiple, sequential sounds, while in some cases sounds can be overlapping. This creates the need for a flexible and robust segmentation mechanism, a) indicating a sound type only when the (identification) confidence level exceeds a threshold, and b) identifying the (temporal) boundaries of the sounds. Furthermore, considering that sounds of very limited duration (e.g. dozens of milliseconds) may have limited value for practical applications, the mechanism should handle the minimal duration of identified sounds. From another point of view, sounds of the same type are typically separated by short periods of noise or silence, e.g. there are blank periods between sequential dog barks. So, it may be more meaningful to consider as a single continuous period rather than as multiple discrete periods of the same sound type. This way adequately adjacent sounds of the same type can be merged. To quantitatively approach these operations we have considered a set of parameters, as in Table 1.

The process applied (as depicted in Figure 3) is the following: The sound files go through preprocessing for homogenizing the sampling frequency (e.g. to 16 or 44.1 KHz) and quantization level. Each sound excerpt (of length FILELEN) is split into frames of fixed length (WINLEN) with overlapping and hop length equal to HOPLEN. The window and overlap lengths can be related to (the reverse of) the sampling rate and with each other (e.g., the hop can be 12.5% to 50% of the window length). The number of frames is approached as: $N = (FILELEN - WINLEN)/(HOPLEN) + 1$. For each frame, the Mel spectrogram is calculated using the STFT (Short Time Fourier Transform), resulting in N figures. The N Mel spectrograms represent the sound file as a set of figures and for each the type of sound is estimated using the CNN. As these estimations may include multiple (overlapping) sound types, the prevailing identification is selected so that there is a single identification per hop.

Based on the set of identified sounds and temporal boundaries, the results are further filtered considering the identification confidence level (comparison with the IDCONF threshold). The sounds to be included and/or excluded are considered. The minimum duration length (MINDUR) is applied after being translated to hops (through its division with the HOPLEN) and removes sound identifications across durations smaller than the minimum accepted. In parallel, if the interval between two identical identifications is less or equal to MINSEP (which is translated to hop number through its division with the HOPLEN), the identification sound excerpts are merged.

3 RESULTS

3.1 Sound classification based on CNNs

Two Sound CNNs, VGGish and YAMNet, have been employed. The dataset was split into the training set, the validation set and the test set by 60%, 20% and 20% respectively. The retraining of these networks was carried out by selecting the values of the hyperparameters according to [19] where evaluation of multiple number combinations has been performed. The optimizer is set to Adaptive Moment Estimation (Adam) [20], the mini batch size to 32, the learning rate to 0.5×10^{-4} and the maximum number of epochs to 10, for both CNNs. The classification accuracy and erroneous identifications per class are depicted in the confusion matrices of Figure 4 for VGGish and YAMNet.

Table 1: Parameters for sound identification and segmentation

Parameter	Description	Value
Minimum sound duration (MINDUR)	MINDUR is the minimum duration of sound region, so that it is recognized	0.5 (sec)
Identification confidence (IDCONF)	IDCONF is used to include, or filter identified sounds	[0, 1]
Minimum temporal distance (MINSEP)	Minimum separation between sequential regions of the same detected sound	0.25 (sec)
Included sound types	Subset of the Audioset ontology types that are included	Sound type subset
Excluded sound types	Subset of the Audioset ontology types that are excluded	Sound type subset
Specificity level	The depth of the identified sound in the structured set	0, 1, 2
Sound window length (WINLEN)	The temporal length of the window, upon which the algorithm is applied	0.5 to 1.5 (sec)
Hop length (HOPLEN)	The hop length that allows the windowing of each sound excerpt	80 to 250 (ms)

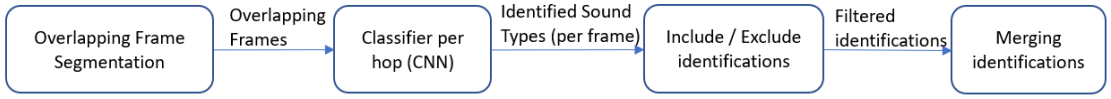

Figure 3: Sound type identification and segmentation

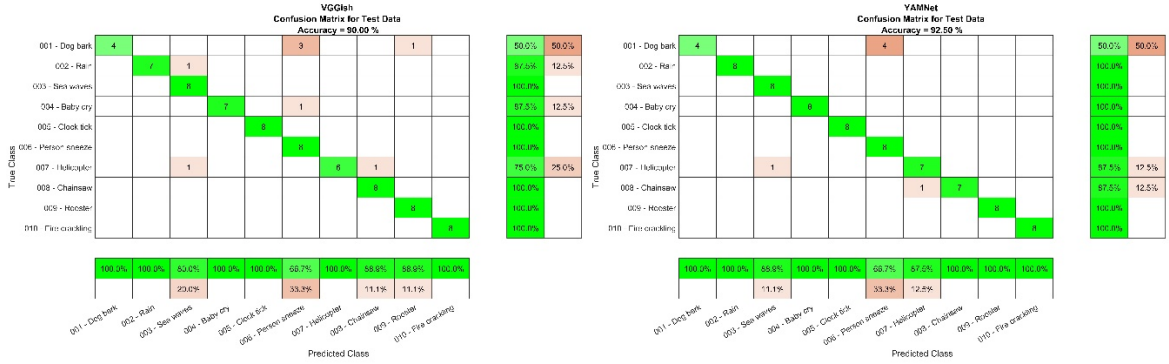

Figure 4: Confusion matrix for VGGish and YAMNet (applied upon ESC-10, after transfer learning)

3.2 Fine- and coarse-grained segmentation scenarios

To verify the behavior of the classification and segmentation algorithm, we consider values for the two key parameters (related to merging and segmentation), i.e., the *Minimum Sound Duration* (MINDUR) and the *Minimum Sound Separation* (MINSEP) from range {0, 0.2, 0.4, 0.6, 0.8, 1} and {0, 0.1, 0.2, 0.3, 0.4, 0.5} respectively. The classification algorithm is run for each of the 36 (6x6) combinations, upon a sound concatenation representative of the dataset (including 2 excerpts per each class). Figure 5 indicates the number of sound identifications for each of the combinations (per discrete minimum duration values in x axis and minimum separation in y axis). The radius of the circles (bubbles) is proportional to the number of sound identifications. When these parameters take their minimum value (MINDUR=MINSEP=0), the sound segmentation takes its more fine-grained form (the number of sound identifications is 32), while for their maximum value, the segmentation process becomes more coarse-grained and the corresponding sound identification number converges to the number of the concatenated discrete sound excerpts (24 sounds). This behavior can be indicative of the *clearness* of the dataset (i.e. whether each sound excerpt consists of a single sound or multiple ones sequential or overlapping).

The scenario can be extended, if we consider sound excerpts representative of each sound type (within the same or different dataset). For the case of ESC-10, we repeat the experiment for each of the classes (after concatenating 20 audio files belonging to the same class). The statistical results, in terms of the number of sound identifications are presented in Table 2 (2nd column indicates the minimum number of sound identifications, 3rd column the maximum, 4th the average and the 5th the standard deviation for the values of MINDUR and MINSEP).

Table 2: Minimum, maximum, average and standard deviation number of sounds for each class

Class	Min	Max	Average	Standard Deviation
Dog bark	14	18	16.8	1.1
Rain	10	22	17.7	3.8
Sea waves	16	23	20.8	1.9
Baby cry	13	33	22.5	6.1
Clock tick	5	13	9	2.4
Person sneeze	27	46	36	6.7
Helicopter	10	16	13.1	1.6
Chainsaw	17	33	27.2	5.4
Rooster	35	37	36.3	0.9
Fire crackling	13	28	22.6	4.9

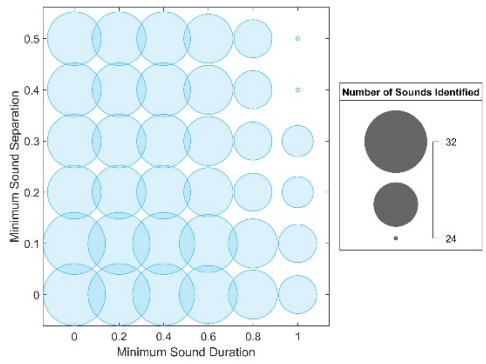

Figure 5: Sound segmentation and identification results for different values of minimum sound duration and separation

Considering that for each class the number of discrete appearances of this sound type is 20, and that all classes are efficiently recognized by the classification algorithm, the minimum number of sound identifications can be an indication of the 'recognizability' – 'distinguishability' of the class. According to Table 2 (column average), rooster and person sneezing are the most recognizable, while clock ticking and helicopter the least ones.

4 CONCLUSIONS

Audio datasets support machine learning frameworks, through algorithm training, and validation. Such datasets can be heterogeneous in terms of technical properties (sampling rate, quantization) and more importantly in terms of sound classes they include. Mapping of similar sound classes belonging to different datasets can be achieved with the assistance (bridging) of the audio ontology (as verified with ESC-10). This mapping allows for a semantic association of sound types (i.e. belonging to the same sub-category). In addition, sound type similarity and comparison can be performed using the value of features extracted from sound excerpts.

In parallel, we have evaluated sound classification techniques using sound CNNs (YAMNet and VGGish) achieving high classification accuracy (90%) in the ESC-10 dataset. The classification algorithm has been extended from discrete sound excerpts of single sound type files towards more complex and realistic sounds of larger duration, which include multiple sound types and need segmentation. A set of operational parameters (minimum sound duration and sound intervals) has been defined which specify the segmentation process. The application of the segmentation algorithm with different values of the operational parameters has verified the possibility of customization from coarse-grained to fine-grained sound segmentation and classification, i.e. resulting in smaller and larger number of sound identifications.

In terms of future work, we will be elaborating the segmentation scenarios with a broader set of sounds as well as streams consisting of overlapping audio types. We also plan to apply these techniques in video footage, consisting of sequential images, where different sentiment / gaze expression takes place [21]. Such an analysis framework can provide a holistic understanding of the smart context events based upon the two most important signals (audio and video).

ACKNOWLEDGMENTS

The authors acknowledge financial support for this work from the University of West Attica.

REFERENCES

[1] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurus, Bryan Seybold, Malcom Slaney, Ron J. Weiss and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In Proceedings of IEEE international conference on acoustics, speech and signal processing (icassp), 131-135. https://doi.org/10.1109/ICASSP.2017.7952132

[2] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In Proceedings of IEEE international conference on acoustics, speech and signal processing (icassp), 3852-3856. https://doi.org/10.1109/ICASSP.2019.8682475

[3] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A Convolutional Representation for Pitch Estimation. In Proceedings of IEEE international conference on acoustics, speech and signal processing (icassp), 161-165. https://doi.org/10.1109/ICASSP.2018.8461329

[4] Karol J. Piczak. 2015. ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 1015–1018. https://doi.org/10.1145/2733373.2806390

[5] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal and Marvin Ritter. 2017. In Proceedings of IEEE international conference on acoustics, speech and signal processing (icassp), 776-780. https://doi.org/10.1109/ICASSP.2017.7952261

[6] Carlos Oscar S. Sorzano, Juan Vargas and A. Pascual Montano. 2014. A survey of dimensionality reduction techniques. arXiv preprint. https://doi.org/10.48550/arXiv.1403.2877

[7] Juan Pablo Bello, Charlie Mydlarz and Justin Salamon. 2018. Sound analysis in smart cities. In Computational Analysis of Sound Scenes and Events; Springer: Cham, Switzerland, 373–397. https://doi.org/10.1007/978-3-319-63450-0_13

[8] Helin Wang, Yuexian Zou, Dading Chong, and Wenwu Wang. 2019. Environmental Sound Classification with Parallel Temporal-Spectral Attention. arXiv preprint. https://doi.org/10.48550/arXiv.1912.06808

[9] Ivan M. Pires, Goncalo Marques, Nuno M. Garcia, Francisco Flórez-Revuelta, Maria Canavarro Teixeira, Eftim Zdravevski, and Susanna Spinsante. 2021. Recognition of Activities of Daily Living Based on a Mobile Data Source Framework. In Bio-inspired Neurocomputing, Springer, Singapore, 321-335. https://doi.org/10.1007/978-981-15-5495-7_18

[10] Pengcheng Wei, Fangcheng He, Li Li and Jing Li. 2020. Research on sound classification based on SVM. Neural Computing and Applications, 32(6), 1593-1607. https://doi.org/10.1007/s00521-019-04182-0

[11] Roneel V. Sharan, Hao Xiong, and Shlomo Berkovsky. 2021. Benchmarking Audio Signal Representation Techniques for Classification with Convolutional Neural Networks. Sensors, 21, 3434, 20-23. https://doi.org/10.3390/s21103434

[12] Igor Kononenko, Edvard Simec and Marko Robnik-Sikonja. 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. https://doi.org/10.1023/A:1008280620621

[13] Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417. https://doi.org/10.1037/h0071325

[14] Eleni Tsalera, Andreas Papadakis, and Maria Samarakou. 2021. Novel principal component analysis-based feature selection mechanism for classroom sound classification. Computational Intelligence, 37(4), 1827-1843. https://doi.org/10.1111/coin.12468

[15] Fatih Demir, Daban Abdulsalam Abdullah, and Abdulkarir Sengur. 2020. A new deep CNN model for environmental sound classification. IEEE Access, 8, 66529-66537. https://doi.org/10.1109/ACCESS.2020.2984903

[16] Fuzhen Zhuang, Zhiyan Qi, Keyu Duan, Dongo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong and Qing He. 2020. A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555

[17] Hamed Azami, Hamid Hassanpour, Javier Escudero, and Saeid Sanei. 2015. An intelligent approach for variable size segmentation of non-stationary signals. Journal of advanced research, 6(5), 687-698. https://doi.org/10.1016/j.jare.2014.03.004

[18] George Almpanidis and Constantine Kotropoulos. 2008. Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion. Speech Communication, 50(1), 38-55. https://doi.org/10.1016/j.specom.2007.06.005

[19] Eleni Tsalera, Andreas Papadakis and Maria Samarakou. 2021. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. Journal of Sensor and Actuator Networks, 10(4), 72. https://doi.org/10.3390/jsan10040072

[20] Nitish Shirish Keskar and Richard Socher. 2017. Improving Generalization Performance by Switching from Adam to SGD. arXiv e-prints. https://doi.org/10.48550/arXiv.1712.07628

[21] Eleni Tsalera, Andreas Papadakis, Maria Samarakou and Ioannis Voyiatzis. 2022. Feature Extraction with Handcrafted Methods and Convolutional Neural Networks for Facial Emotion Recognition. Applied Sciences, 12(17), 8455. https://doi.org/10.3390/app12178455