SOUND EVENT CLASSIFICATION

Chiara Auriemma, Francesca Benesso, Anna Fusari, Filippo Marri

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[chiara.auriemma,francesca1.benesso]@mail.polimi.it
[anna.fusari,filippo.marri]@mail.polimi.it

ABSTRACT

Sound Event Classification (SED) has become an important task in the field of audio processing, with applications ranging from environmental monitoring to human-computer interaction. Aim of this project is to develop a sound event classification system based on a Convolutional Neural Network (CNN) architecture training it on the ESC-50 dataset. At the end, the performances of the model are compared with the ones of a state-of-the-art model (QUALE?). [Da finire, deve essere una sorta di riassunto del progetto, con le tecniche utilizzate e i risultati ottenuti].

***Index Terms*—** Sound Event Classification, Convolutional Neural Network, ESC-50 dataset, performance limitations

1. INTRODUCTION

Sound Event Classification (SED) is a task that involves the identification and classification of specific sound events within an audio signal. This task has gained significant attention in recent years due to its wide range of applications, including environmental monitoring[1], human-computer interaction[2], and multimedia content analysis[3]. The goal of SED is to accurately detect and classify sound events in real-time or from pre-recorded audio data. The process of SED typically involves several steps, including feature extraction, model training, and evaluation[4]. Commonly used features include Mel-frequency cepstral coefficients (MFCCs), spectrograms, and log-mel spectrograms. Model training involves using labeled audio data to train a machine learning model to recognize and classify sound events. Various machine learning algorithms can be used for SED, including support vector machines (SVMs), decision trees, and deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs)[5]. The choice of algorithm depends on the complexity of the task and the available data. Evaluation of the SED system is typically done using standard metrics such as accuracy, precision, recall, and F1-score.

2. METHODOLOGY

3. EVALUATION

3.1. Dataset analysis and preprocessing

The dataset used for this project is the well-known ESC-50 dataset [6], which contains 2000 labeled sound events from 50 different classes, with each class containing 40 samples. Each song of the dataset is available in WAV format with a sample rate of 44.1 kHz and a bit depth of 16 bits.

According to the analysis that will be done on the results inspired by (CHIEDERE ARTICOLOOOO), the type of soound events in the dataset can be divided into three main categories:

- **Transient sounds**: This category includes sounds that have a short duration and are characterized by a sudden onset, such as a dog barking, a door slamming, or a gunshot.
- **Continous sounds**: This category includes sounds that have a longer duration and are characterized by a continuous or sustained sound, such as a car engine running, a train passing, or a river flowing.
- **Intermittent sounds**: This category includes sounds that have a periodic or irregular pattern, such as a clock ticking, a bird chirping, or a phone ringing.

According to what is reported in the paper in which the ESC-50 dataset is presented [6], we higlight how some sounds are more difficult to classify than others, such as the sounds of a washing machine, an helicopter, or an engine due to their similar spectrograms. This happens not only for machines, but for humans too. This will be taken into account in the results section, where we will see how the model performs on different classes of sounds.

It is also important to note that, even though we consider the same class, the variability of the sounds is very high, as we can see by comparing the spectrograms of three different samples of the *dog barking* class.

Furthermore, we underline how some of the ambiental sounds, like the one of the wind, have no univoque structure: by breaking down (phrasl verb...) their spectrograms in their harmonic and percussive components, it is evident that the difference it is not so clear since the two plots are almost equal.

This fast analysis of the dataset has been done to understand the limitations of the model and the difficulties that it could encounter during the training phase.

A part of the dataset, 198 elements (10% of the total), has been separated for testing. The remaining samples have been splitted according to the stratifiedkfold module of sklearn [7] in a training and a validation set.

Drawing inspiration by the Salamon and Bello paper [8], five different techniques have been implemented to process the training set:

- **Time Stretching (TS)**: the audio signal is stretched or compressed in time by a random factor within a specified range. A last boolean parameter crop the processed audio to the original length, so that the model can be trained on the same length of the original audio.
- **Pitch Shifting (PS)**: the audio signal is shifted in pitch by a

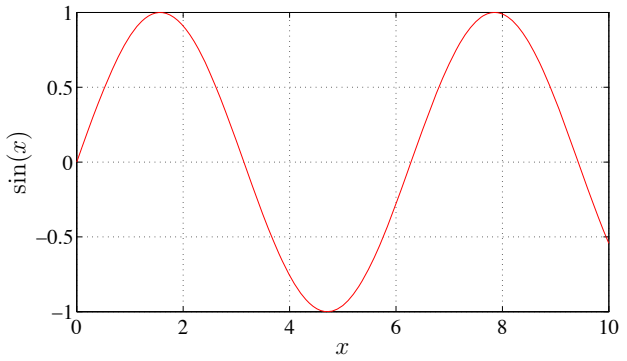

Figure 1: Example of a figure with experimental results.

random factor within a specified range.

- **Background Noise (BN)**: a Gaussian noise is added to the audio signal with a specified SNR range and activation probability.
- **Dynamic Range Compression (DRC)**: the dynamic range of the audio signal is compressed from a certain threshold with a specified ratio, attack time, and release time.
- **Convolution with Impulse Responses (CIR)**: the audio signal is convolved with the *MIT Acoustical Reverberation Scene Statistics Survey* dataset of impulse responses[9] to simulate different acoustic environments. This time again, an activation probability is used to increase variability.

For all the results presented in this paper, the training dataset has been preprocessed using the following parameters:

- TS: factor between 0.8 and 1.25 with an activation probability of 1
- PS: factor between -5 and 5 semitones and an activation probability of 1
- BN: SNR between 5 and 40 dB, activation probability of 1
- DRC: threshold of -20 dB, ratio of 4:1, attack time of 10 ms, release time of 100 ms
- CIR: activation probability of 1

3.2. Evaluation metrics

The evaluation of the sound event classification system is performed using several metrics to assess its performance. Firstly, the test accuracy, the reports and the confusion matrix are computed to evaluate the overall performance of the model.

4. RESULTS

Ci vanno messi tutti i risultati du quello che abbiamo fatto sul dataset: possiamo così giustificare gli errori di alcune classificazioni. *La luna vide dal cielo*
Rosita baciar Manuelo
Con tanto languor, con tanto ardor
Che s'ammantò d'un velo

Since there are many ways, often incompatible, of including images (e.g., with experimental results) in a LaTeX document, an example of how to do this is presented in Fig. 1.

5. ACKNOWLEDGMENT

This work was supported by the Politecnico di Milano, within the framework of the Selected Topics in Music and Acoustic Engineering Course 2025. A special thanks goes to the course instructor, Prof. JULIO JOSÉ CARABIAS ORTI, for being one of the brightest stars in the sky of artificial intelligence. Last but not least, we would like to thank our wallet: without those 24 euros, we would not have been able to run anything. Thank you for your support, we are grateful for your generosity. Grazie a tutt coloro che ci hanno supportato e sopportato.

6. REFERENCES

[1] R. Narasimhan, X. Z. Fern, and R. Raich, "Simultaneous segmentation and classification of bird song using cnn," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 146–150.

[2] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021. [Online]. Available: https://doi.org/10.1007/s00779-020-01389-0

[3] A. Kumar and B. Raj, "Weakly supervised scalable audio content analysis," 2016. [Online]. Available: https://arxiv.org/abs/1606.03664

[4] S. Padmaja and N. Sharmila Banu, "A systematic literature review on sound event detection and classification," in *2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM)*, 2025, pp. 1580–1587.

[5] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent Systems with Applications*, vol. 16, p. 200115, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667305322000539

[6] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM, 2015, pp. 1015–1018.

[7] S. learn developers, "Stratifiedkfold — scikit-learn 1.5.2 documentation," 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[9] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1612524113