DOI: 10.61782/fa.2025.0427

FORUM ACUSTICUM
EURONOISE 2025

HEARING PROTECTION DEVICE WITH AI-BASED SOUND DETECTION AND LOCALIZATION

Lucas Banchero[1*] **Jose Javier López**[1]
[1] Instituto de Telecomunicaciones y Aplicaciones Multimedia (ITEAM), Universitat Politècnica de València, Spain

ABSTRACT

Hearing protection is crucial in heavy industries to prevent auditory injuries from prolonged noise exposure. However, regulations limit noise isolation to ensure workers can hear critical sounds like alarms and warnings, creating a trade-off between protection and situational awareness. We propose an intelligent hearing protection system that combines passive noise isolation with AI-based sound processing. The device integrates external microphones and deep learning models to detect and localize important sounds, allowing workers to stay aware of their surroundings while receiving robust auditory protection. The system uses MEMS microphones on the earmuffs and headband to capture ambient sounds, processed by a central control board. The AI models, designed for emergency sound detection and localization, utilize Transformers and Convolutional architectures to accurately identify and locate critical sounds. Preliminary testing of the hardware prototype shows effective detection and localization of essential sounds in complex environments. As a first approach, this proposal focuses on sound capture, detection, and localization components. Many additional parts are needed for a fully functional system. This innovative approach offers a promising solution for improving workplace safety without compromising auditory health or regulatory compliance.

Keywords: *hearing protection, ai-based sound processing, noise isolation, sound localization, workplace safety*

[*] ***Corresponding author***: *lbanmar@upv.edu.es*

1. INTRODUCTION

In industrial environments and those with high noise exposure, hearing protection is a critical element for ensuring the safety and health of workers. Prolonged exposure to elevated noise levels can lead to noise-induced hearing loss, stress, fatigue, and other adverse health effects [1–3]. For this reason, occupational safety regulations mandate the use of hearing protection devices in certain work environments [4–6].

However, in some cases, the use of conventional hearing protectors can result in a phenomenon known as overprotection. This occurs when the sound insulation is so effective that the worker fails to perceive important sounds for their safety and job performance, such as emergency alarms, moving vehicle signals, or critical verbal instructions [7]. This situation creates a dilemma: protecting the worker from excessive noise without compromising their ability to react to relevant events [8,9].

In the industry, a common approach to mitigating this issue is to mandate the use of different types of hearing protectors depending on the specific noise conditions of each task. This requires companies to provide a range of protective devices tailored to varying levels and types of noise exposure. However, as demonstrated in [8], approximately 85% of workers experience auditory overprotection, which significantly impacts their ability to perceive critical sounds in their environment. This phenomenon leads to reduced situational awareness, increasing safety risks and hindering effective communication in noisy workplaces.

Another widely implemented solution is the integration of communication systems into hearing protection devices. These systems, typically incorporated into protective helmets or earmuffs, enable voice transmission and reception, helping workers maintain verbal communication despite high noise levels. However, while this approach mitigates the overprotection issue regarding speech

perception, it does not address the broader problem of detecting other critical environmental sounds, such as alarms, approaching vehicles, or machinery malfunctions. The selective enhancement of speech alone is insufficient to ensure comprehensive auditory situational awareness, particularly in complex industrial settings where multiple auditory cues contribute to safety and efficiency.

These limitations highlight the need for more adaptive and intelligent hearing protection solutions capable of distinguishing between harmful noise and essential auditory signals. Developing systems that dynamically adjust sound attenuation while preserving critical auditory information is essential for improving both worker safety and operational performance in high-noise environments.

In this context, the integration of artificial intelligence into hearing protection devices opens new possibilities. By utilizing signal processing algorithms and machine learning, it is possible to develop systems that differentiate between harmful noise and sounds of interest, enabling selective attenuation in real-time. This approach aims to provide effective protection without compromising the perception of essential acoustic signals for safety and communication in the workplace environment.

This paper presents an intelligent hearing protection system that employs artificial intelligence and advanced audio processing techniques to address the problem of overprotection. The following sections outline the principles of operation of the system as follows: Section 2 will introduce the proposed hardware model. Section 3 will describe the experiments conducted for data acquisition. Section 4 will focus on data processing and the artificial intelligence structures used for detection and localization. Section 5 will present the obtained results, and Section 6 will discuss potential future research directions for the project.

2. PROPOSED HARDWARE APPROACH

To mitigate the issue of hearing overprotection without compromising the worker's safety, a system is proposed that allows for the capture of external sounds through microphones placed on the outside of the protective earmuffs. In this way, the goal is to preserve the high passive attenuation of ambient noise while offering the user a selection of relevant sounds from the environment.

The proposed approach involves the use of strategically positioned microphones to capture ambient sound and send it to a processing system that would determine which signals should be transmitted to the user. In this sense, the structure of the earmuffs would maximize their passive isolation capacity, minimizing the entry of unwanted noise, while the microphones would provide an alternative audio input that is fully controlled.

Figure 1. MEMS microphone placed in the earmuff of the hearing protection headset.

As part of the experimental development, a prototype has been built based on this proposal using four MEMS MP34DT01-M microphones [10], strategically positioned on the headset: two in each earmuff, as shown in Figure 1 and two on the bridge, as shown in Figure 2.

The incorporation of microphones on the bridge of the headset provides significant advantages concerning sound spatial localization. By placing the microphones in an elevated position, the ability to localize sound in elevation is notably improved. This is because the microphones capture intensity (ILD) and time (ITD) differences between signals received from different elevations more accurately. In configurations where microphones are only placed in the earmuffs, information about the elevation of the sound is limited and more prone to errors, leading to perceptual ambiguities in this axis.

Additionally, the placement of microphones on the bridge helps eliminate the well-known "cone of confusion" and the front/back confusion [11] that occurs when microphones are only used in the earmuffs. This issue arises because two microphones placed on the sides of the head cannot accurately distinguish whether a sound source is located in front or behind the user. Under normal conditions, the human ear can resolve these ambiguities due to the small sound reflections in the folds of the ear [12], which create

unique patterns for each direction of the sound source. However, when using hearing protectors, the two microphones on the earmuffs are unable to discriminate their position.

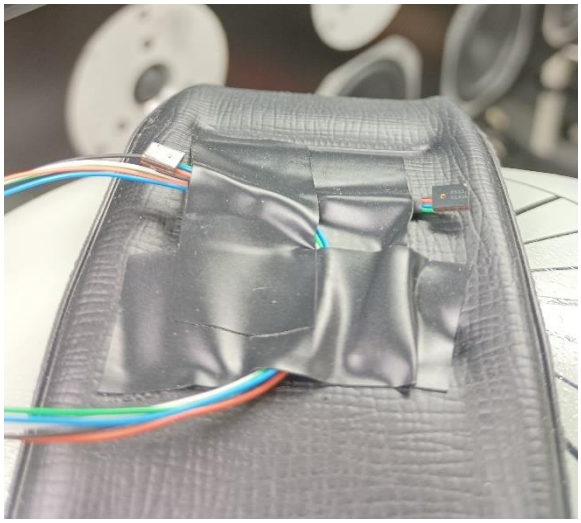

Figure 2. MEMS microphones placed on the headband of the hearing protection headset

The inclusion of microphones on the bridge allows for better differentiation of these signals, significantly improving directional localization capabilities.

The four microphones are connected to a central control board, the XMOS XVF-3000 [13], which enables synchronized signal acquisition. The control board features a USB output that sends the data directly to a PC for processing. This setup ensures that all microphones operate on the same clock, maintaining proper temporal alignment of the captured signals and enabling coherent analysis of the acoustic scene.

In terms of integration, the microphones and the control board are designed to capture and transmit ambient sound in a synchronized manner. The processing unit on the PC receives these signals in real-time, enabling their analysis and subsequent use according to the system's needs.

This approach would allow for the development of an adaptable and robust system, offering effective hearing protection without compromising the perception of essential acoustic events.

3. DATA ACQUISITION AND COLLECTION

The data collection process was carried out in a laboratory equipped with Wave-Field Synthesis (WFS) technology in an acoustically treated environment [14]. This space has been designed to minimize external interference and reflections from walls and ceilings, thus ensuring the accuracy of the measurements. Previous studies have validated and tested the effectiveness of this system in generating highly realistic sound fields, allowing the precise replication of real-world conditions [15].

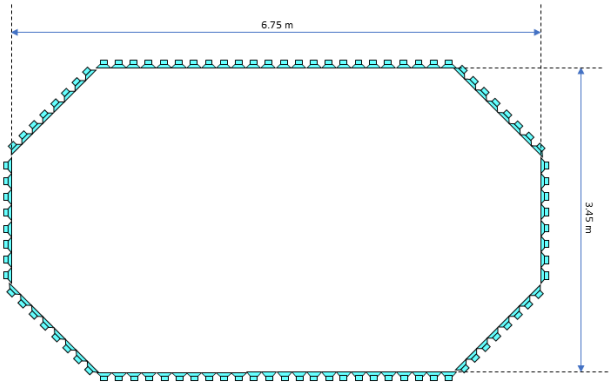

Figure 3. WFS system layout diagram

The sound generation system featured a matrix of 96 speakers arranged in an expanded octagonal configuration, as shown in Figure 3, enabling the creation of complex and controlled sound fields. This setup allowed for the faithful simulation of sound incidence from multiple directions, thereby evaluating the system's performance in various acoustic scenarios.

Since this research represents an initial approach to a functional prototype, the present study focused on the simulation and detection of emergency alerts, such as sirens and horns. To this end, emergency signals and background noise representative of industrial and urban environments were incorporated. The background noise was obtained from recordings made in real-world environments and played through planar waves, ensuring a homogeneous distribution of the acoustic field in all directions. This approach avoided unwanted localization effects and allowed for a realistic replication of a diffuse acoustic environment.

For the simulation of emergency signals, recordings of sirens and horns from two previously validated databases [16,17] were used. Additionally, to assess the system's ability to discriminate between signals of interest and environmental noise, point sources of industrial noise were included in specific locations within the test space. These

sources represented characteristic sounds from tools and machinery, such as chainsaws, running engines, handheld saws, pneumatic hammers, vacuums, and washing machines, among others, extracted from the UrbanSound8K [16] and ESC-50 [18] databases. The inclusion of these point noise sources in the simulated environment allowed for the creation of complex acoustic conditions, with the aim of challenging the artificial intelligence algorithms to differentiate between critical alerts and non-priority noise sources.

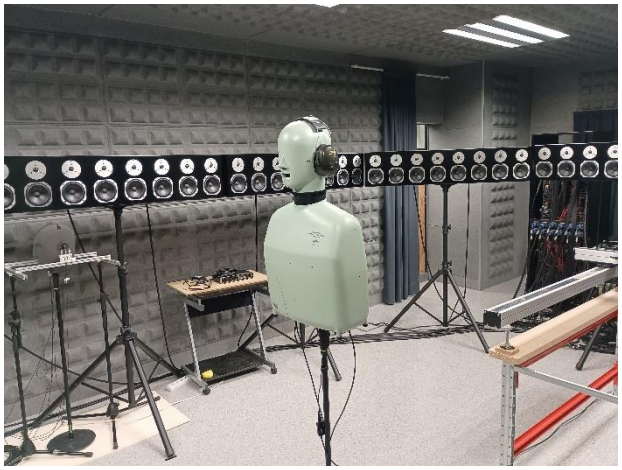

Figure 4. Experimental setup for data collection.

The experimental methodology involved positioning a head and torso simulator for sound quality applications, Model 4100 by Brüel & Kjaer [19], with the headphone prototype placed at the center of the WFS system, as shown in Figure 4. Interest signals were emitted from various directions in the horizontal plane, covering a full 360º range, with the measurement divided into arcs of randomly selected degrees (e.g., every 6º, 8º, 10º, etc.). Once the arc value was defined, the full 360º range was simulated with the respective degree difference. Simultaneously, the industrial background environment was played through planar waves, accompanied by point noise sources placed in specific locations. This approach allowed the system's ability to discriminate relevant signals in a complex acoustic environment to be assessed, challenging the artificial intelligence algorithms to differentiate between emergency signals and non-priority point and non-point noise sources.
Upon completion of the measurement process, 12.900 synchronized four-channel samples were obtained, one for each microphone installed in the measurement prototype.

4. SIGNAL PROCESSING AND AI SYSTEM DEVELOPMENT

4.1 Signal Processing and Model Structure for Emergency Signal Detection

The first step in the proposed system is sound detection, as this process triggers the subsequent stages of processing. Accurate detection of relevant signals is crucial, as it activates the following modules and ensures that only events of interest are processed. As an initial approach, the focus has been placed on identifying emergency sounds, such as sirens and horns, as outlined in Section 3. The methodology adopted in this work is based on strategies presented in previous studies [14], where a deep learning-based approach is employed for the detection of emergency sound events. Specifically, the system analyzes four-second audio segments using a sliding window, allowing continuous updates of sound information without compromising computational efficiency.

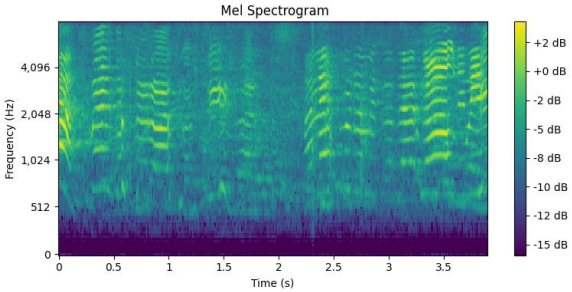

Figure 5. Mel Spectrogram of Siren Sample

For the representation of audio data, Mel spectrograms are used, as this representation preserves both the temporal and spectral information of the signal, facilitating the capture of distinctive features of emergency sounds. The configuration used in this work employs 128 Mel filters, a 32 ms analysis window for the Short-Time Fourier Transform, and a hop length of 10 ms, ensuring an adequate balance between temporal and spectral resolution, as shown in Figure 5.
Once the spectral representation of the audio is obtained, an artificial intelligence model based on Transformers [20] is employed, specifically the Audio Spectrogram Transformer (AST) [21], which has demonstrated superior performance in emergency signal detection compared to other deep learning architectures. Previous studies [22] have tested various configurations, including Convolutional Neural Networks (CNN) such as ResNet and VGG, Recurrent Neural Networks (RNN) such as LSTM, and hybrid architectures that combine CNN for spectral pattern extraction and LSTM for temporal dynamic analysis.

However, transformer-based models have shown greater generalization ability and improved accuracy in classifying emergency sounds.
The AST model has been specifically adapted for this task, modifying its final layers for classification into three categories: siren, horn, and other sounds or noise. To achieve this, the weights of the Multihead Attention blocks have been frozen, and custom classification layers have been added, as studied in [22].
This approach allows the system to differentiate between critical sounds and environmental noise, preventing the unnecessary activation of subsequent modules. Once an emergency signal is detected, the system generates an alert, providing the necessary information for a quick and precise response, thus triggering the activation of the following blocks of the system.

4.2 Signal Processing and Model Structure for Sound Localization

Once the emergency signal is detected, its localization is performed using an artificial intelligence-based system. To achieve this, information extracted from the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [23,24] is used, a parameter that allows the calculation of time delays between microphones. This parameter is crucial for estimating the direction of arrival (DOA) of the sound, as it provides information about the time difference at which the signal reaches each microphone in the array, as shown in Figure 6.

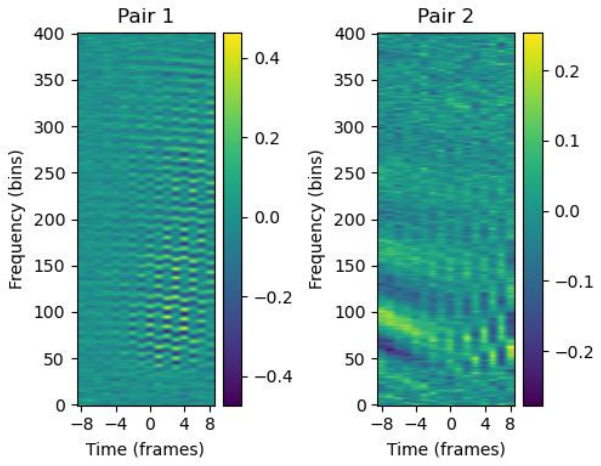

Figure 6. Example of windowed GCC-PHAT between two microphone pairs.

GCC-PHAT is particularly useful in noisy environments, as it applies phase normalization, which enhances the robustness of the time delay estimation by minimizing the impact of interferences and reflections. Additionally, since the microphone separation is relatively small, the time delays are also reduced. As explored in [15], to optimize processing and reduce computational load, a segmentation process of the GCC-PHAT is applied using analysis windows, which allows improving system efficiency without compromising the accuracy of the estimation.
As explored in [22], this parameter is used as input to a convolutional neural network based on ResNet-18, which has been modified for the task of angular localization. The extracted features are then processed by two independent branches, each consisting of a fully connected layer with 512 neurons and PReLU activation, followed by a final layer with Tanh activation. This configuration ensures that the output values are in the range of [-1, 1], which is consistent with the trigonometric values of an angle.
Finally, the direction of arrival of the signal is obtained using the arctangent function, which combines the sine and cosine values to calculate an angle in the range of 0° to 359°. As demonstrated in [22], this method enables precise localization of the sound source without ambiguity.

5. RESULTS

5.1 Detection Performance Results

Given the promising results obtained in [14], an attempt was made to replicate the same detection system with the aim of evaluating its performance in classifying emergency sounds. To this end, advanced Transformer architectures were employed, which have proven highly effective in capturing complex relationships in sequential data such as audio. However, these architectures require substantial computational capacity, presenting an additional challenge when working with large datasets such as AudioSet [25], which consists of over 2.2 million audio samples.
To address this computational demand, the transfer learning strategy presented in [22] was replicated, using a pre-trained model on the AudioSet dataset, covering 527 classes. This technique allowed the use of general acoustic representations without the need to fully retrain the model from scratch, significantly reducing training times and computational resource consumption without compromising classification accuracy.
Following the methodology described in [14], a fine-tuning process was applied in which the weights of the model's

multi-head attention layers were frozen, and only the final classification layers were retrained using the UrbanSound8K dataset [16].
In addition, the classes not corresponding to emergency sounds were grouped under the category "nothing" to simplify the classification task. The 10-fold cross-validation, as recommended in [16], allowed for evaluating the robustness of the model and minimizing the risk of overfitting.
The model training replicated the setup presented in [22], using the cross-entropy loss function and the ADAM optimizer. Furthermore, an adaptive learning rate and early stopping strategy were implemented improving the model's convergence and training efficiency.
During the training phase, the model achieved an average accuracy of 98.23%, reflecting its ability to detect emergency events with high precision. Key metrics such as precision (98.02%), recall (96.32%), and F1-Score (97.16%) corroborate the effectiveness of the replicated approach in distinguishing siren sounds from other environmental noises.
To validate the system, we evaluated its performance using the ESC-50 dataset [18]. The results showed a correct identification rate of 97.5%, confirming the system's effectiveness in detecting emergency sounds across different conditions.

5.2 Localization Performance Results

The localization model was trained using a dataset of 12.900 samples, obtained from the measurements detailed in Section 3. These measurements, conducted in an acoustically controlled environment with Wave-Field Synthesis (WFS) technology, allowed for the precise recreation of realistic scenarios, including emergency signals and background noise representative of industrial and urban environments. The fidelity of the system ensured that the model was exposed to varied and challenging conditions, essential for robust localization.
To maximize the representativeness of the dataset, a split of 80% for training, 10% for validation, and 10% for testing was established. The data partitioning was performed in a stratified manner, ensuring that all incident directions were well represented and avoiding biases in the distribution of angles.
The model training was conducted using the ADAM optimizer, chosen for its ability to handle large volumes of data and stabilize the learning process. Strategies such as early stopping were implemented to halt training when the loss on the validation set ceased to improve, thus preventing overfitting. Additionally, an adaptive learning rate was used to dynamically adjust the magnitude of weight changes, optimizing convergence.
Given that the localization task involves circular regression, the circular cosine loss function [22] was employed, specifically designed for problems where values at the extremes of the range (0° and 359°) are equivalent. This approach allowed for a more accurate evaluation of angular error and avoided inconsistencies that could arise with traditional loss functions.
The model's performance was evaluated in terms of the mean angular error in degrees on the test set, obtaining a value of 7.593°. These results reflect the model's ability to accurately estimate the direction of sound arrival in acoustically complex environments, replicating realistic conditions as described in Section 3.

6. FUTURE DIRECTIONS AND ENHANCEMENTS FOR SYSTEM FUNCTIONALITY

As mentioned at the beginning of this document, the described system is still in an early development phase, presenting a viable prototype of an acoustic protection system with emergency sound detection and localization. However, for this prototype to reach its full potential and become fully functional, a series of improvements and expansions are essential.
One of the first areas to develop in order to enhance the system is the representation of 3D audio. While the system already provides basic localization of detected sounds, it is necessary to refine how this information is presented to the user. Precision in the spatial representation of sounds is crucial for the user to clearly identify the exact location of sounds of interest, even in noisy environments with multiple sound sources. It is essential that the intensity, location, and direction of sounds be intuitive, allowing for a smooth experience that does not overwhelm the user, particularly in high-noise situations.
Another fundamental aspect to address is the implementation of sound cleaning algorithms. While the system is capable of detecting and localizing sounds of interest, an effective process for eliminating unwanted noise that interferes with the signal quality has not yet been developed. Filtering techniques are needed to isolate relevant sounds from interference, ensuring that the user only receives important alerts or sounds, without distractions or confusion. This step is key for the system to be truly useful in noisy industrial environments, where sound clarity and precision are essential.

Additionally, the system must expand its detection capabilities to include a wider variety of sounds of interest. While it primarily focuses on emergency sounds, detecting other types of noises, such as human speech, would make the system more versatile and effective in various situations. To achieve this, it will be necessary to train the system with a broader and more diverse dataset, enhancing the model's ability to classify and localize a wider range of sounds in different contexts.

Testing in real-world high-noise environments is also a crucial step for the system's development. Although the prototype has shown promising results under controlled conditions, real industrial environments present additional challenges. Constant noise and the presence of multiple acoustic sources in these environments require the system to be even more precise and adaptable. Testing in these scenarios will allow for adjustments to the detection and localization algorithms, ensuring the system maintains its reliability and accuracy even in intense noise conditions.

7. CONCLUSIONES

This study introduces an innovative intelligent hearing protection system designed to protect users while avoiding the problem of acoustic overprotection, which can occur when users are completely isolated from their environment. The proposed system combines passive noise isolation with AI-based sound processing to balance auditory protection and situational awareness in industrial environments. By employing MEMS microphones and deep learning models, the system detects and localizes critical sounds, ensuring that workers remain aware of their surroundings while receiving robust auditory protection.

The findings demonstrate that the hardware prototype effectively detects and localizes essential sounds in complex environments, validating the system's capability to distinguish between harmful noise and important auditory signals. The AI models, particularly those based on transformer architectures, exhibited superior performance in emergency sound detection, achieving high accuracy and robustness across various acoustic scenarios. For sound localization, the system utilizes a modified ResNet18 architecture, which significantly enhances the accuracy of spatial localization. Preliminary testing in a controlled environment confirmed the system's potential to improve workplace safety without compromising auditory health or regulatory compliance.

While the current prototype shows promising results, further work is needed to refine the 3D audio representation, implement sound cleaning algorithms, and expand detection capabilities to include a broader variety of sounds. Additionally, testing the system in real-world high-noise environments will be essential to ensure its reliability and effectiveness. This innovative approach offers a promising solution for enhancing workplace safety by providing a balance between effective hearing protection and the ability to perceive critical sounds, thereby improving overall worker safety and performance.

8. ACKNOWLEDGMENTS

This project was supported by the projects SONEM3D (CIAICO/2021/057) by Generalitat Valenciana and STARRING-BIGEAR (PID2022-137048OB-C42) by the Spanish Government.

9. PATENTS

The technology/results/methods described in this paper are being protected through a patent application currently under consideration.

10. REFERENCES

[1] Basner, M.; Babisch, W.; Davis, A.; Brink, M.; Clark, C.; Janssen, S.; Stansfeld, S. Auditory and Non-Auditory Effects of Noise on Health. *The Lancet* 2014, *383*.

[2] Lie, A.; Skogstad, M.; Johannessen, H.A.; Tynes, T.; Mehlum, I.S.; Nordby, K.C.; Engdahl, B.; Tambs, K. Occupational Noise Exposure and Hearing: A Systematic Review. *Int Arch Occup Environ Health* **2016**, *89*, doi:10.1007/s00420-015-1083-5.

[3] Themann, C.L.; Masterson, E.A. Occupational Noise Exposure: A Review of Its Effects, Epidemiology, and Impact with Recommendations for Reducing Its Burden. *J Acoust Soc Am* **2019**, *146*, doi:10.1121/1.5134465.

[4] Acton, W.I. Protection of Workers against Noise and Vibration in the Working Environment. *Occup Environ Med* **1978**, *35*, doi:10.1136/oem.35.1.80-a.

[5] 5OSHA Occupational Noise Exposure: Hearing Conservation Amendments; Final Rule. *Fed Regist* **1981**, *48*.

[6] EU-OSHA *Directive 2003/10/EC - Noise | Safety and Health at Work EU-OSHA*; 2021;

[7] HSE - Noise: Hearing Protection - Overprotection.

[8] Sayler, S.K.; Rabinowitz, P.M.; Galusha, D.; Sun, K.; Neitzel, R.L. Hearing Protector Attenuation and

Noise Exposure among Metal Manufacturing Workers. *Ear Hear* **2019**, *40*, doi:10.1097/AUD.0000000000000650.
[9] Silva, V.A.R.; Guimarães, A.C.; Lavinsky, J.; De Castro, R.F.; Freitas, P.P.; Castilho, A.M.; Chone, C.T.; Crespo, A.N. Are Hearing Protection Devices Used in the Workplace Really Efficient? A Systematic Review. *Work* 2022, *74*.
[10] salvatore giumento This Is Information on a Product in Full Production. MP34DT01-M. **2014**.
[11] Letowski, T.R.; Letowski, S.T. Auditory Spatial Perception : Auditory Localization. *Army research laboratory* **2012**.
[12] Li, S.; Peissig, J. Measurement of Head-Related Transfer Functions: A Review. *Applied Sciences (Switzerland)* 2020, *10*.
[13] XVF3000/XVF3100-TQ128 Datasheet. **2017**.
[14] Banchero, L.; López, J.J. Clasificación de Sonidos En El Exterior de Vehículos Mediante Inteligencia Artificial. In Proceedings of the Tecniacústica; Sociedad Acústica Española: Cuenca, October 2023.
[15] Banchero, L.; Javier López, J. javier Real-World Environment Simulation for Validation of AI Sound Detection and Localization. In Proceedings of the AES Europe 2024; Audio Engineering Society: Madrid, Spain, June 2024.
[16] Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the Proceedings of the 22nd ACM International Conference on Multimedia; Association for Computing Machinery: New York, NY, USA, 2014; pp. 1041–1044.
[17] Asif, M.; Usaid, M.; Rashid, M.; Rajab, T.; Hussain, S.; Wasi, S. Large-Scale Audio Dataset for Emergency Vehicle Sirens and Road Noises. *Sci Data* **2022**, *9*, doi:10.1038/s41597-022-01727-2.
[18] Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In Proceedings of the Proceedings of the 23rd ACM International Conference on Multimedia; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1015–1018.
[19] PRODUCT DATA Sound Quality Head and Torso Simulator Types 4100 and 4100-D.
[20] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems; 2017; Vol. 2017-December.
[21] Gong, Y.; Chung, Y.A.; Glass, J. Ast: Audio Spectrogram Transformer. In Proceedings of the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; 2021; Vol. 1.
[22] Banchero, L.; Vacalebri-Lloret, F.; Mossi, J.M.; Lopez, J.J. Enhancing Road Safety with AI-Powered System for Effective Detection and Localization of Emergency Vehicles by Sound. *Sensors* **2025**, *25*, doi:10.3390/s25030793.
[23] Brandstein, M.S.; Silverman, H.F. Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings; 1997; Vol. 1.
[24] Knapp, C.H.; Carter, G.C. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans Acoust* **1976**, *24*, doi:10.1109/TASSP.1976.1162830.
[25] Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings; 2017.