electronics

Article

Smart Industrial Safety in High-Noise Environments Using IoT and AI

Alessia Bramanti [1], Luca Catarinucci [2], Mattia Cotardo [2], Rosaria Del Sorbo [1], Claudia Giliberti [3], Mazhar Jan [2], Luca Landi [4], Raffaele Mariconte [3], Teodoro Montanaro [2], Federico Paolucci [4], Luigi Patrono [2,*], Davide Rollo [2], Francesco Antonio Salzano [1] and Ilaria Sergi [2]

[1] Department of Medicine, Surgery and Dentistry, University of Salerno, Via S. Allende, 84081 Baronissi, Italy; abramanti@unisa.it (A.B.); rdelsorbo@unisa.it (R.D.S.); frsalzano@unisa.it (F.A.S.)
[2] Department of Engineering for Innovation, University of Salento, Via Monteroni, 1, 73100 Lecce, Italy; luca.catarinucci@unisalento.it (L.C.); mattia.cotardo@studenti.unisalento.it (M.C.); mazhar.jan@studenti.unisalento.it (M.J.); teodoro.montanaro@unisalento.it (T.M.); davide.rollo1@studenti.unisalento.it (D.R.); ilaria.sergi@unisalento.it (I.S.)
[3] Department of Technological Innovations and Safety of Plants, Products and Anthropic Settlements, INAIL, Via Roberto Ferruzzi 38/40, 00143 Rome, Italy; c.giliberti@inail.it (C.G.); r.mariconte@inail.it (R.M.)
[4] Department of Engineering, University of Perugia, Via G. Duranti, 93, 06125 Perugia, Italy; luca.landi@unipg.it (L.L.); federico.paolucci@dottorandi.unipg.it (F.P.)
* Correspondence: luigi.patrono@unisalento.it; Tel.: +39-0832-297330

Abstract

High noise levels in industrial workplaces pose significant challenges to occupational safety, particularly with hearing protection and effective communication. Traditional hearing protection devices, while effectively attenuating harmful noise, often compromise situational awareness by excessively isolating workers from the acoustic environment and preventing the perception of critical auditory cues (e.g., emergency alarms), thereby introducing additional safety risks. This paper presents a smart industrial safety system that integrates Internet of Things (IoT) and artificial intelligence (AI) and is based on intelligent hearing protection devices to (a) selectively attenuate hazardous industrial noise while (b) preserving human speech and (c) reproduce targeted audio notifications to workers near malfunctioning or hazardous machinery. A real-time voice activity detection (VAD) model is employed to distinguish vocal components from background noise to adaptively control digital signal processing filters. Furthermore, indoor localization enables the delivery of targeted audio messages to workers in proximity to relevant events. Experimental evaluations on embedded hardware demonstrate that the selected VAD model operates well within real-time constraints and effectively supports dynamic noise filtering. Objective evaluation of the filtering stage using Mean Opinion Score (MOS), signal-to-noise ratio (SNR), and Harmonics-to-Noise Ratio (HNR) shows consistent quality improvements across all tested conditions, with MOS gains up to +118%, SNR increases between +10.4 and +29.0 dB, and HNR improvements up to +6.22 dB, indicating enhanced speech intelligibility and preservation of voice harmonic structure even under high-noise scenarios. Robustness validation of the VAD module across varying acoustic conditions confirms reliable speech detection performance, achieving perfect classification at +10 dB SNR, very high accuracy at 0 dB (98.3%, ROC AUC 0.998), and stable operation even at −7 dB SNR (79.8% accuracy, ROC AUC 0.878). The proposed architecture achieves a balanced trade-off between hearing protection and speech intelligibility while enhancing the effectiveness of safety communications in noisy industrial environments.

Academic Editor: Domenico Ursino

Received: 11 February 2026
Revised: 11 March 2026
Accepted: 16 March 2026
Published: 20 March 2026

Keywords: industrial safety; smart hearing protection; adaptive noise filtering; voice activity detection; Internet of Things; artificial intelligence; occupational noise; wearable devices

1. Introduction

Workplace safety represents a fundamental pillar of modern industrial systems, as safeguarding workers' well-being is essential to sustain high levels of productivity and quality of life [1]. To this aim, over the past few decades, advanced technologies such as artificial intelligence (AI), the Internet of Things (IoT), and Cloud Computing have been integrated into industrial production processes. This perspective is fully aligned with the vision of Society 5.0 [2], a recent paradigm that aims to enhance human well-being through the intelligent and responsible use of technology, including significant improvements in occupational safety. Nevertheless, despite continuous technological progress, workplace accidents still occur, highlighting the persistent need for effective protective solutions.

One of the most problematic situations that undermine worker safety in industrial domains concerns the noise levels to which workers can be exposed during their working hours.

According to the World Health Organization (WHO), approximately 360 million people worldwide are affected by severe hearing loss, confirming that hearing impairment is a major global health issue. Noise levels of 85 dB(A) are widely recognized as the maximum tolerable threshold beyond which the risk of hearing damage significantly increases [3]. Prolonged exposure to noise levels exceeding regulatory safety thresholds can result in hearing loss, increased stress, and other long-term health consequences for workers [4]. One of the most well-documented and consequential manifestations of this process is the Occupational Noise-Induced Hearing Loss (ONIHL), a gradual and irreversible condition caused by prolonged exposure to high noise levels in the workplace. It typically begins by affecting high-frequency hearing, particularly around 4 kHz, before progressively extending to adjacent frequencies. ONIHL accounts for approximately 16% of disabling hearing loss cases among adults worldwide [5]. Moreover, numerous studies have demonstrated that many industrial environments regularly expose workers to hazardous noise levels, including construction sites [6], textile mills [7], pharmaceutical companies [8], and steel factories [9]. Beyond hearing impairment, prolonged noise exposure has also been associated with extra-auditory effects such as mood disorders, sleep disturbances, and cardiovascular diseases [6].

Given that noise-induced hearing damage can be irreversible, prevention strategies are of paramount importance. These strategies typically include continuous noise monitoring, administrative controls, and the use of personal protective equipment (PPE). Regulatory bodies have established specific exposure limits to mitigate risk. The Occupational Safety and Health Administration (OSHA) defines a permissible exposure limit (PEL) of 90 dBA over an 8 h workday, with allowable exposure time halved for every 5 dBA increase. In contrast, the National Institute for Occupational Safety and Health (NIOSH) recommends a more conservative limit of 85 dBA over the same duration [10]. OSHA's hearing conservation program (HCP) emphasizes systematic noise assessment, periodic audiometric testing, the use of appropriate hearing protection devices, worker training, and the continuous evaluation of protective measures. When workplace noise levels exceed 85 dBA, employers are required to implement a comprehensive hearing conservation program to ensure adequate protection for workers [10].

In response to the growing demand for more effective hearing protection in noisy industrial environments, several research initiatives are focusing on the development of

intelligent hearing PPE that surpasses the limitations of traditional solutions that usually apply passive noise cancellation through specifically designed headphones. In fact, while conventional hearing protection devices are effective in reducing noise exposure, they often lead to sensory isolation and significantly hinder communication. This limitation can prevent workers from perceiving critical auditory cues, such as emergency alarms, verbal warnings from colleagues, or sounds generated by nearby moving machinery. The complete suppression of auditory input may therefore create hazardous situations, reducing workers' ability to respond promptly to emergencies or avoid potential risks [11–13].

In line with the principles of Industry 4.0 [14–16], the present work aims to design, implement, and evaluate a smart hearing protection system based on IoT and AI technologies to selectively attenuate harmful or disruptive noise while preserving relevant safety-related sounds, such as human speech. Furthermore, the system delivers targeted voice notifications to workers located near malfunctioning machines, thereby enhancing the effectiveness of safety communications. Beyond the described capabilities, two core requirements have been taken into consideration in the design of the proposed system: its responsiveness and its acceptability. Specifically, any hearing protection becomes effective only if the workers accept to use it and if it is able to filter noise while preserving, in real time, safety-related sounds. Therefore, the architecture has been designed according to a "lightweight" philosophy that enables real time functions on IoT small devices that the workers could better accept in their daily tasks. Therefore, it has been designed to work on embedded hardware platforms characterized by limited computational resources and low cost.

The proposed system delivers alarm notifications directly via smart headphones by generating and reproducing a synthesized version of the alarm, ensuring clear and immediate perception without further increasing ambient noise levels. This feature ensures compliance with the most important international safety-related standards. Within this framework, international standards such as ISO 7731:2003 [17] and ISO 11429 [18] provide important guidelines for the design of auditory danger signals in public and occupational environments. ISO 7731:2003 specifies that auditory warning signals should exceed ambient noise levels by at least 15 dB in hazardous areas to ensure reliable perception. Compliance with such a similar requirement in industrial settings may lead to further acoustic overload caused by playing messages at a volume that is higher than already elevated background noise. Similarly, ISO 11429, aligned with the European standard EN 981 [19], offers a comprehensive framework for both auditory and visual hazard signals, but its applicability remains limited in highly specialized industrial contexts. Therefore, the approach proposed in this paper improves both safety and ergonomic compliance, effectively overcoming a key drawback of conventional auditory signaling systems.

From a technological perspective, the system leverages communication protocols such as MQTT, a lightweight communication protocol suitable for embedded and constrained systems [20,21] and Bluetooth Low Energy (BLE) [22,23] beacons to enable accurate indoor localization [24,25] and real-time dissemination of safety-related messages. The smart hearing protection devices are integrated into a broader software architecture that collects data from industrial machinery and provides a centralized dashboard for infrastructure management. In parallel, the headphones are responsible for implementing intelligent noise filtering mechanisms. A key component of the proposed solution is the integration of an AI-based voice activity detection (VAD) model. The selected model, Silero VAD [26], is capable of distinguishing human speech from background noise even in highly noisy industrial environments and in real time, despite other similar solutions available in the literature. Such knowledge is used to dynamically and adaptively filter noise through a digital signal processor (DSP), adjusting attenuation levels based on ambient noise conditions and the

detected presence of speech. This ensures an effective balance between hearing protection and situational awareness. The system operates in real time by continuously acquiring environmental audio signals through onboard microphones. To assess the performance of the proposed approach, a dedicated dashboard has been developed, allowing users to test the VAD model on audio samples containing industrial noise and to evaluate the effectiveness of the adaptive filtering strategy.

Contributions and positioning. Unlike traditional hearing protection devices and existing intelligent PPE solutions that primarily focus either on passive attenuation or on isolated signal processing techniques, this work proposes an integrated cyber–physical system that jointly addresses auditory protection, situational awareness, and safety communication in industrial environments. The main contributions of this paper can be summarized as follows:

- Technical novelty: We introduce an adaptive hearing protection mechanism that combines AI-based VAD with real-time DSP-driven noise attenuation, enabling selective suppression of harmful noise while preserving speech intelligibility in highly non-stationary industrial acoustic scenarios. The system operationalizes VAD outputs to dynamically control attenuation policies, rather than using speech detection solely for monitoring purposes.
- System-level design: We design and implement a distributed IoT-enabled architecture that integrates smart headphones, industrial machinery, and a centralized management platform. The system supports context-aware alarm delivery through personalized audio streams, enabling compliance with auditory safety standards without increasing environmental noise emissions.
- End-to-end validation framework: We provide an experimental infrastructure that combines real-time audio acquisition, adaptive filtering evaluation, and IoT-based event dissemination, enabling quantitative assessment of both signal-level performance and system-level responsiveness.
- Cost-effectiveness and accessibility: Design of a system based on low-cost off-the-shelf components and algorithm optimization to ensure real-time performance on embedded devices with limited computational power.

These elements collectively define a unified approach that advances current intelligent hearing protection systems by integrating adaptive audio processing, wearable computing, and industrial IoT communication within a single operational framework.

The remainder of the paper is organized as follows. Section 2 reviews the state of the art in noise reduction techniques, intelligent hearing protection systems, and VAD approaches, with particular emphasis on AI-based solutions for noisy environments. Section 3 discusses the industrial regulations and standards relevant to occupational noise exposure and auditory safety. Section 4 presents the overall system architecture, detailing the interaction between smart headphones, industrial machinery, and the central server. Section 5 describes the system implementation, focusing on hardware components, audio processing pipelines, and communication mechanisms, and Section 6 reports the experimental validation of the proposed solution, including real-time performance evaluation, adaptive noise filtering effectiveness, and IoT-based alarm delivery. Finally, Section 7 discusses the results and their implications, while Section 8 concludes the paper and outlines future research directions.

2. State of the Art

Prolonged exposure to high noise levels has well-documented detrimental effects, which has driven the development of increasingly advanced noise reduction techniques. Early approaches were mainly based on passive hearing protection devices, such as headphones or earmuffs, which attenuate sound by introducing a physical barrier between the

source and the listener [27]. While effective in certain conditions, passive methods are limited in their ability to adapt to complex acoustic environments, leading to the adoption of more sophisticated solutions based on Active Noise Cancellation (ANC). ANC systems rely on the generation of anti-noise signals [28] and make use of fixed or adaptive filters whose impulse response is continuously updated to match changing noise characteristics [29]. Additional signal processing strategies include spectral subtraction techniques, which estimate and suppress noise components in the frequency domain [30], as well as approaches based on the Wavelet Transform that enable adaptive time–frequency analysis comparable to the Short-Time Fourier Transform (STFT) [31].

More recent ANC systems increasingly rely on AI-based models, which require informative audio representations extracted through time–frequency analysis, most commonly using the Short-Time Fourier Transform (STFT) [32]. Among the most widely adopted features, Mel-Frequency Cepstral Coefficients (MFCCs) provide compact representations of perceptually relevant spectral characteristics and are widely used in speech detection and noise reduction systems [33]. MFCCs belong to the broader class of cepstral-domain features, which also includes Linear Prediction Cepstral Coefficients (LPCCs) and Perceptual Linear Prediction Cepstral Coefficients (PLP) [34], and they are often combined with complementary time- and frequency-domain descriptors to improve robustness in complex acoustic environments [35].

Since audio features can be interpreted as grayscale images, MFCCs are particularly well suited for training convolutional neural networks (CNNs) [36]. Although CNNs are highly effective in extracting local spectro-temporal patterns, recent studies have highlighted their limited capability in modeling long-range temporal dependencies when used in isolation. To address this limitation, hybrid architectures combining CNNs with recurrent layers have been proposed. For instance, CNN-BiLSTM models have demonstrated improved performance in VAD and speech-related tasks by jointly exploiting local feature extraction and bidirectional temporal modeling [37]. Similarly, LSTM-based and hybrid LSTM-thresholding approaches have shown robustness in speech activity detection by explicitly modeling temporal dynamics over auditory features [38]. More recently, transformer-based architectures leveraging self-attention mechanisms have been introduced for audio processing, enabling the modeling of globally distributed semantic relationships across long temporal contexts and achieving state-of-the-art performance in challenging multilingual scenarios [39].

Within this evolving landscape, CNN-based approaches remain relevant due to their computational efficiency and suitability for embedded or real-time ANC applications. Moreover, they can be effectively integrated with recurrent or attention-based modules to enhance global context modeling. In this context, Autoencoders (AEs) play a key role in noise cancellation by learning compact latent representations that enable signal reconstruction, with Denoising Autoencoders (DAEs) being especially effective in reconstructing clean signals from noisy inputs [40].

The features and representations discussed above constitute the foundation of many modern noise cancellation systems and have been extensively employed in recent works that explore architectures closely related to the one proposed in this study. Several contributions in the literature combine IoT technologies, AI techniques, and PPE, demonstrating effective noise reduction capabilities while also revealing specific limitations. For instance, the system presented by Bhope et al. [41] integrates classical adaptive filtering with a neural network trained on MFCCs to classify acoustic environments and switch between Street, Workspace, and General Ambience modes, while the ANC stage relies on a Least Mean Square algorithm [42] to generate phase-inverted anti-noise signals. Similarly, the architecture described in [43] employs an embedded controller that transmits noise signals

to a coprocessor, where a 2D CNN trained on Log-Mel Spectrograms selects the most effective control filter for updating system coefficients. In the domain of pedestrian safety, the eNext earphones [44] use MFCC- and spectrogram-based features to detect emergency signals such as sirens or alarms and separate them from background noise. This separation process is supported by Blind Source Separation (BSS) techniques [45], which operate in the frequency domain through STFT-based analysis to decouple overlapping audio sources. By extending the concept of selective auditory perception, Veluri et al. [46] propose a binaural system that leverages STFT representations to train a neural network capable of isolating specific environmental sounds, such as birdsong, within complex noisy scenes. Finally, the work presented in [47] introduces a targeted noise cancellation approach based on complex STFT features and latent embeddings acquired during a short enrollment phase, allowing a neural network to learn a speaker's vocal characteristics and isolate their voice from surrounding interference even in dynamic environments.

A further fundamental component in noise filtering pipelines is VAD, whose purpose is to distinguish speech from background noise within an audio signal. VAD is commonly used as a preliminary stage in denoising systems to ensure that filtering operations primarily suppress noise while preserving speech intelligibility, thereby improving the overall effectiveness of subsequent processing stages [48,49]. An example is provided in the work presented by Kos [50], where spectral-domain noise subtraction is guided by the output of a neural network-based VAD model. Numerous VAD techniques have been proposed, ranging from traditional signal processing approaches to advanced machine learning-based solutions [51]. Classical methods include Zero Crossing Rate (ZCR) VAD, which identifies speech through sign changes in the signal, and Short Time Energy (STE) VAD, which evaluates signal energy over short temporal windows. More advanced statistical techniques, such as Gaussian Mixture Model (GMM) VAD, improve discrimination between speech and noise, while rVAD combines energy, ZCR, GMM, and the Laplacian Mixture Model (LMM) within a likelihood-based classification framework. With the advent of deep learning, neural network-based VAD approaches have become increasingly prominent, exploiting Deep Neural Networks (DNNs) trained on labeled datasets to achieve highly accurate speech detection.

3. Industrial Regulation

Since occupational noise exposure and auditory warning systems are strictly regulated in industrial environments, the design of the proposed system must comply with existing safety standards. For this reason, this section summarizes the main international regulations relevant to occupational noise and hearing protection, which constitute the regulatory framework guiding the system design.

The protection of workers from occupational noise exposure is governed by regulatory frameworks that establish mandatory exposure limits and specify technical requirements for PPE.

3.1. Noise Exposure Limits

European Directive 2003/10/EC [52] defines three critical exposure action values for occupational noise:

- Lower exposure action values of LEX,8h = 80 dB(A) and ppeak = 112 Pa (135 dB(C));
- Upper exposure action values of LEX,8h = 85 dB(A) and ppeak = 140 Pa (137 dB(C));
- Exposure limit values of LEX,8h = 87 dB(A) and ppeak = 200 Pa (140 dB(C)).

Hearing protection must be provided when exposure exceeds lower action values and becomes mandatory above upper action values. In the United States, OSHA establishes a permissible exposure limit of 90 dBA over 8 h, with exposure time halved for every

5 dBA increase [53]. NIOSH recommends the more conservative limit of 85 dBA with a 3 dB exchange rate [54]. Furthermore, the Machinery Directive 2006/42/EC requires manufacturers to minimize noise emissions and declare sound pressure levels (SPLs) at workstations or sound power level in technical documentation [55].

3.2. Standards for Hearing Protection and Warning Signals

This paragraph summarizes the most important standards dealing with hearing protection and warning signals with the aim of providing the most important regulations acting in the field.

The ISO 4869-1:1990 [56] standard establishes measurement methods for hearing protector attenuation, while the EN 352 series [57] specifies performance requirements and conformity assessment procedures for earmuffs and earplugs. These standards define minimum attenuation performance across frequency bands, providing the basis for noise reduction ratings that guide PPE selection. Moreover, ISO 7731:2003 [17] requires auditory warning signals to exceed ambient noise by at least 15 dB to ensure reliable perception. However, achieving this signal-to-noise ratio in high-noise industrial environments through increased alarm volume creates additional acoustic overload. In addition, ISO 11429:1996 [18] provides a framework for integrating auditory and visual danger signals.

3.3. Regulatory Implications for Adaptive Hearing Protection

Traditional passive hearing protection devices provide effective noise attenuation but simultaneously reduce all sounds, including speech communication and safety-relevant acoustic cues. This creates a conflict between compliance with exposure limits and maintenance of situational awareness. The existing regulatory framework does not explicitly address intelligent hearing protection systems that selectively attenuate noise while preserving speech intelligibility. The system proposed in this work addresses these requirements through adaptive filtering that ensures compliance with Directive 2003/10/EC exposure limits [52] by dynamically adjusting attenuation based on measured SPLs. Alarm notifications are delivered directly through integrated headphone speakers, satisfying ISO 7731 signal-to-noise requirements [17] without increasing workplace noise levels. This approach overcomes the fundamental limitation of traditional auditory signaling systems in high-noise environments while maintaining the attenuation performance required by EN 352 standards [57].

4. Architecture Design

Figure 1 reports the architecture designed for the system proposed in this paper. The architecture has been conceived not only to enable adaptive noise filtering and context-aware safety communication, but also to address the regulatory requirements governing occupational noise exposure and auditory warning systems discussed in Section 3. In particular, the system design aims to ensure adequate hearing protection in accordance with Directive 2003/10/EC and EN 352 standards, while preserving the perception of safety-related sounds and alarms as required by ISO 7731. The following paragraphs detail each component of the architecture and summarize the most relevant functionalities provided by the different modules.

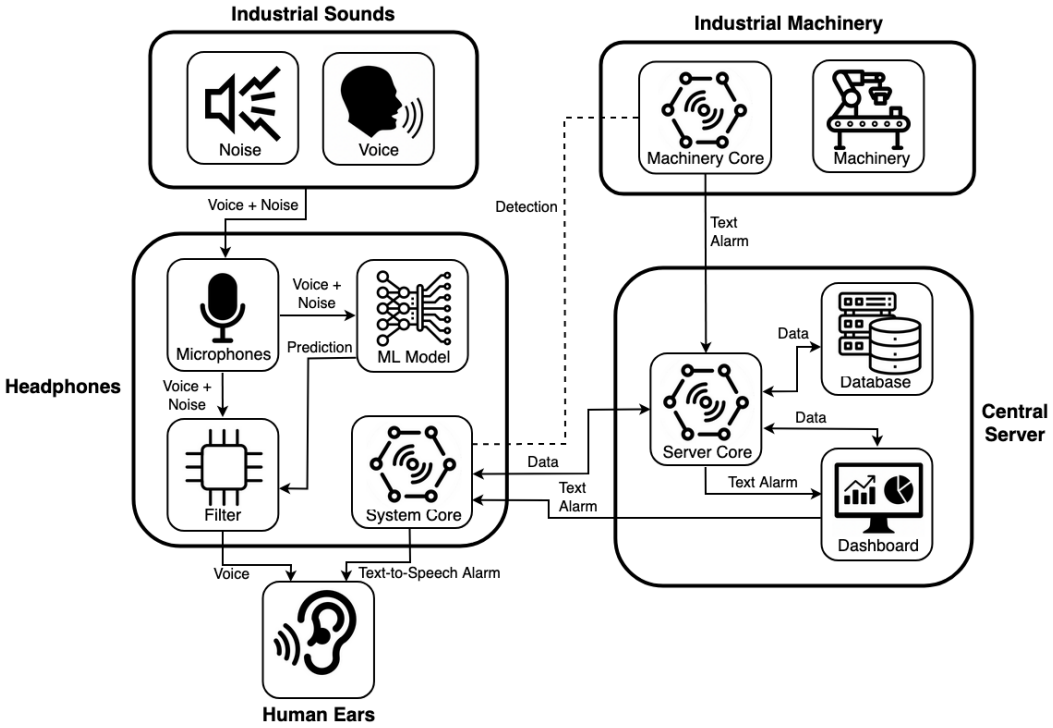

Figure 1. High-level architecture of the proposed industrial safety system. The diagram illustrates the interaction between the industrial acoustic environment, smart headphones, industrial machinery, and the central server. Environmental sounds, composed of noise and voice, are captured and processed by the headphones through voice detection and adaptive filtering. Localization, data exchange, and alarm management are handled via communication with smart machinery and a local central server, which supervises data storage, monitoring, and safety notifications through a dedicated dashboard.

4.1. Industrial Sounds

The "Industrial Sounds" subsystem represents the acoustic environment within an industrial workplace. This environment is mainly characterized by two distinct sound categories. The first category consists of noises generated by industrial machinery, such as motors, presses, lathes, milling machines, conveyors, and compressors. These sounds are typically high-intensity, continuous or impulsive, and potentially harmful to human hearing, and therefore need to be attenuated. The second category is represented by human voices, namely the speech of nearby workers, which must be preserved to remain clearly audible by nearby colleagues in order to maintain situational awareness and increase overall safety. Within the industrial environment, these two sound categories are naturally superimposed. As a result, the acoustic waves propagating through the workspace contain both noise and speech information merged into a single audio signal. This combined sound wave is the input that reaches the workers and, consequently, the smart headphones.

4.2. Headphones

The "Headphones" subsystem constitutes the core filtering and notification interface worn by workers during their activities. Once workers arrive at their workstation, they are required to wear and activate the headphones. This subsystem is composed of several interconnected blocks, each with a specific role in audio acquisition, processing, and reproduction.

4.2.1. Microphones

Although the architectural diagram depicts a single microphone for clarity, the actual system employs three microphones. Two microphones are positioned at ear level, one for each ear. Their purpose is to capture the ambient audio signal, which contains both

voice and noise components, with the aim of preserving the situational awareness, usually guaranteed by binaural listening (the right ear will hear a potentially different sound with respect to the left ear). These signals are the ones that will later be filtered and reproduced through the ear cups, reaching the worker's ears. A third microphone is, instead, positioned on the worker's head, capturing an additional audio stream that is exclusively used as input for the machine learning algorithm responsible for voice detection. The information extracted from this audio stream, combined with the measured SPL in decibels of the environment, is used to dynamically determine the optimal filtering level to be applied to the audio signals captured by the ear-level microphones.

4.2.2. Machine Learning Model

The "ML Model" block implements a VAD algorithm. It receives as input the audio signal acquired by the third microphone. This signal is divided into temporal chunks in order to enable real-time processing. For each chunk, the model estimates a continuous confidence score representing the likelihood of vocal presence, rather than producing a binary speechnon-speech decision. The resulting confidence value, together with the environmental decibel level, is forwarded to the filtering stage. These parameters are used to dynamically modulate the attenuation level of the DSP, allowing the system to adaptively balance noise suppression and speech preservation. Consequently, the VAD output is not used as a hard filtering gate, but as a control signal that continuously adjusts the filtering strength, preventing abrupt transitions between fully filtered and unfiltered audio.

Our work aimed at the development and validation of a unified IoT-AI architecture for industrial safety; therefore, we chose the Silero VAD model over more complex alternatives, such as DAE-based architectures, as a deliberate compromise between real-time performance and embedded implementation feasibility. The Silero VAD model was selected for its superior performance and robustness compared to traditional open-source baselines like WebRTC. In fact, a recent comparative study [58] demonstrates that the neural-based Silero model significantly outperforms WebRTC across diverse real-world audio streams, achieving higher ROC AUC and average precision scores. This choice aligns with our "lightweight" philosophy presented in the introduction, prioritizing low-latency operation and minimal computational requirements while maintaining sufficient detection accuracy for the industrial acoustic environments targeted by our system. In addition, its exploitation allowed us to overcome one of the most limiting problems of experiments in the field, which is the availability of data. Based on the state-of-the-art analysis presented in the previous sections, a lack of a comprehensive dataset of industrial noises emerged, which could be used to train other more complex ML models and then used for predictions. The use of the Silero VAD model, instead, allowed the authors to achieve good results without recording millions of hours of noise coming from a hypothetically infinite number of machines.

4.2.3. Filter

The "Filter" block is an electronic filtering module that operates on the audio streams acquired by the two ear-level microphones. Based on the information provided by the VAD predictions and the measured decibel levels, the filter dynamically adapts its attenuation characteristics over fixed time intervals, for example every half a second. The objective of this block is to suppress harmful industrial noise while preserving human voices. The filtered audio signal, enriched with intelligible speech, is then reproduced through the ear cups, allowing the worker to perceive a safe and intelligible version of the surrounding acoustic environment.

4.2.4. System Core

The "System Core" block includes all the modules that enable the headphones to interact with external entities, such as industrial machinery and the central server. It acts as a communication and coordination layer between the wearable device and the rest of the system. One module is dedicated to the detection of smart industrial machinery in the surrounding environment. Its purpose is to allow each pair of headphones to determine its relative position, identifying which machine the worker is currently closest to. A second module manages communication with the central server. Through this module, the headphones transmit localization information derived from machinery detection, enabling the server to maintain an up-to-date mapping of worker positions. The same communication channel is also used to receive additional data from the server, such as industrial environment configurations required for correct system operation. Finally, an alarm management module is responsible for handling safety notifications. This module receives textual alarm messages sent by the safety manager through the dashboard. These messages are converted from text to synthesized speech using a text-to-speech engine and are immediately reproduced in the ear cups. This mechanism ensures that critical safety information is clearly delivered to workers, even while noise filtering is active.

4.3. Industrial Machinery

The "Industrial Machinery" subsystem represents the set of machines operating within the industrial environment, such as lathes, milling machines, presses, robotic arms, and conveyor systems. Each machine, in addition to its standard operational components, is equipped with an additional communication unit referred to as the "Machinery Core". This module enables the machine to be detected by nearby smart headphones and to communicate with the central server. When a malfunction or anomalous condition occurs, the machinery generates a textual alarm message through its Machinery Core. This message is transmitted to the central server, where it is stored in the database and visualized on the dashboard. In this way, machine-level events are seamlessly integrated into the global safety monitoring infrastructure.

4.4. Central Server

The Central Server is the central element of the system and is responsible for data management, coordination, and alarm handling. The server is designed to operate locally within the industrial facility, without relying on cloud infrastructure, in order to guarantee low latency and immediate responsiveness. This design choice is essential for safety-critical applications.

4.4.1. Server Core

The "Server Core" block contains the modules responsible for receiving data from external entities and storing them in the database. Examples include localization data sent by the headphones and status or alarm messages received from industrial machinery. In addition, the Server Core retrieves configuration data from the database and transmits it to the headphones, enabling proper adaptation to the industrial environment.

4.4.2. Database

The Database stores all system-related information, including worker localization data, machinery data, headphone configurations, and general industrial environment parameters. It represents the persistent knowledge base of the entire system.

4.4.3. Dashboard

The Dashboard is the interface continuously monitored by the safety manager. It aggregates data retrieved from the database and textual alarms received from industrial machinery, presenting them through a clear, intuitive, and informative visualization. Beyond visualization, the dashboard supports data management operations, including the modification of existing information and the creation of new entries, such as CRUD operations on workers, headphones, and machinery. It also enables the management of alarm notifications. Textual alarms generated by machines can be selectively forwarded to workers located in their vicinity, based on localization data. Additionally, the dashboard allows the safety manager to send customized service messages to individual workers, which are then synthesized and delivered through the headphones.

5. Implementation

This section describes the implementation details related to each component of the presented system.

5.1. System Component Implementation

This subsection describes the implementation of the wearable device and of the embedded system that enables real-time audio processing, adaptive noise reduction, localization, and safety communication. The focus is on the practical realization of the architecture previously introduced, highlighting hardware choices, signal processing pipelines, and software integration.

Regarding audio acquisition, low-cost electret lavalier microphones (widely used small capacitor microphones) were selected based on preliminary experimental results [59] and prior validation campaigns comparing frequency response and directivity patterns against professional measurement equipment [60], because they demonstrated sufficient accuracy for speech detection applications while maintaining cost-effectiveness and compact design suitable for integration into the headphone housing. The selected microphones provide a fixed sampling rate of 48 kHz and are used to capture the acoustic environment surrounding the worker. As described in the architectural design, two microphones are positioned at ear level and are dedicated to audio reproduction after filtering, while a third microphone is used exclusively to feed the VAD pipeline.

Once captured, the audio stream is organized to be compatible with the selected VAD model, namely Silero VAD. This model was chosen due to its suitability for real-time applications and its high computational efficiency. Performance evaluations reported in the official Silero VAD documentation [26] show that, on high-end CPUs, the model is capable of processing audio chunks significantly faster than real time. To further validate its applicability in our scenario, additional tests were conducted directly on a Raspberry Pi 4 [61] using audio sampled at 48 kHz and chunk sizes equivalent to 32 ms. In this configuration, Silero VAD processed each chunk in approximately 2.5 ms, resulting in a real-time speed well above the requirements of the system and confirming its feasibility for continuous, low-latency operation on embedded hardware.

For real-time inference, Silero VAD requires an input sampling rate of 16 kHz. Since the microphones are connected to an ADC that natively operates at 48 kHz, the audio stream is initially opened at this rate and segmented into chunks of 1536 samples, corresponding to three times the standard 512-sample window used at 16 kHz. Each chunk is then converted from `int16` to `float32` format and normalized with respect to its maximum absolute value. After this conversion, the chunk is resampled to 16 kHz and passed to the VAD model. The output of the model is a floating-point confidence score in the range $[0, 1]$, representing

the estimated probability of voice activity within the analyzed time window. The complete processing chain is summarized in Algorithm 1.

Algorithm 1 Audio processing in real time with Silero VAD model

```
procedure AUDIOPROCESSING
    Input: Audio stream at 48 kHz
    Output: Confidence score between 0 and 1 for each chunk
    current_sr ← 48,000
    target_sr ← 16,000
    chunk_size ← 1536
    stream ← open_audio_stream(current_sr)
    sample ← 0
    while True do
        chunk ← stream[sample : sample + chunk_size]
        sample ← sample + chunk_size
        chunk ← convert_to_float32(chunk)
        chunk ← normalize(chunk)
        chunk ← resample(chunk, target_sr)
        confidence_score ← SileroVAD(chunk)
        yield confidence_score
    end while
end procedure
```

The confidence score produced by the VAD model for each chunk is combined with the equivalent SPL, measured in decibels, to determine the filtering level applied by the digital signal processors. The DSPs accept a discrete filtering level ranging from 0 to 7, where higher values correspond to stronger noise attenuation. To reduce temporal fluctuations and improve stability, the voice activity confidence v is not computed on a single chunk but averaged over multiple consecutive chunks. Specifically, the system processes $N = \text{round}((f_s/C) \cdot \alpha)$ chunks, where f_s = 48,000 Hz is the sampling rate, $C = 1536$ samples is the chunk size, and $\alpha = 0.5$ is a temporal factor. This results in $N = \text{round}((48{,}000/1536) \cdot 0.5) = \text{round}(15.625) = 16$ chunks being processed together. Given that each chunk has a duration of $C/f_s = 1536/48{,}000 = 0.032$ s, the total analysis window spans approximately $16 \cdot 0.032 = 0.512$ s, or roughly half a second. The voice activity confidence v is therefore computed as follows:

$$v = \frac{1}{N} \sum_{i=1}^{N} c_i \tag{1}$$

where c_i is the confidence score returned by the Silero VAD model for the i-th chunk.

The mapping from continuous acoustic features to the discrete filtering domain is then defined as follows:

$$L = \text{round}(7 \cdot ((1 - v) \cdot d + k \cdot v)) \tag{2}$$

where $v \in [0, 1]$ is the voice activity confidence defined earlier, $d \in [0, 1]$ is the normalized SPL, where the SPL has been computed as the RMS-based sound pressure level in dBFS over the same 16 chunks where the voice confidence was averaged, and k is a tuning parameter. The SPL computed over the 16 chunks is mapped to the unit interval through min–max normalization: the upper reference value is fixed at 0 dB, while the lower bound is determined empirically for the specific acquisition device. In practice, the minimum reference value corresponds to the SPL measured by the microphone under complete silence conditions, thereby capturing the sensor noise floor. This calibration step ensures that the normalized variable d reflects relative acoustic intensity with respect to the operational dynamic range of the employed microphone.

The formulation can be interpreted as a bounded decision function implementing a trade-off between two competing objectives: environmental noise attenuation and speech preservation. Specifically, the term $(1 - v) \cdot d$ models the expected benefit of attenuation under the assumption that noise suppression is desirable when speech presence is unlikely, while the term $k \cdot v$ introduces a compensatory factor that limits attenuation when voice activity is detected. The expression is therefore a convex combination of noise-driven and speech-driven control components, constrained to the unit interval and subsequently quantized to match the discrete control interface of the DSP. This design enforces three desirable properties for real-time embedded operation: boundedness, monotonicity with respect to noise level, and attenuation reduction in the presence of speech.

The specific parametric form was selected to satisfy these structural constraints while maintaining constant-time computation and numerical stability on resource-limited hardware. The value of k is not theoretically fixed but empirically tuned using the debug dashboard described in Section 6.1.1, to reflect application-specific safety requirements and perceptual tolerances. Calibration was performed by analyzing system responses across representative acoustic scenarios and selecting values that maximize speech intelligibility while maintaining effective noise reduction. The comparative behavior shown in Figure 2 illustrates how different values of k implement distinct operating points along this trade-off curve.

Alternative formulations are possible: adaptive strategies could estimate k online based on contextual statistics of the acoustic environment, while learning-based approaches could replace the analytical mapping with a regression model trained to predict optimal attenuation levels from acoustic features. However, such approaches introduce additional computational cost, training requirements, and reduced interpretability. The proposed formulation was therefore adopted as a deterministic control law that provides predictable behavior, low latency, and explicit tunability suitable for safety-critical embedded deployment.

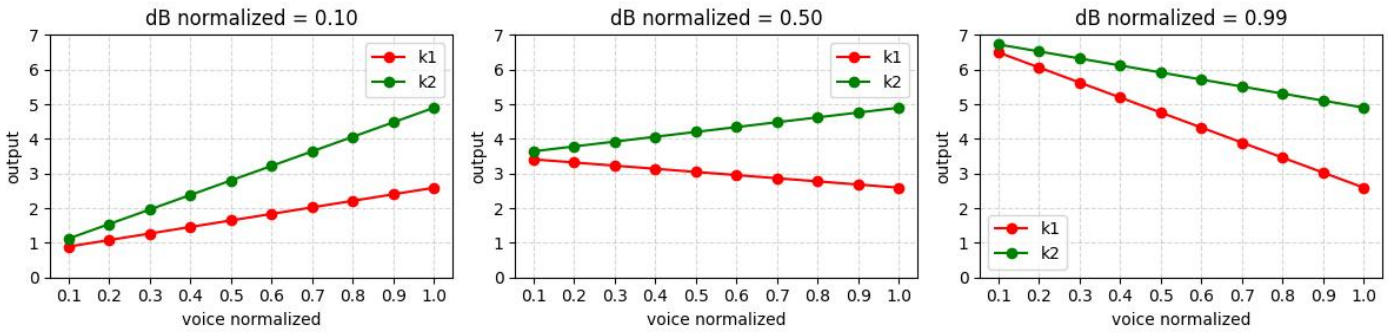

Figure 2. Behavior of the filtering level quantization as a function of the normalized voice activity confidence for different normalized decibel levels. Each subplot corresponds to a fixed noise condition, while the two curves highlight the effect of different tuning parameters ($k_1 = 0.37$ and $k_2 = 0.7$). Lower values of k emphasize speech preservation by reducing the filtering level more aggressively as voice confidence increases, whereas higher values of k favor stronger noise attenuation even in the presence of voice activity.

The quantized filtering level L is then translated into a hardware configuration for the DSPs through a set of GPIO pins. Each filtering level is encoded as a 3-bit value, which is mapped directly to three digital output pins. These pins are mirrored on two independent DSPs, one for each ear cup, ensuring symmetrical and consistent audio processing on both sides. The presence of two separate DSPs is required to maintain true stereo processing and to allow independent but synchronized control of the left and right acoustic channels, which is essential for spatial perception and user comfort. As previously described, to ensure stable auditory perception and prevent rapid oscillations in filtering strength, the system

does not update the DSP configuration on a per-chunk basis. Instead, VAD confidence scores are aggregated over half-second time windows and mapped to one of eight discrete filtering levels. This temporal aggregation strategy smooths short-term fluctuations while maintaining responsiveness to changes in acoustic conditions.

The mapping between the 3-bit configuration and the DSP behavior is defined by the tables reported below. Table 1 shows how the combination of logic levels on the three control pins selects one of the eight available filtering configurations. Table 2 reports the corresponding attenuation values applied by the DSPs in terms of tone reduction and white noise reduction.

Table 1. DSP configuration levels based on control pin states.

Level	N2	N1	N0
1	0 V	0 V	0 V
2	0 V	0 V	+3.3 V
3	0 V	+3.3 V	0 V
4	0 V	+3.3 V	+3.3 V
5	+3.3 V	0 V	0 V
6	+3.3 V	0 V	+3.3 V
7	+3.3 V	+3.3 V	0 V
8	+3.3 V	+3.3 V	+3.3 V

Table 2. Noise reduction characteristics associated with each DSP level.

Level	Tone Reduction	White Noise Reduction
1	4 dB	8 dB
2	5 dB	12 dB
3	6 dB	16 dB
4	8 dB	20 dB
5	16 dB	25 dB
6	21 dB	30 dB
7	25 dB	35 dB
8	65 dB	40 dB

All the audio processing, including VAD inference, SPL computation, and filtering level quantization, is executed on a Raspberry Pi 4. This platform was selected due to its ease of use, extensive software support, and direct access to GPIO pins, which allows seamless interfacing with the electronic board controlling the DSPs. Furthermore, the Raspberry Pi natively integrates both Bluetooth and WiFi modules, which are fundamental for system operation.

The Bluetooth module is used to continuously scan the industrial environment for BLE beacons associated with smart machinery. Each machine is equipped with one or more beacons, and when a beacon is detected, its received signal strength indicator (RSSI) value is compared against a predefined threshold. Depending on whether the RSSI is above or below this threshold, the device subscribes to either a `near` or a `far` MQTT topic on the local central server. From that point on, the Raspberry Pi periodically publishes messages to the subscribed topic, containing its own MAC address and the identifier of the detected machine. This mechanism allows the server to maintain an up-to-date view of worker proximity to machinery.

The WiFi module is used to interact with the backend APIs exposed by the local central server. Through these APIs, the Raspberry Pi retrieves the mapping between beacon MAC addresses and machinery identifiers, as well as the RSSI thresholds associated with each

beacon. This information is required to correctly interpret BLE scan results and to apply machine-specific proximity logic.

The industrial machines considered in this system are smart machines equipped with programmable logic controllers (PLCs) and network connectivity modules [62]. They were simulated using digital twins implemented as Python 3/Qt Creator-based virtual clients running on Raspberry Pi devices, each emulating the operations of the Data Collector Manager (DCM) via HTTP POST requests. Each digital twin can simulate machine operations and complex scenarios in real time.

For physical implementation, a modified CNC lathe was integrated into the system, equipped with a PLC that communicates directly with the system supervisor. The DCM by D.Electron acts as middleware, collecting, formatting, and transmitting data to the backend server via HTTP POST requests. When a malfunction or hazardous condition occurs, the PLC transmits real-time information regarding machine downtime, alarms, control display status, maintenance needs, smart service functionalities, and executed programs. All transmitted data follow a structured format containing the machine's serial number, data type, numeric value, description, and timestamp.

The local central server was implemented using Spring Boot for the backend and Angular for the frontend. In addition to the APIs used by the Raspberry Pi to retrieve beacon configurations, the backend exposes a complete set of services for system management, including visualization and modification of workers, machines, beacons, thresholds, and environmental configurations.

Access to the dashboard is restricted to the safety manager through an authentication mechanism. The dashboard allows real-time monitoring of all system entities and enables the creation and editing of custom messages to be delivered to workers. Each pair of headphones is associated with a dedicated MQTT topic, identified by its MAC address. By correlating MQTT messages published on the `near` topic with the known mapping between workers and headphones, the dashboard can determine which workers are located near a specific machine. The safety manager can then select these workers and send them personalized messages or forward machine-generated alarms.

When a pair of headphones receives an alarm or a customized message, the filtering system is temporarily disabled by toggling a dedicated control pin. During this interval, the system relies exclusively on passive attenuation, ensuring that the synthesized alarm message is reproduced with maximum clarity. The passive acoustic barrier is dimensioned to provide full baseline hearing protection on its own, preventing harmful noise components from reaching the ear even in the absence of active noise reduction. In this operating mode, no detrimental sound energy bypasses the protective attenuation, and the hearing protection performance remains compliant with the measurement criteria defined by the ISO 4869 framework and the product requirements of the EN 352 series. The selected commercial device complies with the EN 352 standard and provides a noise reduction rating (NRR) of 25 dB, which represents a global attenuation parameter characterizing the overall noise reduction capability of the hearing protector across typical exposure conditions.

Once the message playback is completed, the active filtering system is re-enabled and normal operation resumes. This design choice prevents residual noise components from overlapping with the alarm message. In particular, impulsive noises are only partially mitigated by the DSPs, since active noise reduction algorithms require a short temporal window to adapt their filtering response. This limitation arises from the adaptive nature of DSP-based noise reduction, which requires a finite convergence time to estimate the acoustic transfer function and adjust filter coefficients. Under typical industrial conditions, this adaptation process ranges from several hundreds of milliseconds to a few seconds, making the system inherently less effective against highly impulsive or rapidly varying noise

components. Consequently, transient acoustic events may not be sufficiently attenuated and could interfere with the intelligibility of synthesized alarm messages. Moreover, the system is explicitly designed to preserve external voices to maintain situational awareness, which could otherwise interfere with message perception. Temporarily disabling active filtering ensures that neither residual impulsive noise nor intentionally preserved speech masks the alarm, while passive attenuation continues to provide standards-compliant hearing protection and guarantees clear and unambiguous delivery of safety-critical information.

5.2. Data Flow

Industrial environments are typically characterized by the coexistence of high-intensity noise generated by machinery and the voices of workers communicating with each other. This scenario highlights the challenge of achieving effective noise attenuation while preserving situational awareness in hearing protection systems.

The proposed smart headphones primarily rely on passive noise isolation, physically shielding the worker from external sounds. As described in the previous subsection, the device is equipped with three microphones. Two microphones are dedicated to the acquisition of environmental audio, one for each ear, enabling binaural perception and supporting situational awareness, while the third one is specifically used for the VAD algorithm and for the estimation of the SPL expressed in decibels.

Based on the VAD algorithm outputs—namely the estimated probability of speech presence in the acquired audio stream—and the measured SPL, the audio signals captured by the two ear-level microphones are adaptively processed by a DSP-based filter. The filtering parameters are dynamically adjusted in real time to attenuate harmful industrial noise while preserving speech components when vocal activity is detected. The resulting filtered audio stream is then reproduced to the worker through the headphones. In this way, the worker is protected from excessive noise exposure while remaining capable of perceiving colleagues' voices, which may be essential for warnings or emergency communications.

In parallel with audio processing, the smart headphones also support worker localization within the industrial environment. At system startup, each device retrieves the current configuration of the industrial facility from the central server, including information about installed machinery and the associated BLE beacons. When a worker approaches a machinery, the beacon detection module embedded in the hearing protection system identifies the proximity condition based on the RSSI exceeding a predefined threshold. Upon detection, the device publishes a JSON-formatted message to the system-wide MQTT topic `user-near-machinery`. The message payload contains two fields: `userId`, an integer identifying the worker associated with the headphones, and `machineryId`, an integer identifying the specific machinery. This structured format enables the central server to maintain a precise mapping between workers and machinery in real time. Conversely, when the RSSI falls below the threshold, indicating that the worker has moved away from the machinery, the device publishes a similarly structured message to the complementary topic `user-far-machinery`, carrying the same `userId` and `machineryId` fields. Through this bidirectional notification mechanism, the central server, which hosts the MQTT broker, maintains an up-to-date view of workers' positions relative to industrial machinery. The use of JSON encoding ensures interoperability and facilitates integration with other system components or third-party monitoring tools.

In addition to proximity-based topics, the MQTT infrastructure defines a dedicated topic for each headphone, uniquely identified by its serial number. Each device is permanently subscribed to its own topic, enabling direct and individualized communication with the corresponding worker. This mechanism supports the delivery of personalized messages or targeted safety alerts.

When a malfunction or hazardous condition is detected by an industrial machinery, the central server is notified. The safety manager can then forward an alarm to all workers considered to be at risk, namely those located in the proximity of the affected machinery. The alarm message is published to the individual MQTT topics of the relevant headphones as a JSON payload containing two fields: `message`, a string carrying the textual content of the alert, and `language`, a string specifying the language code (e.g., ''`it`'' for Italian, ''`en`'' for English) to ensure proper text-to-speech synthesis. Upon reception of the alarm, the headphones convert the textual notification into synthesized speech using a text-to-speech module. Simultaneously, a control signal is sent to the DSP filter to temporarily mute the reproduction of external audio, even if previously filtered. This ensures that the alarm message is delivered with the highest priority and maximum intelligibility.

The presence of the central server, integrating both backend and frontend components, plays a crucial role in the overall system. It is responsible for managing configuration data, coordinating communication, tracking worker positions, and handling safety notifications. As such, the central server represents the core element upon which the operation, supervision, and configurability of the entire system depend.

The overall data flow and interactions among system components described in this section are summarized in Figure 3.

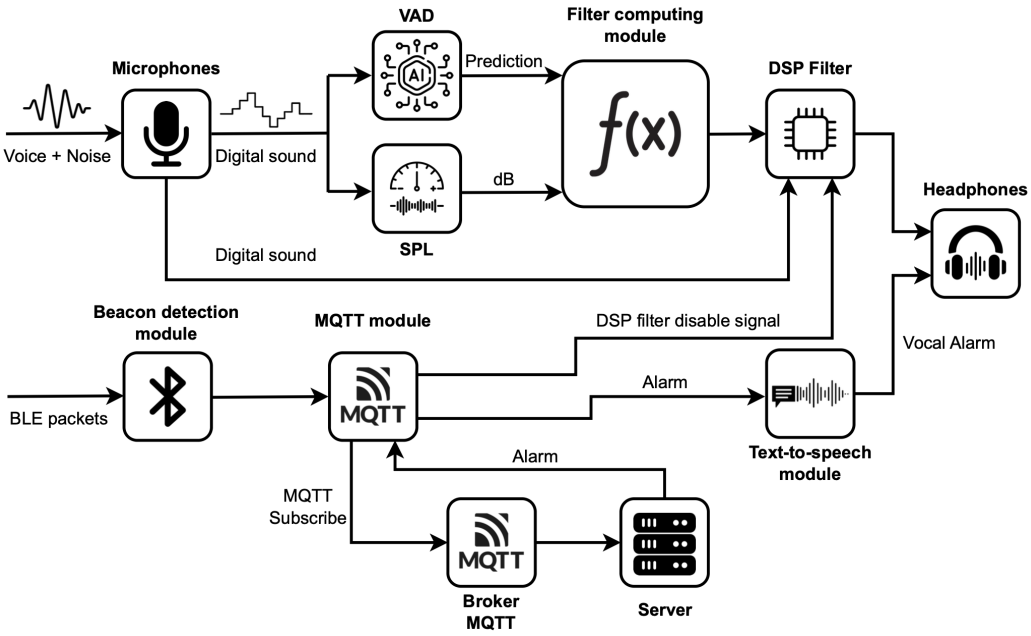

Figure 3. Overview of the overall system data flow.

6. System Validation and Evaluation

The experimental campaign was designed to provide a comprehensive validation of the proposed smart hearing protection system, addressing both algorithmic performance and system-level safety requirements. The validation focused on three main aspects:

- Real-time feasibility and robustness of the VAD-driven adaptive audio processing pipeline.
- Effectiveness of the adaptive noise filtering strategy in preserving speech intelligibility under high industrial noise conditions.
- Responsiveness and reliability of the IoT-based safety notification infrastructure.

All experiments were conducted using a prototype implementation on low-power embedded hardware. Specifically, a Raspberry Pi 4 was used as the wearable processing unit, interfaced with smart headphones equipped with a DSP-based adaptive filter.

Representative industrial acoustic scenarios were reproduced by combining recorded machinery noise, human speech, and simulated alarm events, enabling controlled yet realistic validation conditions.

6.1. Real-Time Audio Processing and VAD Evaluation

The evaluation was conducted on a corpus of audio tracks characterized by heterogeneous signal-to-noise conditions, including continuous industrial noise, intermittent speech, and mixed scenarios with overlapping speech and machinery sounds. All recordings were acquired directly in real industrial environments and on operational machinery. The machines and acoustic contexts included: toolroom ambient environment, inactive press environment, milling machine, uninterruptible power supply unit, hydraulic/pneumatic systems, robotic cell, two presses, shaker, general workshop environment, saw, assembly area environment, lathe, compressed air emission, and handheld grinding machine. Each audio file in the corpus was systematically labeled using a structured naming convention embedding four attributes: (i) machine or acoustic source, (ii) recorder distance from the machine, (iii) speaker distance from the machine, and (iv) speaker identity. This labeling scheme enabled controlled comparative analysis across acoustic configurations and ensured traceability of experimental conditions.

6.1.1. Debug Dashboard

A dedicated debug dashboard, illustrated in Figure 4, was developed to support real-time visualization of both the VAD confidence evolution and the corresponding SPL, enabling qualitative and quantitative inspection of model behavior. The dashboard was designed as a validation and analysis tool for the proposed adaptive filtering system. It allows users to load an arbitrary audio track and process it through the pipeline while providing synchronized visualization of VAD predictions and SPL measurements over time, thereby enabling detailed inspection of their temporal correlation. The interface also provides interactive control over the DSP stage. Users can configure a fixed filtering level and replay the processed audio to directly evaluate the perceptual impact of static noise attenuation. Alternatively, an adaptive filtering mode can be enabled, in which filtering parameters are dynamically adjusted according to Equation (2). In this configuration, users can listen to the audio stream while adaptive filtering is applied, allowing immediate perceptual assessment of system behavior. Crucially, the dashboard enabled systematic tuning of the parameter k defined in Equation (2). By analyzing the filtering effects produced across the labeled industrial audio corpus while varying k, it was possible to compare perceptual quality, noise attenuation effectiveness, and VAD stability under realistic operating conditions. Based on this empirical evaluation, the optimal parameter value for the considered application domain was determined to be $k = 0.37$. Finally, the dashboard supports exporting the resulting curves in PDF format and saving a tabular representation of the computed values in CSV format. The exported CSV includes both Silero VAD outputs and SPL measurements sampled at one-second resolution. An example of stored values for a specific audio file is reported in Table 3. Overall, the debug dashboard played a central role in validating the adaptive filtering approach, providing integrated visual and auditory feedback that enabled systematic evaluation, parameter tuning, and performance verification of the proposed system.

Table 3. Example output of per-second averaged VAD and decibel levels for an audio track.

	1 s	**2 s**	**3 s**	**4 s**	**5 s**	**6 s**	**7 s**
VAD:	0.622	0.965	0.996	0.994	0.983	0.987	0.996
dB:	94.9	98.7	98.7	98.6	96.3	96.7	96.2

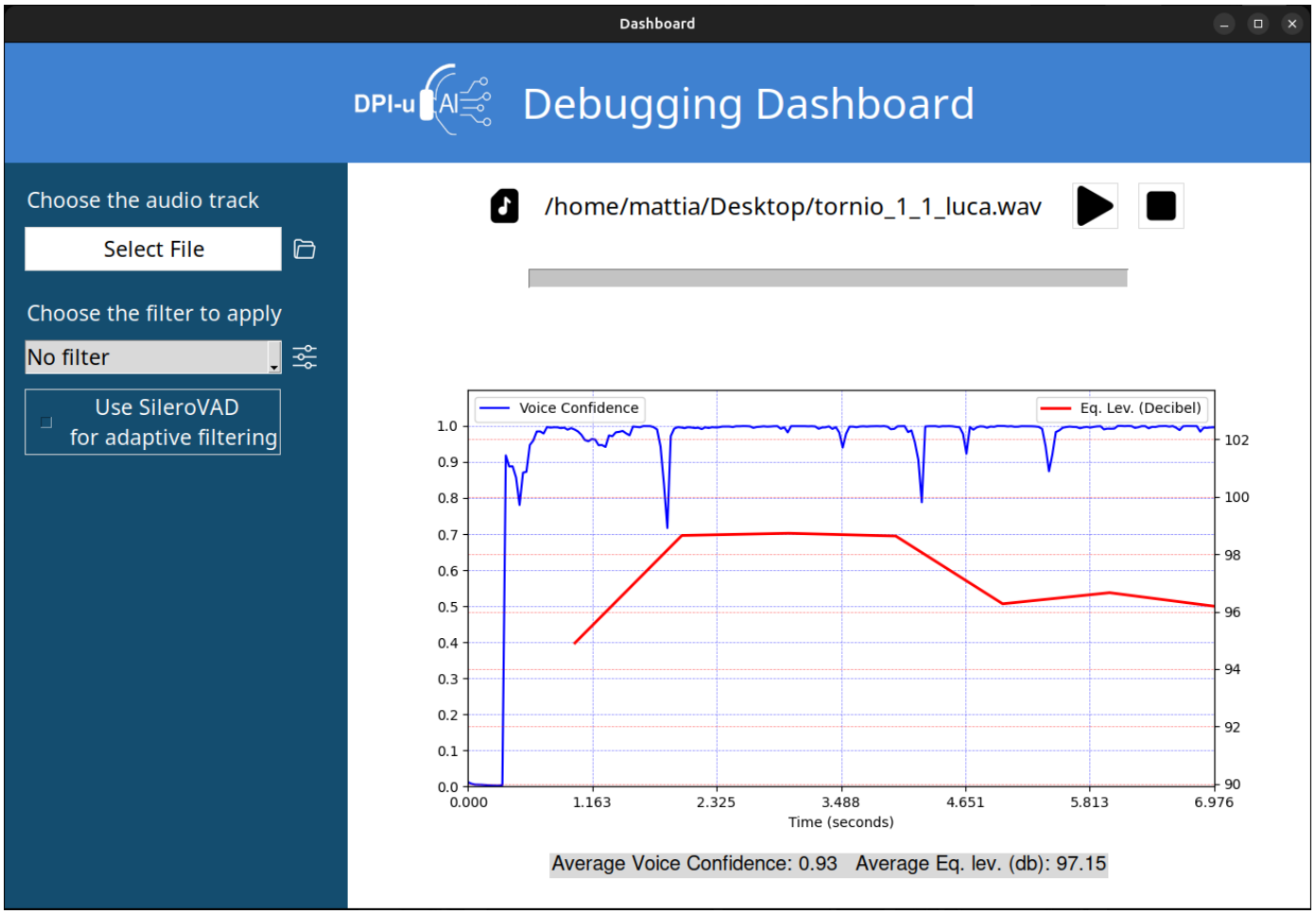

Figure 4. Graphical representation of the audio analysis: Silero VAD output (blue) and decibel-equivalent sound pressure level, SPL (red).

6.1.2. Real-Time Evaluation

The VAD model demonstrated stable performance across all evaluated scenarios, consistently tracking speech activity even under conditions of intense industrial noise. The confidence scores exhibited strong temporal alignment with actual speech segments, enabling reliable discrimination between speech and noise-dominated intervals.

A fundamental requirement of the proposed system is real-time execution on low-power wearable hardware. Accordingly, the processing time of the VAD algorithm was explicitly evaluated on the target platform. Experimental measurements conducted on the Raspberry Pi 4 indicate that Silero VAD processes each 32 ms audio chunk in approximately 2.5 ms, corresponding to a real-time speed (RTS) of about 12.8. This result confirms a substantial computational margin, as the processing time is more than an order of magnitude shorter than the chunk duration. Consequently, the VAD module fully satisfies real-time constraints while preserving sufficient computational resources for additional operations, including audio resampling, DSP control, communication, and dashboard interaction. Overall, these findings validate that the proposed audio processing pipeline is well suited for continuous real-time operation on embedded wearable devices, without introducing perceptible latency or degrading system responsiveness.

6.1.3. Robustness Evaluation

To systematically evaluate the robustness of the Silero VAD model under varying acoustic conditions, a comprehensive validation pipeline was developed to analyze system performance at different signal-to-noise ratios (SNRs).

The evaluation was conducted on a controlled dataset comprising multi-speaker speech audio samples acquired in the presence of industrial noise at three SNR levels: −7 dB, 0 dB, and +10 dB. To ensure a balanced dataset, samples containing exclusively noise (negative class) were also included, with durations corresponding to those of samples containing speech. All audio files were normalized to a uniform peak level (−1.0 dBFS) to prevent discrimination based on absolute amplitude and ensure controlled experimental conditions. The final dataset comprises 2616 samples for each SNR condition, equally distributed between speech+noise samples (1308) and noise-only samples (1308), for a total of 7848 audio files across the three experimental conditions.

The Silero VAD model was evaluated on the complete dataset using 32 ms windows (512 samples at 16 kHz). For each audio file, frame-by-frame confidence scores were aggregated using arithmetic mean to produce a file-level score. For each SNR condition, ROC curves were computed and optimal decision thresholds were selected by maximizing Youden's J statistic ($J = \text{TPR} - \text{FPR}$). Performance metrics include accuracy, precision, recall, F1-score, ROC AUC, and PR AUC, calculated at the optimal threshold.

Table 4 summarizes VAD performance under different SNR conditions. The results show expected behavior: performance is excellent at high SNR (+10 dB), with perfect accuracy, precision, and recall (1.0) and an ROC AUC equal to 1.0. At 0 dB, the system maintains very high performance, with 98.3% accuracy, 99.1% precision, 97.5% recall, and ROC AUC of 0.998. The most critical condition is represented by −7 dB SNR, where accuracy drops to 79.8%, precision stands at 88.3%, and recall at 68.7%, with an ROC AUC of 0.878 and a PR AUC of 0.901.

Table 4. VAD performance metrics across different SNR conditions.

SNR Level	Optimal Threshold	Accuracy	Precision	Recall	F1-Score	ROC AUC	PR AUC
−7 dB	0.060	0.798	0.883	0.687	0.772	0.878	0.901
0 dB	0.124	0.983	0.991	0.975	0.983	0.998	0.998
+10 dB	0.786	1.000	1.000	1.000	1.000	1.000	1.000

The ROC curves (Figure 5) and precision–recall curves (Figure 6) clearly visualize the progressive performance degradation as SNR decreases. It is interesting to note how the optimal threshold varies significantly between conditions: from 0.060 at −7 dB to 0.786 at +10 dB, suggesting the need for dynamic threshold adaptation based on operational acoustic conditions.

The confusion matrices for each SNR condition (Figures 7–9) provide further details on error distribution. At −7 dB (Figure 7), most errors consist of false negatives (undetected speech), with 410 vocal samples erroneously classified as noise, consistent with the reduced recall (68.7%). This behavior is typical of VAD systems under extreme noise conditions, where the speech signal is masked by background noise. At 0 dB (Figure 8), errors are drastically reduced, with only 33 false negatives and 12 false positives, demonstrating excellent discriminative capability. At +10 dB (Figure 9), classification is perfect, with no errors on either class.

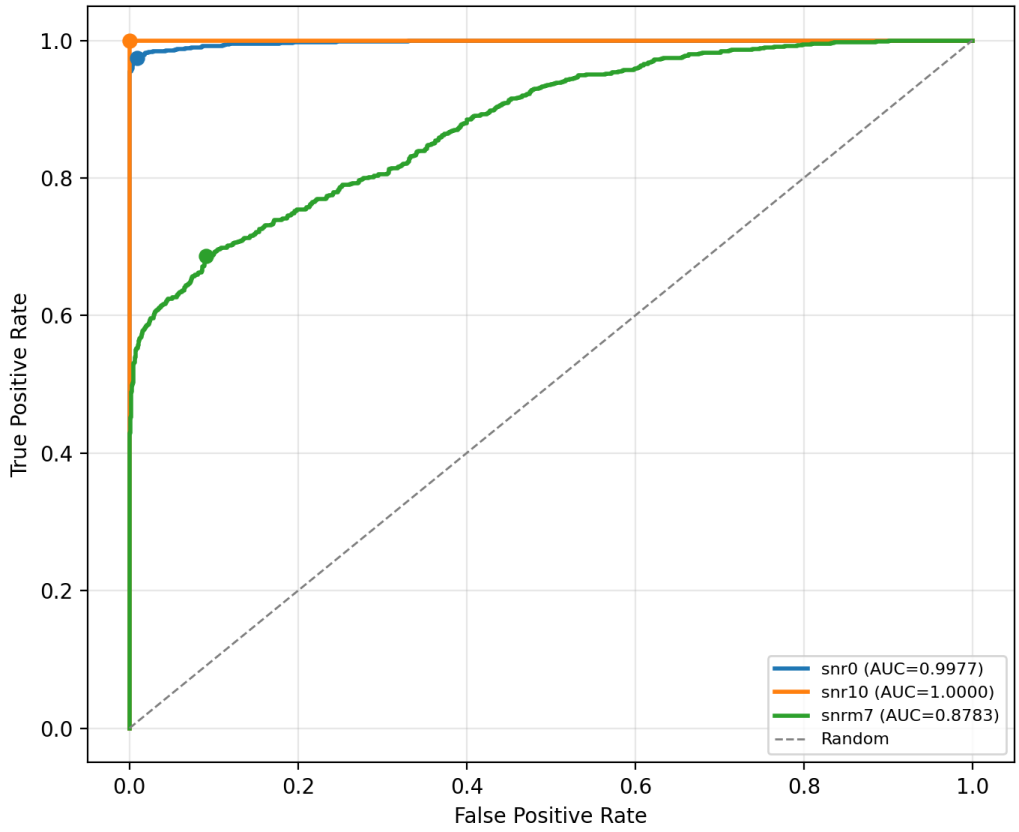

Figure 5. ROC curves for Silero VAD under three SNR conditions. Markers indicate optimal operating points selected via Youden's J statistic.

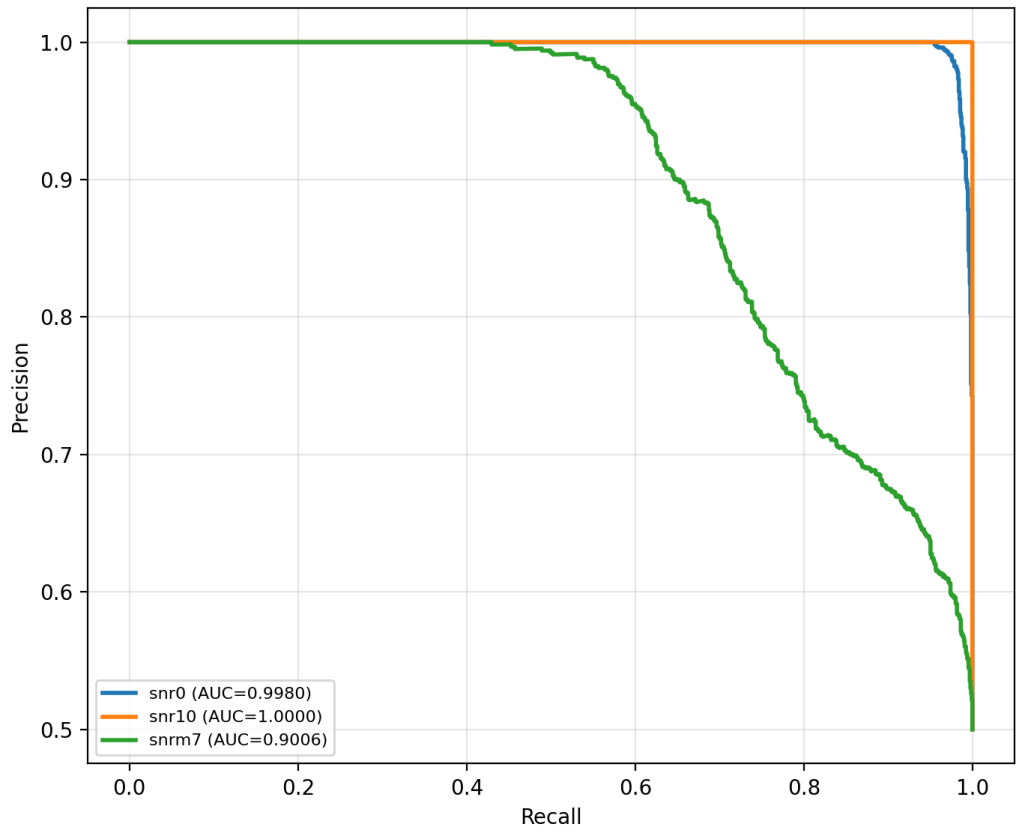

Figure 6. Precision–recall curves for the three tested SNR conditions.

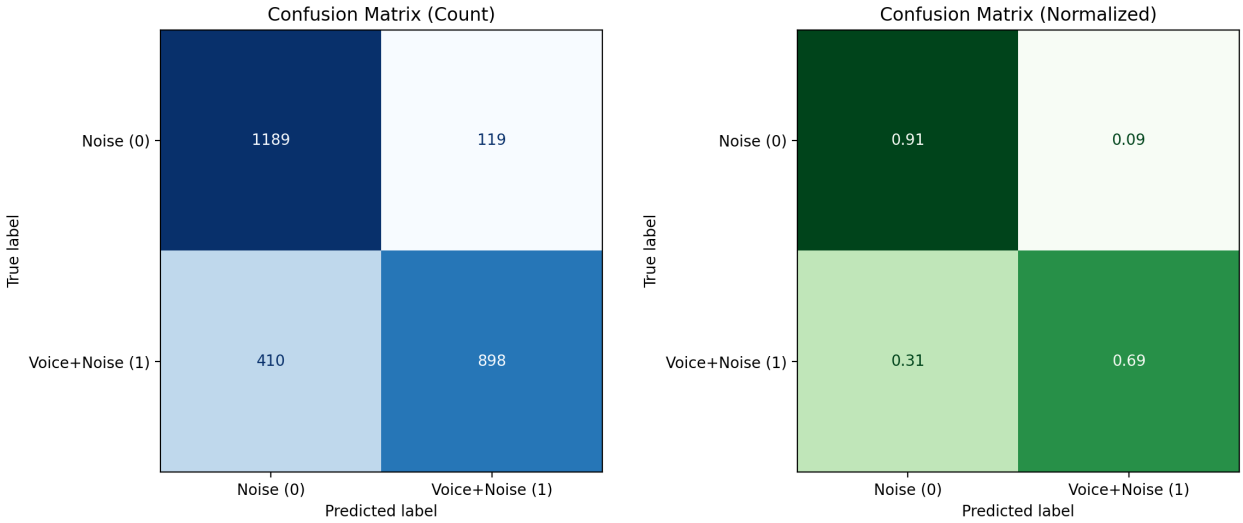

Figure 7. Confusion matrices for SNR −7 dB condition. **Left**: absolute counts; **right**: normalized values.

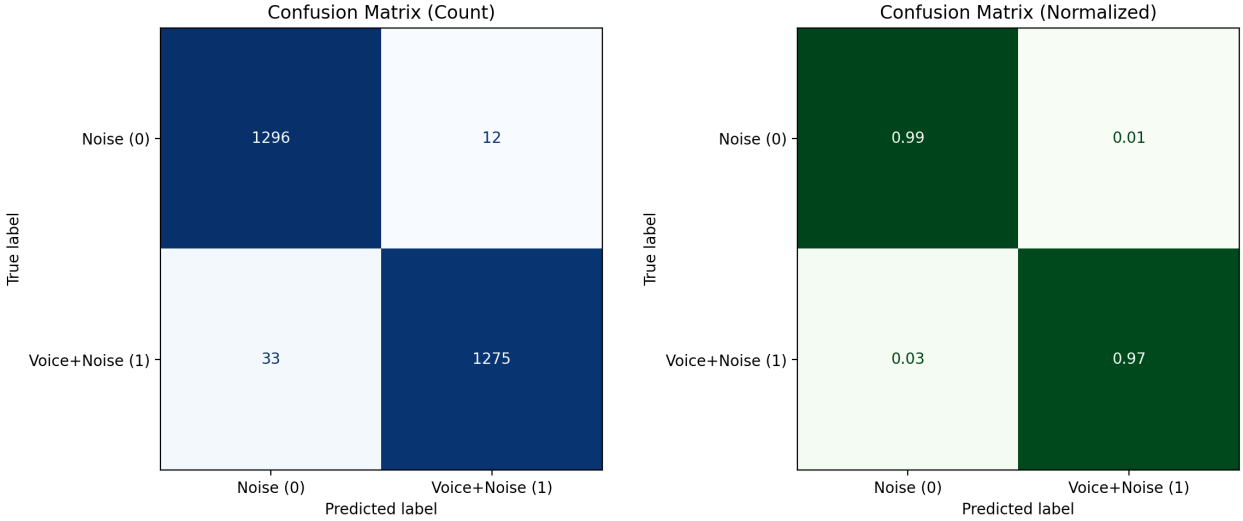

Figure 8. Confusion matrices for SNR 0 dB condition.

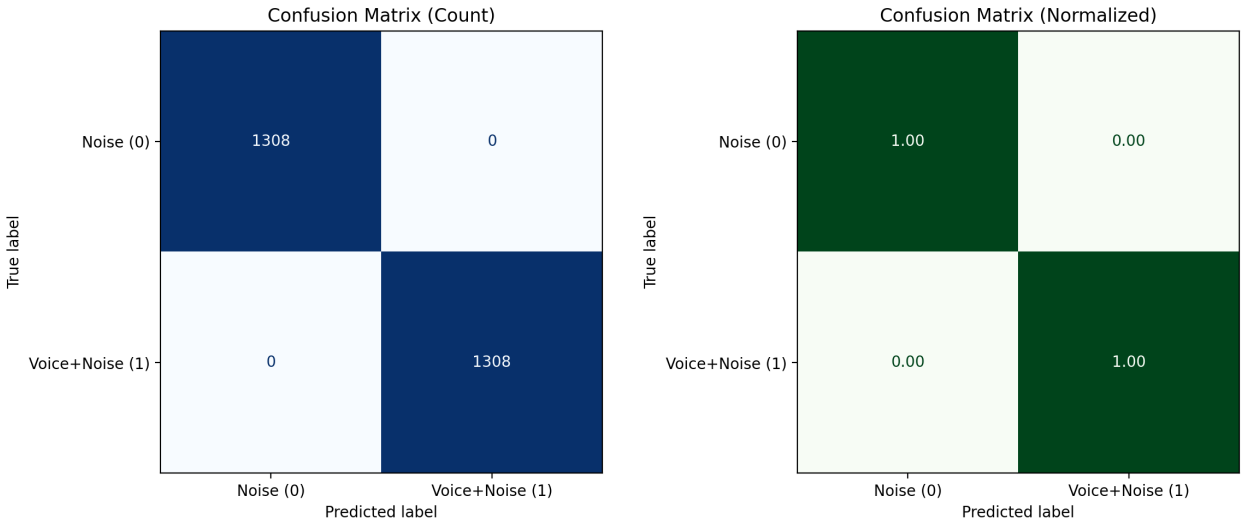

Figure 9. Confusion matrices for SNR +10 dB condition.

6.2. Validation of Adaptive Noise Filtering

A second experimental phase investigated the interaction between the VAD output and the DSP-based adaptive noise filtering mechanism. Based on the estimated speech activity, the system dynamically selected the appropriate filtering level to balance noise attenuation and speech intelligibility. During intervals dominated by background noise, higher attenuation levels were applied to maximize hearing protection. Conversely, when speech was detected, the system reduced filtering strength to preserve voice clarity and situational awareness.

6.2.1. Objective Quality Metrics Assessment

To quantitatively validate the effectiveness of the adaptive filtering system, a comprehensive evaluation was conducted using three established objective quality metrics: Mean Opinion Score (MOS), signal-to-noise ratio (SNR), and Harmonics-to-Noise Ratio (HNR). These metrics provide complementary perspectives on speech quality and intelligibility. The benchmark evaluation utilized a diverse set of audio recordings featuring a speaking subject with various background noise conditions, including both industrial and generic acoustic environments:

- Industrial machinery noise: lathe operation, electric saw, jackhammer;
- Generic noise profiles: white noise, pink noise.

Each audio sample was processed in two conditions: without filtering and with adaptive filtering, enabling direct comparison of the system's performance across different acoustic scenarios. In order to provide a comprehensive understanding of system performance, the validation was conducted using the following objective and subjective evaluation metrics:

- **MOS**: This represents a standardized measure of perceived speech quality, traditionally obtained through subjective listening tests where human evaluators rate quality on a scale from 1 (bad) to 5 (excellent). In this evaluation, MOS was computed using the ITU-T P.563 standard [63], which provides an objective, single-ended method for predicting speech quality without requiring a reference signal. The P.563 algorithm analyzes multiple perceptual dimensions including:
  - Vocal tract characteristics and naturalness;
  - Speech level consistency and temporal structure;
  - Spectral clarity and distortion artifacts;
  - Background noise characteristics;
  - Interruptions and unnatural silence patterns.

  The algorithm combines these features through a non-linear mapping function developed by the ITU-T, which aggregates the individual perceptual dimensions into a single quality score. This mapping function was calibrated using extensive subjective test databases containing thousands of speech samples rated by human listeners, establishing the relationship between objective measurements and perceived quality. The resulting model produces MOS predictions that correlate with human perception across diverse degradation conditions.
- **SNR**: This quantifies the ratio between the speech signal power and the background noise power, expressed in decibels (dB):

$$\mathrm{SNR} = 10 \times \log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \tag{3}$$

  Higher SNR values indicate cleaner speech with less noise interference. The P.563 framework estimates SNR by distinguishing speech-active regions from noise-only segments, computing the power differential between these intervals. This metric directly reflects the filtering system's ability to attenuate background noise while preserving speech energy.

- **HNR**: This measures the ratio between periodic (harmonic) and aperiodic (noise) components in voiced speech segments, expressed in decibels:

$$\text{HNR} = 10 \times \log_{10}\left(\frac{P_{\text{harmonics}}}{P_{\text{noise}}}\right) \quad (4)$$

This metric is particularly sensitive to voice quality degradation, as speech production naturally generates strong harmonic structure during phonation. Higher HNR values indicate clearer, more natural-sounding speech. The calculation involves autocorrelation analysis of voiced frames to separate periodic signal components from residual noise.

6.2.2. Experimental Results

The objective metrics demonstrate substantial improvements across all test conditions following adaptive filtering application. Tables 5–7 present the comparative results for each metric:

- **MOS Results**: The MOS improvements are particularly striking in high-noise scenarios. The jackhammer and pink noise conditions, which initially scored at the minimum MOS value of 1.0 (indicating severely degraded quality), showed dramatic improvements of 104% and 118% respectively after filtering. The "Lathe2" recording achieved the highest post-filtering MOS of 2.903, representing an 84.7% improvement. Even in the challenging white noise condition, which exhibited the smallest relative gain, the filtering system still provided measurable quality enhancement.
- **SNR Results**: The SNR measurements reveal exceptional noise reduction performance. The most remarkable improvement occurred in the lathe recording, where SNR increased by 29.03 dB—representing a noise power reduction of approximately 800-fold. All test conditions showed substantial SNR gains ranging from 10.44 dB to 29.03 dB. These improvements translate to significantly enhanced speech intelligibility, as SNR increases of this magnitude move speech from barely intelligible to clearly understandable ranges.
- **HNR Results**: HNR analysis confirms that the adaptive filtering preserves and enhances voice quality characteristics. The pink noise condition showed the most dramatic improvement (+6.22 dB), recovering from a negative HNR value (indicating noise dominance over harmonics) to a positive value, reflecting clear harmonic structure. All filtered samples achieved HNR values above 4 dB, indicating well-preserved voice naturalness. The saw recording reached the highest HNR of 6.451 dB post-filtering, demonstrating excellent harmonic clarity despite the challenging acoustic environment.

Table 5. Mean Opinion Score (MOS) comparison between unfiltered and filtered audio samples.

Audio Sample	MOS (No Filter)	MOS (Filter)	Improvement
Jackhammer	1.000	2.042	+104.2%
Lathe	1.983	2.627	+32.5%
Pink Noise	1.000	2.181	+118.1%
Saw	2.079	2.863	+37.7%
Lathe2	1.572	2.903	+84.7%
White Noise	1.364	1.512	+10.8%

Table 6. Signal-to-noise ratio (SNR) comparison between unfiltered and filtered audio samples.

Audio Sample	SNR (No Filter) [dB]	SNR (Filter) [dB]	Improvement [dB]
Jackhammer	8.628	19.069	+10.44
Lathe	6.042	35.074	+29.03
Pink Noise	6.354	20.451	+14.10
Saw	8.614	20.539	+11.93
Lathe2	8.131	20.010	+11.88
White Noise	5.083	16.519	+11.44

Table 7. Harmonics-to-Noise Ratio (HNR) comparison between unfiltered and filtered audio samples.

Audio Sample	HNR (No Filter) [dB]	HNR (Filter) [dB]	Improvement [dB]
Jackhammer	2.416	5.760	+3.34
Lathe	2.843	5.537	+2.69
Pink Noise	−1.907	4.310	+6.22
Saw	2.611	6.451	+3.84
Lathe2	1.841	5.533	+3.69
White Noise	1.000	4.796	+3.80

6.3. IoT-Based Alarm Delivery and System Responsiveness

In addition to audio processing, the experimental campaign evaluated the performance of the safety notification subsystem. The objective was to measure the end-to-end latency required for an alarm generated by industrial machinery to reach the headphones of a worker located in the corresponding risk area. It is important to note that, due to existing regulations, the safety manager is responsible for forwarding the alert, and therefore the evaluation focuses on the system-induced delays once the alert has been issued.

Alarm events were generated by simulated machinery malfunctions and propagated through the backend system. Workers' wearable devices subscribed to the relevant MQTT topics based on proximity information derived from BLE beacons. For each alarm, timestamps corresponding to alarm forwarding and reception at the wearable device were recorded, enabling precise computation of transmission delays.

It is worth clarifying the methodological implications of employing simulated machinery malfunctions. The simulation layer was designed to replicate the exact communication behavior of the corresponding real industrial devices. In particular, the structure, payload format, timing characteristics, and protocol-level properties of the messages transmitted to the backend infrastructure are identical to those produced by the physical machines during normal operation. Consequently, from the perspective of the IoT communication pipeline, the alarm generation process is functionally indistinguishable from a real deployment scenario. Therefore, the use of simulated malfunction events does not introduce bias in the evaluation of system responsiveness. Since the measured metric concerns end-to-end transmission delay from alarm issuance to wearable reception, and since this process depends exclusively on network communication, message brokering, and device subscription mechanisms, the validity of the results is not affected by whether the triggering condition originates from a physical fault or a faithfully emulated one. The simulation was adopted solely to ensure repeatability, safety, and controlled experimental conditions, without altering the behavior of the underlying communication infrastructure.

Table 8 reports some experimental results, assuming that the safety manager forwards the alarm immediately upon receiving it, without introducing additional delays.

Table 8. Alarm transmission times and reception delays.

Alarm	Sent	Received	Delay (s)
1°	20:26:57.294	20:26:58.916	1.62
2°	20:28:34.623	20:28:36.205	1.58
3°	20:30:12.851	20:30:14.589	1.73
4°	20:31:43.532	20:31:45.344	1.81
5°	20:32:15.747	20:32:17.082	1.34
6°	20:35:08.928	20:35:10.371	1.45
7°	20:36:56.196	20:36:57.959	1.76
8°	20:38:24.629	20:38:26.267	1.64
9°	20:40:10.104	20:40:11.523	1.41
10°	20:43:29.364	20:43:31.315	1.95

Alarm delivery experiments showed that the average end-to-end delay between alarm generation and reception at the worker's wearable device was approximately 1.6 s, with limited variability across trials. This response time is compatible with industrial safety requirements, ensuring timely notification without compromising system reliability.

The results confirm the effectiveness of combining BLE-based proximity detection with MQTT-based publish/subscribe communication for selective and efficient dissemination of safety alerts.

6.4. Functional Validation via Dashboard Interaction

Finally, the experimental workflow included functional validation of the integrated system through interaction with the dashboard. During tests, the dashboard was used to monitor machinery status, visualize worker proximity, inspect alarm notifications, and verify correct message routing to at-risk users.

This validation phase ensured that all system components—audio processing, localization, communication, and notification delivery—operated cohesively, confirming correct end-to-end behavior in realistic operating conditions.

The dashboard screenshots provide functional validation of the system by highlighting its main operational features. Figure 10 shows the main dashboard interface, which displays the list of registered machinery and their current status. Figure 11 illustrates the reception of an alarm notification generated by the system following the detection of a hazardous event (both are messages generated on the machinery side). Finally, Figure 12 presents the interface used to forward the alarm message to workers identified as being at risk, enabling timely notification and intervention.

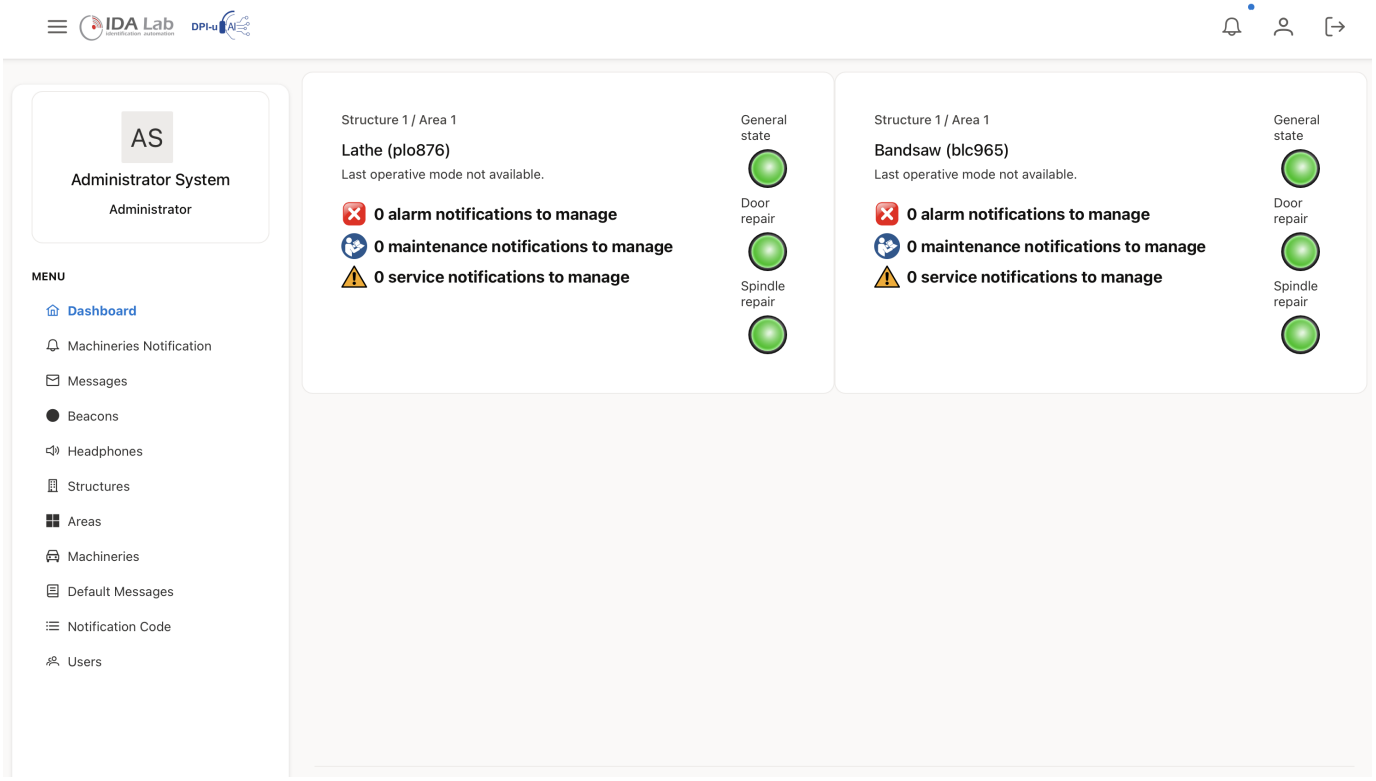

Figure 10. Dashboard main screen.

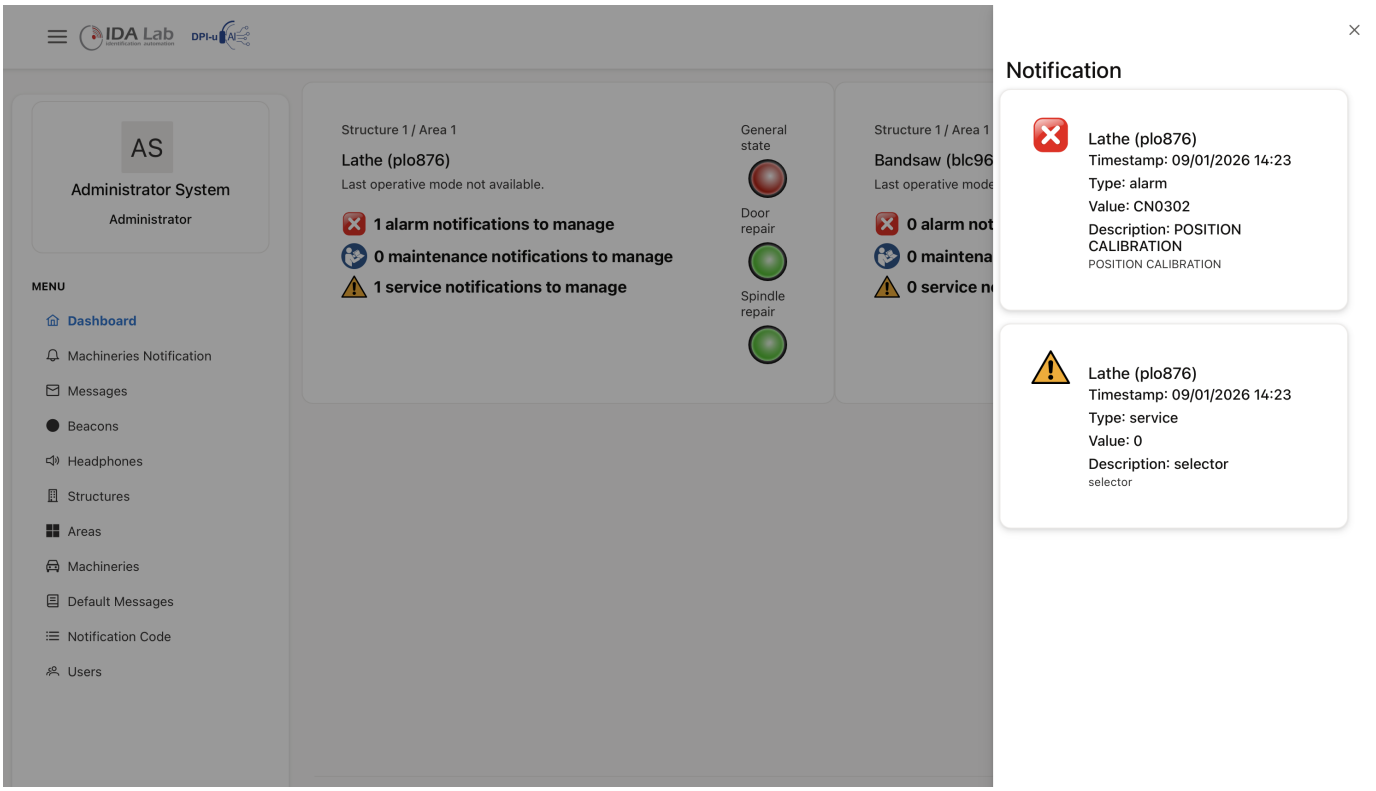

Figure 11. Alarm notification.

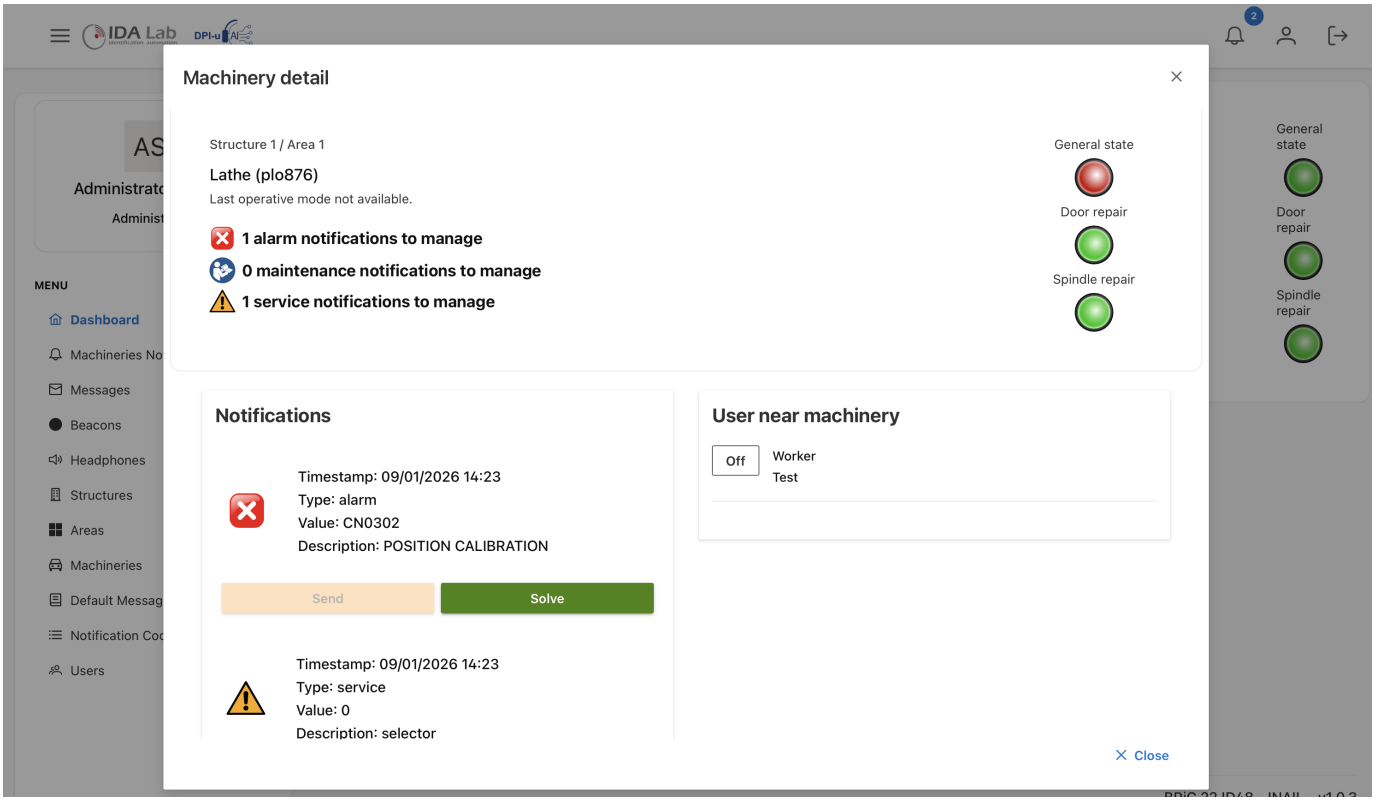

Figure 12. Send alarm to at-risk worker.

The integrated validation demonstrated that all system components operate coherently within a unified architecture. The combination of adaptive hearing protection, real-time speech detection, worker localization, and targeted safety notifications confirms the feasibility of the proposed solution as an intelligent hearing PPE for noisy industrial environments.

7. Discussion

The proposed system addresses a critical limitation of conventional hearing protection solutions used in industrial environments, namely the trade-off between effective noise attenuation and the preservation of situational awareness. Traditional passive or fully suppressive active noise reduction approaches tend to isolate workers from their surroundings, potentially obscuring important auditory information such as spoken warnings, alarms, or cues generated by nearby machinery. The architecture presented in this work overcomes this limitation by introducing an adaptive and context-aware hearing protection mechanism driven by AI-based voice detection.

As already highlighted in the previous sections, a key strength of the proposed approach lies in the integration of a VAD model within the audio processing pipeline: we adopted a VAD-driven DSP architecture rather than deep learning models like DAEs for two main reasons. First, training robust models requires massive industrial datasets that are not publicly available; in fact, while we collected real data in real industrial environments, it was only sufficient for validation: training any ML model based on audio files requires recording millions of hours of noise coming from a hypothetically infinite number of machines. Second, complex neural networks introduce significant latency on low-power hardware, while our "lightweight" approach ensures near-zero delay, providing the instantaneous response required for real-time worker safety. By continuously estimating the probability of speech presence and combining this information with measured SPLs, the system dynamically adjusts the degree of noise attenuation applied by the DSPs. The design defined by Equation (2) enables aggressive noise reduction in the absence of speech, while progressively relaxing filtering when vocal activity is detected, thus preserving intelligibility.

The parameterization of Equation (2) reflects two additional design choices. First, the variable d is obtained by min–max normalization of the measured SPL, with 0 dB as the upper reference and the microphone noise floor, measured under silence conditions, as the lower bound, ensuring device-aware scaling of acoustic intensity. Second, the parameter k is empirically tuned through controlled evaluation across representative acoustic scenarios to balance noise attenuation and speech preservation while maintaining stable real-time behavior on resource-constrained hardware.

The effectiveness of this adaptive filtering strategy has been quantitatively validated through objective quality metrics. The comprehensive evaluation using MOS, SNR, and HNR measurements demonstrates that the system successfully achieves its dual objectives of hearing protection and speech intelligibility preservation. SNR improvements exceeding 10 dB in all test cases directly confirm the noise reduction capability, while MOS enhancements indicate that these technical improvements translate to perceptually meaningful quality increases. Moreover, the HNR results demonstrate that the filtering process does not introduce artifacts that degrade voice naturalness; instead, by removing noise that masks harmonic structure, the system enhances the clarity of speech characteristics. The consistent improvements across diverse acoustic conditions, including industrial machinery noise and generic noise profiles, validate the robustness of the VAD-driven adaptation strategy.

The results demonstrate that the Silero VAD model maintains excellent performance under moderate and high-SNR conditions ($\geq$0 dB), with ROC AUC greater than 0.99. Even in the most unfavorable condition (-7 dB), the system achieves an ROC AUC of 0.878, indicating a still significant discriminative capability. However, the reduced recall (68.7%) at -7 dB suggests that in extremely noisy environments it may be necessary to adopt more conservative decision thresholds.

From an implementation perspective, the use of low-cost embedded hardware, such as the Raspberry Pi, demonstrates the feasibility of deploying AI-driven hearing protection systems without relying on high-performance computing resources or cloud infrastructure. Local processing ensures low latency and increased reliability, which are essential requirements for safety-critical industrial applications. The experimental results obtained from real-time inference tests confirm that the selected VAD model is computationally efficient and suitable for continuous operation on embedded platforms.

Another relevant contribution of this work is the integration of the smart headphones into a broader IoT-based safety infrastructure. Through BLE-based localization and MQTT communication, the system enables context-aware safety notifications, ensuring that alarm messages are delivered only to workers who are actually exposed to a specific hazard. This targeted communication strategy reduces cognitive overload and improves the effectiveness of safety interventions compared to traditional broadcast alarm systems.

Despite these advantages, some limitations remain. The evaluation presented in this work focuses primarily on system feasibility, architectural design, and real-time performance, while large-scale experimental validation in real industrial environments is left for future work. In addition, although the system preserves speech intelligibility, the perception of other relevant non-speech sounds, such as atypical machinery noises that may indicate incipient failures, is not explicitly addressed and could be investigated in future extensions. Moreover, the authors are working on a new version of the developed system that will better handle impulsive noise in future work. In fact, as expected, experiments highlighted that the efficacy of DSP-based noise filtering can be challenged by this kind of noise.

In addition, it is worth mentioning the evidence that emerged from the system regarding the lack of a standardized, publicly available benchmark dataset specifically tailored for high-noise industrial scenarios. In fact, as emerged especially from the state of the art, most existing intelligent PPE solutions are either proprietary commercial products with closed architectures or research prototypes that rely on specialized, non-public datasets. Therefore, the absence of a standardized, publicly available benchmark dataset specifically tailored for high-noise industrial scenarios, combining human speech with non-stationary machinery noise, prevents the presentation in the literature of works similar to the one proposed in this paper and also limits a fair "head-to-head" comparison of objective metrics like MOS or SNR.

8. Conclusions

This paper presented a smart industrial safety system designed to improve hearing protection and safety communication in high-noise industrial environments through the combined use of IoT and AI technologies. The proposed solution introduces intelligent hearing protection devices capable of dynamically adapting noise attenuation based on real-time VAD and environmental SPLs. By selectively preserving human speech while attenuating harmful noise, the system effectively addresses the long-standing trade-off between hearing protection and situational awareness.

The integration of the smart headphones within a localized IoT infrastructure enabled additional safety functionalities, including worker localization and the delivery of targeted audio notifications in response to machinery malfunctions or hazardous conditions. The adoption of lightweight communication protocols and on-device AI inference ensured low latency and reliable operation, making the system suitable for safety-critical industrial contexts. Experimental results demonstrate that the selected VAD model meets real-time performance requirements on embedded hardware and supports stable and responsive adaptive filtering behavior.

Overall, the proposed architecture represents a promising step towards more intelligent, context-aware personal protective equipment aligned with the principles of Industry 4.0 and Society 5.0.

Despite the encouraging results, some limitations remain. In particular, the adaptive DSP filtering exhibits reduced effectiveness when dealing with highly impulsive or rapidly varying noise, due to the finite convergence time required by the adaptive filtering process. Moreover, the lack of standardized public datasets specifically designed for high-noise industrial environments limits large-scale comparative evaluations and hinders direct benchmarking against alternative approaches.

In addition, although speech quality improvements were assessed using the objective ITU-T P.563 model to estimate the MOS and provide a reliable approximation of perceptual quality, the work highlighted the need for subjective listening tests, which the authors are already working on as future work, with the aim of involving human evaluators to further validate the perceived speech intelligibility improvements.

Additionally, other future work will involve extensive field testing in real industrial environments. Because industrial noise is often stationary and structured, future approaches will exploit characteristic noise profiles for more precise adaptive filtering, with large-scale deployments across diverse settings assessing these optimizations, including the influence of site-specific acoustics and task-dependent communication. Future work will also evaluate energy consumption and introduce software optimizations to increase computational efficiency and battery life. Moreover, while Silero VAD was selected as the most suitable model for the proposed architecture based on the analyzed literature and its proven efficiency in real-time embedded applications, future research will aim to provide a more comprehensive quantitative validation. Specifically, a performance comparison against other established baseline methods, such as rVAD and WebRTC VAD, will be conducted under the same experimental conditions to further benchmark their effectiveness specifically within challenging industrial settings. A current limitation is the finite convergence time of adaptive DSP noise reduction, which reduces effectiveness against transient or highly impulsive noise. Addressing this through faster adaptation or hybrid filtering will be a primary objective to improve robustness in smart industrial hearing protection systems. In particular, future developments will investigate the integration of a fast-response protective mechanism capable of detecting sudden high-energy acoustic events and temporarily bypassing active amplification or filtering stages. When a predefined amplitude threshold is exceeded, the system could operate in a protective fallback mode in which active processing is momentarily disabled and the device relies exclusively on its passive attenuation, which already guarantees baseline hearing protection.

Author Contributions: Conceptualization, A.B., L.C., M.C., R.D.S., C.G., M.J., L.L., R.M., T.M., F.P., L.P., D.R., F.A.S. and I.S.; methodology, M.C., L.L., T.M., F.P., L.P. and D.R.; validation, M.C., L.L., T.M., F.P., L.P. and D.R.; investigation, M.C., L.L., T.M., F.P., L.P. and D.R.; data curation, M.C. and D.R.; writing—original draft preparation, M.C., L.L., T.M., F.P., L.P. and D.R.; writing—review and editing, A.B., L.C., M.C., R.D.S., C.G., M.J., L.L., R.M., T.M., F.P., L.P., D.R., F.A.S. and I.S.; visualization, M.C., L.L., T.M., F.P., L.P. and D.R.; software, M.C. and D.R.; supervision, L.C., C.G., L.L., R.M. and L.P.; funding acquisition, A.B., L.C., L.L., L.P. and F.A.S.; Project administration, A.B., L.C., L.L., R.M., L.P. and F.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Istituto Nazionale Assicurazione Infortuni sul Lavoro (INAIL) (Italy) (https://www.inail.it) within the project "DPI-u AI-Ingegnerizzazione di un dispositivo di protezione individuale uditivo intelligente tramite algoritmi di intelligenza artificiale integrato con realtà aumentata e architettura IoT" through the Call for Project "Bando BRiC. Bando Ricerche in Collaborazione (BRiC–2022)–Piano Attività di Ricerca 2022–2024".

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to the data are audio recordings of a real industrial environment. It was done in a real industry, therefore, the data cannot be published.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siskova, V.; Juricka, M. The effect of sound on job performance. In *Proceedings of the 2013 IEEE International Conference on Industrial Engineering and Engineering Management*; IEEE: Piscataway, NJ, USA, 2013; pp. 1679–1683. [CrossRef]
2. Society 5.0. Available online: https://tech4future.info/societa-5-0-super-smart-society/ (accessed on 5 March 2026).
3. Ding, T.; Yan, A.; Liu, K. What is noise-induced hearing loss? *Br. J. Hosp. Med.* **2019**, *80*, 525–529. [CrossRef]
4. Raja, S.; Ganguly, T. Impact of exposure to noise on the hearing acuity of employees in a heavy engineering industry. *Indian J. Med. Res.* **1983**, *78*, 100–113.
5. Nandi, S.S.; Dhatrak, S.V. Occupational noise-induced hearing loss in India. *Indian J. Occup. Environ. Med.* **2008**, *12*, 53–56. [CrossRef]
6. Chen, K.H.; Su, S.B.; Chen, K.T. An overview of occupational noise-induced hearing loss among workers: Epidemiology, pathogenesis, and preventive measures. *Environ. Health Prev. Med.* **2020**, *25*, 65. [CrossRef]
7. Bhattacharya, S.K.; Saiyed, H.N.; Roy, A.; Chatterjee, S.K. Hearing acuity in weavers of a textile mill. *Indian J. Med. Res.* **1981**, *74*, 779–785. [PubMed]
8. Bhattacharya, S.K.; Tripathi, S.R.; Kashyap, S. A study of heat and noise problems in a drug and pharmaceutical firm in India. *Ind. Health* **1990**, *28*, 203–207. [CrossRef]
9. Agarwal, G.; Nagpure, P.S.; Gadge, S.V. Noise Induced Hearing Loss in Steel Factory Workers. *Int. J. Occup. Saf. Health* **2016**, *4*, 34–43. [CrossRef]
10. World Health Organization. Occupational Safety and Health Administration. Available online: https://www.osha.gov/noise (accessed on 5 March 2026).
11. Song, Y. Active Noise Cancellation and Its Applications. *J. Phys. Conf. Ser.* **2022**, *2386*, 012042. [CrossRef]
12. Untwale, A.N.; Degaonkar, K.S. Survey on noise cancellation techniques of speech signal by adaptive filtering. In *Proceedings of the 2015 International Conference on Pervasive Computing (ICPC)*; IEEE: Piscataway, NJ, USA, 2015; pp. 1–4. [CrossRef]
13. Blessy, J.; Christopher, C.S. A Survey on Filtering Methods Used to Remove Noise in Speech and Music Signal. In *Proceedings of the 2022 6th International Conference on Electronics, Communication and Aerospace Technology*; IEEE: Piscataway, NJ, USA, 2022; pp. 140–145. [CrossRef]
14. Corallo, A.; Crespino, A.M.; Del Vecchio, V.; Gervasi, M.; Lazoi, M.; Marra, M. Evaluating maturity level of big data management and analytics in industrial companies. *Technol. Forecast. Soc. Change* **2023**, *196*, 122826. [CrossRef]
15. Madonna, M.; Monica, L.; Anastasi, S.; Di Nardo, M. Evolution of cognitive demand in the human–machine interaction integrated with industry 4.0 technologies. *Wit Trans. Built Environ.* **2019**, *189*, 13–19. [CrossRef]
16. Fadda, E.; Perboli, G.; Rosano, M.; Mascolo, J.E.; Masera, D. A Decision Support System for Supporting Strategic Production Allocation in the Automotive Industry. *Sustainability* **2022**, *14*, 2408. [CrossRef]
17. *ISO 7731:2003*; Ergonomics–Danger Signals for Public and Work Areas–Auditory Danger Signals. International Organization for Standardization: Geneva, Switzerland, 2003. Available online: https://www.iso.org/standard/38617.html (accessed on 5 March 2026).
18. *ISO 11429:1996*; Ergonomics–System of Auditory and Visual Danger Signals for Public and Work Areas. International Organization for Standardization: Geneva, Switzerland, 1996. Available online: https://www.iso.org/standard/19369.html (accessed on 5 March 2026).
19. *UNI EN 981:2009*; Machinery Safety-Auditory and Visual Warning and Information Systems. UNI: Rome, Italy, 2009. Available online: https://store.uni.com/uni-en-981-2009 (accessed on 5 March 2026).
20. Radogna, A.V.; Siciliano, P.A.; Sabina, S.; Sabato, E.; Capone, S. A Low-Cost Breath Analyzer Module in Domiciliary Non-Invasive Mechanical Ventilation for Remote COPD Patient Monitoring. *Sensors* **2020**, *20*, 653. [CrossRef]
21. Gervasi, A.; Cardol, P.; Meyer, P.E. Open-hardware wireless controller and 3D-printed pumps for efficient liquid manipulation. *HardwareX* **2021**, *9*, e00199. [CrossRef]
22. Anedda, M.; Floris, A.; Girau, R.; Fadda, M.; Ruiu, P.; Farina, M.; Bonu, A.; Giusto, D.D. Privacy and Security Best Practices for IoT Solutions. *IEEE Access* **2023**, *11*, 129156–129172. [CrossRef]
23. Chietera, F.P.; Colella, R.; Catarinucci, L. The Promising Role of 3D-printed Dielectric Resonator Antennas in the IoT Framework. In Proceedings of the 2021 6th International Conference on Smart and Sustainable Technologies (SpliTech), Bol and Split, Croatia, 8–11 September 2021; pp. 1–4. [CrossRef]

24. Bilbao-Jayo, A.; Almeida, A.; Sergi, I.; Montanaro, T.; Fasano, L.; Emaldi, M.; Patrono, L. Behavior modeling for a beacon-based indoor location system. *Sensors* **2021**, *21*, 4839. [CrossRef]
25. Bilbao-Jayo, A.; Cantero, X.; Almeida, A.; Fasano, L.; Montanaro, T.; Sergi, I.; Patrono, L. Location Based Indoor and Outdoor Lightweight Activity Recognition System. *Electronics* **2022**, *11*, 360. [CrossRef]
26. Silero VAD Github Official Repository. Available online: https://github.com/snakers4/silero-vad (accessed on 5 March 2026).
27. Vér, I.L.; Beranek, L.L. *Noise and Vibration Control Engineering: Principles and Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2005. [CrossRef]
28. Kuo, S.M.; Morgan, D.R. Active noise control: A tutorial review. *Proc. IEEE* **1999**, *87*, 943–973. [CrossRef]
29. Dewasthale, M.M.; Kharadkar, R. Acoustic noise cancellation using adaptive filters: A survey. In *Proceedings of the 2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*; IEEE: Piscataway, NJ, USA, 2014; pp. 12–16. [CrossRef]
30. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech, Signal Process.* **1979**, *27*, 113–120. [CrossRef]
31. Hernández, O.; Olvera, E. Noise cancellation on ECG and heart rate signals using the undecimated wavelet transform. In *Proceedings of the 2009 International Conference on eHealth, Telemedicine, and Social Medicine*; IEEE: Piscataway, NJ, USA, 2009; pp. 145–150. [CrossRef]
32. Mertins, A. Short-time Fourier analysis. In *Signal Analysis*; Wiley: Hoboken, NJ, USA, 1996; pp. 196–210. [CrossRef]
33. Kopparapu, S.K.; Laxminarayana, M. Choice of Mel filter bank in computing MFCC of a resampled speech. In *Proceedings of the 10th International Conference on Information Science, Signal Processing and Their Applications (ISSPA 2010)*; IEEE: Piscataway, NJ, USA, 2010; pp. 121–124. [CrossRef]
34. Sharma, G.; Umapathy, K.; Krishnan, S. Trends in audio signal feature extraction methods. *Appl. Acoust.* **2020**, *158*, 107020. [CrossRef]
35. Hossan, M.A.; Memon, S.; Gregory, M.A. A novel approach for MFCC feature extraction. In *Proceedings of the 2010 4th International Conference on Signal Processing and Communication Systems*; IEEE: Piscataway, NJ, USA, 2010; pp. 1–5. [CrossRef]
36. Zhang, H.; McLoughlin, I.; Song, Y. Robust sound event recognition using convolutional neural networks. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2015; pp. 559–563. [CrossRef]
37. Wilkinson, N.; Niesler, T. A hybrid CNN-BiLSTM voice activity detector. In *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2021; pp. 6803–6807. [CrossRef]
38. Korkmaz, Y.; Boyacı, A. Hybrid voice activity detection system based on LSTM and auditory speech features. *Biomed. Signal Process. Control* **2023**, *80*, 104408. [CrossRef]
39. Karan, B.; van Vüren, J.J.; de Wet, F.; Niesler, T. A Transformer-Based Voice Activity Detector. In Proceedings of the Interspeech, Kos, Greece, 1–5 September 2024. [CrossRef]
40. Lu, X.; Tsao, Y.; Matsuda, S.; Hori, C. Speech enhancement based on deep denoising autoencoder. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; Volume 2013, pp. 436–440. [CrossRef]
41. Bhope, R.; Talele, K.; Huang, T. Adaptive Ambiance Mode For Noise Cancelling Headphones. In *Proceedings of the 2023 IEEE Industrial Electronics and Applications Conference (IEACon)*; IEEE: Piscataway, NJ, USA, 2023; pp. 231–236. [CrossRef]
42. Krishna, A.; Ravinchandra, L.; Fei, T.K.; Yong, L.C. Active noise reduction using LMS and FxLMS algorithms. *J. Phys. Conf. Ser.* **2019**, *1228*, 012064. [CrossRef]
43. Shi, D.; Lam, B.; Ooi, K.; Shen, X.; Gan, W.S. Selective fixed-filter active noise control based on convolutional neural network. *Signal Process.* **2022**, *190*, 108317. [CrossRef]
44. Anjum, M.N. eNext: An IoT and AI Driven Solution to the Plugged-Ear Pandemic. *IEEE Internet Things J.* **2023**, *10*, 11940–11941. [CrossRef]
45. Makino, S.; Araki, S.; Mukai, R.; Sawada, H. Audio source separation based on independent component analysis. In *Proceedings of the 2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No. 04CH37512)*; IEEE: Piscataway, NJ, USA, 2004; Volume 5, pp. 668–671. [CrossRef]
46. Veluri, B.; Itani, M.; Chan, J.; Yoshioka, T.; Gollakota, S. Semantic hearing: Programming acoustic scenes with binaural hearables. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023; pp. 1–15. [CrossRef]
47. Veluri, B.; Itani, M.; Chen, T.; Yoshioka, T.; Gollakota, S. Look Once to Hear: Target Speech Hearing with Noisy Examples. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; pp. 1–16. [CrossRef]
48. Lee, S.J.; Kwon, H.Y. A preprocessing strategy for denoising of speech data based on speech segment detection. *Appl. Sci.* **2020**, *10*, 7385. [CrossRef]

49. Bai, L.; Zhang, Z.; Hu, J. Voice activity detection based on deep neural networks and Viterbi. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *231*, 012042. [CrossRef]
50. Kos, M. Noise reduction algorithm for robust speech recognition using minimum statistics method and neural network VAD. In *Proceedings of the 2007 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference Focused on Speech and Image Processing, Multimedia Communications and Services*; IEEE: Piscataway, NJ, USA, 2007; pp. 284–287. [CrossRef]
51. Patil, R.M.; Patil, C. Unveiling the State-of-the-Art: A Comprehensive Survey on Voice Activity Detection Techniques. In *Proceedings of the 2024 Asia Pacific Conference on Innovation in Technology (APCIT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5. [CrossRef]
52. European Parliament and Council. Directive 2003/10/EC of the European Parliament and of the Council of 6 February 2003 on the Minimum Health and Safety Requirements Regarding the Exposure of Workers to the Risks Arising from Physical Agents (Noise). 2003. Available online: https://osha.europa.eu/it/legislation/directives/82 (accessed on 5 March 2026).
53. Occupational Safety and Health Administration. *Occupational Noise Exposure (Standard No. 1910.95)*; Technical Report; U.S. Department of Labor: Washington, DC, USA, 2008. Available online: https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.95 (accessed on 5 March 2026).
54. National Institute for Occupational Safety and Health. *Criteria for a Recommended Standard: Occupational Noise Exposure, Revised Criteria 1998*; Technical Report DHHS Publication No. 98-126; NIOSH: Cincinnati, OH, USA, 1998. Available online: https://stacks.cdc.gov/view/cdc/6376 (accessed on 5 March 2026).
55. European Parliament and Council. Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on Machinery, and Amending Directive 95/16/EC. 2006. Available online: https://eur-lex.europa.eu/eli/dir/2006/42/oj/eng (accessed on 5 March 2026).
56. *ISO 4869-1:1990*; Acoustics—Hearing Protectors—Part 1: Subjective Method for the Measurement of Sound Attenuation. Technical Report; International Organization for Standardization (ISO): Geneva, Switzerland, 1990. Available online: https://www.iso.org/standard/10850.html (accessed on 5 March 2026).
57. European Committee for Standardization. *Hearing Protectors—Safety Requirements and Testing*; Technical Report EN 352 Series; CEN: Brussels, Belgium, 2002. Available online: https://standards.iteh.ai/catalog/standards/cen/ef8e29f8-2bc6-4f00-925d-92cd768e9f0b/en-352-6-2020a1-2024 (accessed on 5 March 2026).
58. McKinnon, M.; Khaki, S.; Reddy, C.; Chandan, K.A.; Huang, W. Window Size Versus Accuracy Experiments in Voice Activity Detectors. *arXiv* **2026**, arXiv:2601.17270. [CrossRef]
59. Paolucci, F.; Landi, L.; Mariconte, R.; Giliberti, C.; Salzano, F.A.; Patrono, L.; Catarinucci, L. Industrial warning system with active devices for signal reception and dynamic noise attenuation using artificial intelligence algorithms. In Proceedings of the 35th European Safety and Reliability Conference (ESREL2025) and the 33rd Society for Risk Analysis Europe Conference (SRA-E 2025), Stavanger, Norway, 15–19 June 2025. [CrossRef]
60. Awan, S.N.; Shaikh, M.A.; Desjardins, M.; Feinstein, H.; Abbott, K.V. The effect of microphone frequency response on spectral and cepstral measures of voice: An examination of low-cost electret headset microphones. *Am. J. Speech-Lang. Pathol.* **2022**, *31*, 959–973. [CrossRef]
61. Raspberry Pi Official Website. Raspberry Pi 4-Model B. 2023. Available online: https://www.raspberrypi.com/products/raspberry-pi-4-model-b/ (accessed on 5 March 2026).
62. Lattanzi, L.; Raffaeli, R.; Peruzzini, M.; Pellicciari, M. Digital twin for smart manufacturing: A review of concepts towards a practical industrial implementation. *Int. J. Comput. Integr. Manuf.* **2021**, *34*, 567–597. [CrossRef]
63. Union, I.T. ITU-T Recommendation P.563: Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications. 2004. Available online: https://www.itu.int/rec/T-REC-P.563 (accessed on 5 March 2026).