DOI: 10.61782/fa.2025.0340

FORUM ACUSTICUM EURONOISE 2025

ENHANCED ENVIRONMENTAL SOUND EVENT CLASSIFICATION THROUGH TRANSFER LEARNING WITH CLAP MODEL

Francesco Artuso[1*] **Geremia Pompei**[2] **Andac Akbaba**[3]
Gino Iannace[3] **Francesco D'Alessandro**[4] **Francesco Fidecaro**[1]
Gaetano Licitra[2] **Luca Fredianelli**[2]
[1] Department of Physics, University of Pisa, Italy
[2] Institute for Chemical-Physical Processes of the Italian Research Council, Pisa, Italy
[3] Department of Architecture and Industrial Design, University of Campania, Aversa, Italy
[4] iPOOL S.r.l., Via Antonio Cocchi 3, 56121 Pisa, Italy

ABSTRACT

Long-term noise monitoring is essential to ensure compliance with regulations. This process requires the removal of spurious sounds unrelated to the target source or to the typical soundscape of the monitored area. Traditionally, such tasks relied on manual labelling by operators, but recent advancements in data-driven methodologies highlight that it is time to automate the process using cutting-edge machine learning techniques. Pre-trained models, widely available in literature, are trained on extensive datasets covering numerous classes and serve as a foundation for developing specialized machine learning models fine-tuned for specific tasks or subsets of classes. This study presents a Transfer Learning approach to leverage the knowledge of the Contrastive Language-Audio Pre-training (CLAP) model for a classification task focused on a subset of its original classes. Although the CLAP model has demonstrated adaptability across a broad range of classes with good results, the findings of this study suggest that the application of Transfer Learning can enhance classification accuracy for the selected subset of environmental sound classes.

Keywords: *environmental noise, sound event classification, pre-trained models, transfer learning, clap*

**Corresponding author*: *francesco.artuso@phd.unipi.it.*
***Copyright**: ©2025 Artuso et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

1. INTRODUCTION

The evaluation of noise sources is essential for monitoring and mitigation of environmental noise levels in order to protect public health [1]. The primary contributors to environmental noise typically include road traffic [2–4], railways [5,6], airports [7], and industrial activities [8]. In real monitoring scenarios, however, in addition to the noise originating from the target source, a substantial number of unintended or anomalous events, referred to as spurious sounds, is also collected. In this context, accurately attributing the correct amount of responsibility to each source is crucial. Consequently, the identification and subsequent exclusion of spurious events from measurements becomes a key task. At present, this process is generally performed by acoustic experts through visual inspection of time histories and spectral representations. While effective in some cases, this method is highly time-consuming and susceptible to human error. These limitations are particularly evident in the case of wind turbine noise, where long-term measurements are necessary. To address these challenges, it can be useful to adopt new data-driven methodologies, capable of efficiently processing large volumes of data. In this regard, machine learning and deep learning offer promising potential to enhance and accelerate the detection of spurious events. The specific name of the task just described is Sound Event Detection (SED), which aims not only to identify the class of the event but also to determine its onset and offset within a longer audio recording. In this article, we address a preliminary step to this task, commonly referred to as Sound Event Classification (SEC), where audio segments are pre-

trimmed to contain only the sound event to be classified. This classification step serves as a foundation for building a model capable of accurately recognizing sound classes, which can then be integrated into a more complex pipeline to perform event detection.

One of the main difficulties related to this task lies in the highly heterogeneous nature of these anomalous occurrences, which makes their identification very complex. The principal categories of identifiable sources include anthropogenic sounds, transportation infrastructure noise, natural and animal-generated sounds, agricultural machinery, and industrial operations. Nevertheless, even this macro-level classification does not fully capture the variability found within individual categories, highlighting the need for adaptable classification approaches. The traditional strategy for addressing these tasks involves designing a neural network architecture and training it from scratch to differentiate among a fixed set of predefined target classes. While this method is effective in scenarios with a limited scope, it lacks generalization capability. Specifically, if the initial set of target classes were to be extended, the model would require complete retraining to accommodate the new classes. Moreover, in order to achieve good classification performance, especially on heterogeneous and complex datasets such as in this case, it is necessary to train the neural network on a very large amount of data. This ensures that the model can learn a comprehensive representation of the wide variability in data and scenarios it will be required to recognize. To address the need for a solution that is both adaptable to diverse contexts and trained on sufficiently varied data, the use of pre-trained models has recently become widespread. Pre-trained models already possess knowledge of the general features typical of the task they were designed for, such as image recognition, natural language processing, or audio classification. They can be used either directly or as a starting point to develop specialized models tailored to a specific set of classes or more targeted tasks. In this context, the Contrastive Language-Audio Pretraining (CLAP) [9, 10] model has been developed. It is a deep learning model designed to learn joint representations of audio and text. In order to achieve this, CLAP transforms both an audio measurements and its corresponding textual description into vectors of the same dimension, called embeddings. During training, the model pulls the embeddings of matching audio-text pairs closer together, while pushing apart those of unrelated pairs. This contrastive learning enables the model to capture semantic relationships, organizing the embedding space so that similar sounds and descriptions are near each other. To reinforce the selection of this model, it is worth mentioning that in [11] it has been demonstrated that, on a subset of five classes extracted from the ESC-50 dataset [12], the performances achieved using embeddings extracted by CLAP overcome the performances obtained using the classic features Mel-Frequency Cepstral Coefficients and Gammatone Frequency Cepstral Coefficients, both in clustering and classification task.

A more detailed description of the CLAP model will be provided in Section 2, along with an explanation of the transfer learning approach adopted in this study. Subsequently, in Section 3 the preliminary outcomes obtained with the fine-tuned model will be presented and compared with the results of the original CLAP model on the same dataset. Finally, Section 4 will briefly analyze these results and outline the main conclusions.

2. METHODOLOGY

2.1 CLAP model

The CLAP model consists of an audio encoder and a text encoder, designed to extract two embeddings of equal dimensionality d from the audio measurement and its corresponding textual description, respectively. The audio encoder is based on a Transformer architecture named Hierarchical Token-Semantic Audio Transformer (HTS-AT), introduced in [13]. The goal of CLAP is to compare these embeddings and optimize the representation space such that audio and text embeddings with similar semantic content are pulled closer together, while those with differing semantics are pushed farther apart. Analyzing the structure of the model in detail, the input of the audio encoder is the Mel-Spectrogram of the audio, that is the spectrogram computed using the Mel scale, designed to mimic human auditory perception, which is nonlinear. This scale converts frequencies so that equal distances on the scale correspond to equally perceived differences in sound, with greater sensitivity to lower frequencies and less to higher ones. The structure of CLAP model is depicted in Figures 1 and 2.

The Mel-Spectrogram of dimensions (T, F), where T is the temporal dimension and F is the number of frequency bins, is cut into different patches of dimensions $(P \times P)$, as it can be seen in Figure 1. Each element of the sequence is then passed to a module composed by a Convolutional Neural Network (CNN), to be transformed transformed in a so-called token of dimen-

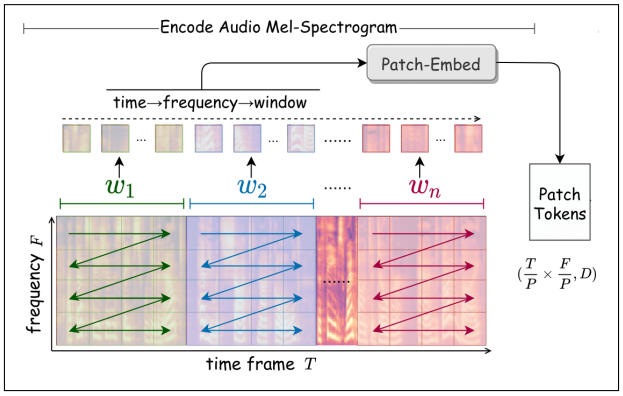

Figure 1. Mel-spectrogram encoding as input for HTS-AT. Architecture and Figure have been presented in [13].

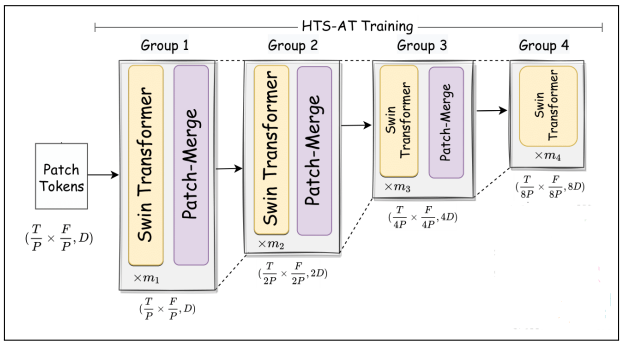

Figure 2. Sequence of SWIN Transformers and Patch-Merge in HTS-AT. Architecture and Figure have been presented in [13].

sions $(1, D)$. Then, all the tokens are arranged in a single structure of dimensions $\left(\frac{T}{P} \times \frac{F}{P}, D\right)$ where $\frac{T}{P} \times \frac{F}{P}$ is the number of tokens. The tokens are then fed into a sequence of transformer-encoder groups, as shown in Figure 2. The transformer architecture differs from previous models used in sequence analysis, like recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), as it doesn't rely on recurrent structures. Instead, it uses a self-attention mechanism to dynamically adjust the relevance of different tokens based on their contextual dependencies, allowing the model to focus on the most relevant parts of the sequence and capture complex relationships within the data. After the Transformer, the patch-merge module is applied to decrease the number of patches and increase the dimension in the latent space, combining adjacent patches. This procedure is repeated sequentially for other groups of transformer-merger, until the i^{th} audio is transformed into the embedding E_a^i of dimension $d = 1024$. By means of an encoder also every textual descriptions is transformed into a d-dimensional embedding E_t^i. Assuming to have N pairs of audio and text, it is possible to arrange the embeddings into two matrices E_a and $E_t \in R^{N \times d}$. The audio and text embeddings obtained through this procedure are then used to compute a similarity measurement, namely a scalar product between the vectors as an estimate of their distance in the hyperspace created by the model. The similarity matrix is equal to

$$\text{SIM} = \tau(E_t \cdot E_a^\top) \,, \tag{1}$$

where τ is a scaling factor. The matrix contains the similarity between N matching pairs in the diagonal and between $N^2 - N$ non-matching pairs off-diagonal. The goal of the training stage is to bring closer matching pairs and separate non-matching ones. The loss function is therefore designed so that, by minimizing it, the similarity between matching pairs (i.e., along the diagonal) is increased, while the similarity between non-matching pairs (i.e., off-diagonal elements) is decreased. The two loss functions ℓ_{text} and ℓ_{audio} are equal to

$$\ell = -\frac{1}{N} \sum_{i=0}^{N} \log \text{diag}(\text{softmax}(\text{SIM})) \,. \tag{2}$$

The terms of the matrix softmax(SIM) can be interpreted as the probability of matching between the corresponding audio-text pairs. The total loss is equal to:

$$\mathcal{L} = 0.5(\ell_{text} + \ell_{audio}) \,. \tag{3}$$

In the latent hyperspace of the model, minimizing this loss function results in bringing closer audio and text embeddings concerning the same semantic area.

2.2 Transfer learning

The CLAP model, once trained, generates a latent space where elements are arranged based on their semantic relationships. This structure allows the model to be used for classification tasks without additional training. To classify a new sound using the original model, a set of candidate class labels must be provided. Since the pre-training

objective was to align audio and text embeddings that refer to the same concept, the model performs classification by evaluating the similarity between these embeddings. Specifically, it computes the similarity score between the audio embedding and each of the class label embeddings, selecting the class with the highest score as the predicted output. Directly employing the pre-trained model for classification offers clear advantages, including its ability to recognize a vast range of classes. However, performance on an arbitrary set of classes is not guaranteed to be optimal. Since the goal of this work is not to ensure applicability to an unlimited set of categories, but rather to achieve high performance on a well-defined subset, a transfer learning strategy was adopted. The idea is to leverage the robust foundational knowledge of the CLAP model and fine-tune it to specialize in the selected subset of classes. The goal is to enhance accuracy within a limited set of sound classes, compared to the broader but less specialized performance of the original model. To achieve this, only the initial portion of the CLAP architecture, specifically the component responsible for embedding extraction, was retained. A new linear layer was then trained to project the resulting embeddings of dimension $d = 1024$ into a lower-dimensional space corresponding to the number of target classes of dimension $n_{\text{classes}} = 21$. Each output value from this layer represents the predicted likelihood that the input sound belongs to one of the predefined categories. The linear transformation applied by the layer to map the input x into the output y is equal to:

$$y = xW^{\top} + b \,, \tag{4}$$

where W and b are the learnable weights and bias of the layer, of dimension (n_{classes}, d) and (n_{classes}) respectively.

2.3 Dataset

The dataset was constructed using both measurements collected online, primarily from the website freesound.org [14], and those recorded by the authors. The classes were selected based on the authors' experience with monitoring measurements, primarily conducted in non-urban or sparsely populated areas with some human presence. The classes belong to five main categories: anthropogenic sounds, transportation infrastructure noise, natural and animal-generated sounds, agricultural machinery, and industrial operations. The chosen classes are 21, and they are reported in Table 1. It has been observed that varying the segment lengths within the same dataset during classification can result in different accuracy outcomes. Consequently, multiple versions of the dataset were created, with measurements segmented into durations ranging from 1 to 10 seconds in 1-second increments. For each dataset version, four distinct sets were constructed: the training set, the validation set, the early stopping set, and the test set. The early stopping set is essential to prevent overfitting, as training is stopped when the accuracy on the early stopping set does not improve for 10 consecutive epochs. Each dataset version consisted of 500, 100, 100, and 100 measurements for the training, validation, early stopping, and test sets, respectively. In cases where the total duration of the dataset for a given segment length was insufficient to meet these numbers, data augmentation was applied through the addition of gaussian noise.

Table 1. Classes of sounds chosen for the dataset.

Classes of Sounds	
Airplane	Bells
Birds	Cats
Chicken coop	Cicadas and crickets
Clacson	Crows and seagulls
Dogs	Glass breaking
Helicopter	Lawn mower and brush cutter
Music	Sirens and alarms
Thunder, fireworks and gunshot	Train
Vacuum cleaner, fan and hairdryer	Vehicle idling
Vehicle pass-by	Voices
Workshop	

3. RESULTS

In this Section the results obtained using directly the pre-trained model are compared with the results obtained with the fine-tuned model. The length of the measurement in seconds plays a crucial role in classification performance. It has been observed that varying the lengths of the segments from the same dataset during classification can lead to different accuracy results. This variation is likely due to the structural differences in the spectrograms. To accurately capture recurring patterns in a sound's spectrogram,

a minimum duration may be required. Hyperparameters were tested with lengths ranging from 1 to 10 seconds, in 1-second increments. Table 2 shows the accuracy and loss of the original pre-trained model on the validation set, depending on the length of the input audio. Various experiments were then conducted using the fine-tuned model, with different hyperparameters adjusted during training. The best result is reported in Table 3.

Table 2. Accuracy and Loss results of the original CLAP model on the validation set at varying of the audio length hyperparameter.

Audio length	Accuracy	Loss
4 s	72.93 %	2.83
7 s	72.89 %	2.82
8 s	70.55 %	2.83
3 s	70.23 %	2.82
9 s	69.49 %	2.83
10 s	69.35 %	2.83
5 s	68.32 %	2.83
6 s	66.81 %	2.83
2 s	64.02 %	2.84
1 s	60.42 %	2.86

Table 3. Accuracy and Loss result of the fine-tuned model.

Audio length	Accuracy	Loss
4 s	91.83 %	0.28

4. DISCUSSION AND CONCLUSIONS

This study presents the Contrastive Language-Audio Pretraining (CLAP) model, which integrates inputs from different domains into a shared space where their vector representations can be compared and manipulated. The model was fine-tuned using a Transfer Learning approach to perform Sound Event Classification (SEC) on a specific dataset, focusing on key environmental sources relevant to noise monitoring. A linear layer was applied to the sound embeddings generated by the original model to adapt it for the classification task. The results demonstrate a significant improvement in accuracy, increasing from 72.93 % (the best performance achieved by the original model) to 91.83 % (the best performance achieved by the fine-tuned model) on the validation set. These findings highlight the effectiveness of the fine-tuned model in the classification task, showing a substantial performance gain over the original model. Further analysis and exploration of additional fine-tuning strategies are warranted to continue improving these results. The next step in this research will involve incorporating the fine-tuned model into a detection pipeline for Sound Event Detection (SED), which is the ultimate objective in real-world scenarios.

5. ACKNOWLEDGMENTS

The paper was supported by the PRIN 2022 project 20223LMSZN, COMBINE "Sustainable condition monitoring of wind turbines using acoustic signals and machine learning techniques" and by the Ministry of University and Research (MUR) as part of the PON 2014-2020 "Research and Innovation" resources – Green/Innovation Action – DM MUR 1061/2022.

6. REFERENCES

[1] European Environment Agency., *Environmental noise in Europe, 2020.* LU: Publications Office, 2020.

[2] E. M. Andersson, M. Ögren, P. Molnár, D. Segersson, A. Rosengren, and L. Stockfelt, "Road traffic noise, air pollution and cardiovascular events in a Swedish cohort," *Environmental Research*, vol. 185, p. 109446, June 2020.

[3] L. Fredianelli, S. Carpita, M. Bernardini, L. G. Del Pizzo, F. Brocchi, F. Bianco, and G. Licitra, "Traffic Flow Detection Using Camera Images and Machine Learning Methods in ITS for Noise Map and Action Plan Optimization," *Sensors*, vol. 22, p. 1929, Mar. 2022.

[4] G. Licitra, M. Bernardini, R. Moreno, F. Bianco, and L. Fredianelli, "CNOSSOS-EU coefficients for electric vehicle noise emission," *Applied Acoustics*, vol. 211, p. 109511, Aug. 2023.

[5] F. Bunn and P. H. T. Zannin, "Assessment of railway noise in an urban setting," *Applied Acoustics*, vol. 104, pp. 16–23, Mar. 2016.

[6] G. Licitra, L. Fredianelli, D. Petri, and M. A. Vigotti, "Annoyance evaluation due to overall railway noise and vibration in Pisa urban areas," *Science of The Total Environment*, vol. 568, pp. 1315–1325, Oct. 2016.

[7] M. Espey and H. Lopez, "The Impact of Airport Noise and Proximity on Residential Property Values," *Growth and Change*, vol. 31, pp. 408–419, Jan. 2000.

[8] P. C. Eleftheriou, "Industrial noise and its effects on human hearing," *Applied Acoustics*, vol. 63, pp. 35–42, Jan. 2002.

[9] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes Island, Greece), pp. 1–5, IEEE, June 2023.

[10] B. Elizalde, S. Deshmukh, and H. Wang, "Natural Language Supervision for General-Purpose Audio Representations," Feb. 2024. arXiv:2309.05767 [cs, eess].

[11] F. Artuso, F. Fidecaro, F. D'Alessandro, G. Iannace, G. Licitra, G. Pompei, and L. Fredianelli, "Identifying optimal feature sets for acoustic signal classification in environmental noise measurements," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. INTER-NOISE24, (Nantes, France), pp. 7540–7549, Institute of Noise Control Engineering, 2024.

[12] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, (Brisbane Australia), pp. 1015–1018, ACM, Oct. 2015.

[13] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," Feb. 2022. arXiv:2202.00874 [cs, eess].

[14] Freesound Team, "Freesound." `https://freesound.org`, 2024.