Evaluating and predicting the audibility of acoustic alarms in the workplace using experimental methods and deep learning

François Effa, Jean-Pierre Arz, Romain Serizel, Nicolas Grimault

Evaluating and Predicting the Audibility of Acoustic Alarms in the Workplace Using Experimental Methods and Deep Learning

F. Effa[1,2,3*], J.-P. Arz[1], R. Serizel[2], N. Grimault[3]

1. Institut National de Recherche et de Sécurité, Rue du Morvan, F-54500 Vandœuvre-lès-Nancy, France
2. Université de Lorraine, CNRS, INRIA, LORIA, Campus Scientifique, 615 Rue du Jardin-Botanique, F-54506 Vandœuvre-lès-Nancy, France
3. Centre de Recherche en Neurosciences de Lyon, CNRS, Neurocampus, 95 Bd Pinel, F-69500 Bron, France

*Corresponding author e-mail: francois.effa@gadz.org

ABSTRACT

Occupational noise exposure is a widespread concern, impacting millions of workers. The present research focuses on the audibility of acoustic alarms to ensure worker safety while minimizing exposure to unnecessarily high alarm levels. It introduces a laboratory experiment carried on normal-hearing participants to assess the perceived audibility of acoustic alarms in various workplace noise conditions. The experiment aimed to enhance comprehension of the audibility of acoustic alarms at supra-threshold levels, sought to facilitate the formulation of improved guidelines for alarm design. The results reveal the inappropriateness of the most commonly employed alarm level setting criterion of the ISO 7731 international standard, leading to excessive alarm levels in highly noisy work environments. Based on our data, we propose a revised value for this criterion. In addition, an acoustical analysis of the sounds used in the experiment shows that alarms that are more salient are perceived as more audible, thereby providing leads for alarm design. The study also introduces an innovative technique using a convolutional neural network model to predict the audibility of alarms in noise. Moving beyond generic arbitrary criteria, this data-driven approach leverages knowledge from perceptually annotated examples sourced from our contributed dataset. Evaluation on the experimental data and further analysis of the model outputs demonstrate solid alignment of the model predictions with human perception.

Keywords: Psychoacoustics, Alarms, Audibility, Occupational noise, Convolutional Neural Network, Dataset

1. INTRODUCTION

Exposure to high-level occupational noise affects millions of workers around the world. In France, more than 3 million workers (10%) are subject to prolonged exposure to hazardous occupational noise levels. This number is close to 3.7 million in Canada (15%) and 22 million in the U.S. (14%) [1,2], making noise one of the most prevalent occupational risk factors in these countries. The deleterious effects of noise exposure on hearing [3,4] and more generally on health [5,6] have been widely studied for years. In addition, occupational noise has also been discussed as a contributing factor to workplace accidents, especially when it impairs communication and perception of acoustic alarms [7–10].

In many occupational settings, acoustic alarms are used to alert workers to hazardous situations that may require immediate action. Therefore, the audibility of these sounds plays a critical role in ensuring worker safety. The ISO 7731 international standard dedicated to auditory danger signals for public and work areas [11] requires that acoustic alarms be “clearly audible”, meaning that the masked threshold has to be “distinctly exceeded”. In order to eliminate any uncertainty around the terms “clearly” and “distinctly”, the standard specifies three level criteria for alarms relative to ambient noise. These criteria rely on time-averaged objective measures. Meeting at least one of them is necessary to comply with the standard. The first criterion imposes a minimum difference of 15 dB between the respective A-weighted levels of the alarm and background noise. It leads, however, to unnecessary high alarm levels, especially in work environments with a high noise level. This is acknowledged by the standard itself, which describes the requirements as “sufficient but not always necessary” for alarms to be properly heard and recognized. Numerous studies confirm this observation and demonstrate the absence of consensus on an audibility criterion expressed as an overall fixed signal-to-noise ratio (SNR) [12–17]. In an

experiment performed on normal-hearing listeners, Żera and Nagórski asked participants to adjust the level of acoustic alarms in noise so that they were perceived as "clearly audible". Results showed that the SNR decreased continuously from 15 dB to about – 2 dB when the noise level was raised from 60 to 90 dB [12]. Dolan and Rainey suggested a lower SNR limit of – 10 dB for perception of train horns, corresponding to a 50% detection rate [13]. More recently, two experimental studies carried on reverse alarms in different workplace noises established that the SNR criterion of 15 dB or higher leads to excessive alarm levels, while the criterion imposing a minimum SNR of 0 dB, specified in the ISO 9533 standard for earth-moving machinery [14], seems adequate for reverse alarms [16,17]. The second and third criteria of the ISO 7731 standard are based on the computation of an effective masked threshold, respectively computed from one octave-band and one-third octave-band spectra of the alarm and background noise. Although leading to more appropriate alarm levels than the global SNR criterion of at least 15 dB [15], the differences can be negligible in the case of pure tones or sounds with a largely dominant frequency component. Furthermore, the practical application of these last two criteria is less frequent due to their higher level of complexity. As a result, the criterion based on a global SNR of at least 15 dB remains predominantly used in various workplaces, thereby exposing workers to excessive alarm levels.

In the interests of occupational safety and health, it is crucial to establish a suitable level for acoustic alarms to ensure they are effectively heard without being overly loud. Low alarm levels may be too weak to properly transmit emergency information and prompt a quick response to control or eliminate the danger. Conversely, in addition to being a source of annoyance, very high levels can lead to permanent damage to the hearing of workers. Moreover, a sudden increase of the sound pressure level in the work area caused by loud acoustic alarms is likely to induce startle reactions, which could endanger workers [11,12]. In that respect, the absence of consistent guidelines established in the field to ensure appropriate levels of acoustic alarms poses a significant problem [16]. This can be addressed by adjusting the level of acoustic alarms through listening tests. However, it requires recruiting volunteers whose hearing status matches the target population and presenting them with stimuli under well-controlled conditions, which can be demanding and time-consuming. Most importantly, the experimental approach is stimulus-dependent, as the audibility of an acoustic alarm depends on various parameters such as the temporal envelope and long-term power spectrum of the alarm, the background noise spectrum, and the interaction between these factors [18]. Therefore, the slightest change in the sound environment or the acoustic alarm should necessitate a new experimental assessment. Hence, the use of predictive approaches is more convenient. In that perspective, many auditory models have already been developed, some of which show convincing performance in predicting detection thresholds of complex time-varying target sounds in complex backgrounds [19–21]. Notwithstanding their accuracy, those models are of limited use when determining the appropriate level for acoustic alarms in complex noisy environments. There are three reasons for this. First, many of those models are exclusively designed to predict detection thresholds and cannot provide any relevant information regarding the perception of target sounds at supraliminal levels. Second, these models are intrusive, in that they require the separate consideration of the alarms and background noises. Finally, the decision stage of a model capable of determining the proper level of an acoustic alarm should require prior knowledge about what is considered "clearly audible", which is largely unknown. As such, one of the objectives of the present study is to better assess and discuss what is meant by "clearly audible". In addition, we propose a solution based on the use of a neural network model to address the issue of adjusting alarm levels. This approach offers the advantage of being data-driven, meaning that the decision criterion regarding the audibility of a sound can be implicitly contained in the perceptual data used to train the model with no need to be formulated at design stage.

Recent advancements in deep learning have had a significant impact in the field of ambient sound analysis, facilitating the accurate and efficient recognition of environmental sounds. Deep neural networks such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated a great ability to learn complex patterns in acoustic data and adapt to new environments. This has led to the predominant use of deep learning methods in a wide variety of ambient sound analysis applications. Such applications include acoustic scene classification [22], sound event detection [23], sound source localization [24], or anomalous sound detection [25]. In the area of auditory perception and cognition, a number of researchers have explored deep neural networks to draw parallels between the internal representations of the human brain and those in deep network models [26–28]. Several studies also investigated similarities between human and model behavior in auditory tasks like word recognition, musical genre recognition, sound localization, fundamental frequency estimation [28–30]. In two recent papers, we proposed a proof of concept of an automatic approach to evaluate the audibility of acoustic alarms in noise using a CNN trained on perceptually annotated data [31,32]. In the present study, we provide a more detailed presentation of this approach, including evaluation of the model on a larger and more elaborate dataset, as well as an in-depth analysis of the perceptual data and model performance.

The purpose of this article is two-fold. Firstly, it seeks to provide better understanding of the audibility of acoustic alarms at supra-threshold levels, which can help formulate guidelines for alarm design. This will be achieved by the means of a laboratory experiment carried on normal-hearing subjects and involving a large variety of acoustic alarms and workplace noise conditions. Secondly, it introduces a low-complexity CNN-based approach to predict the audibility of acoustic alarms in complex noisy environments. The remainder of the paper is structured as follows. In Section 2, we define the notion of audibility referred to throughout the article, taking care to clarify any semantic discrepancies with other uses of this term found in the literature. Section 3 is dedicated to the description of the experimental procedure developed to evaluate the audibility of acoustic alarms in noise, followed by an analysis of the experimental results. Section 4 presents the collection of a perceptually annotated dataset, as well as the learning and evaluation of a CNN model on the perceptual data of Section 3 to make audibility predictions. In Section 5, we conduct a comprehensive comparative analysis between the model and human responses, encompassing examination of machine learning performance metrics and psychoacoustical interpretation of the model outputs. We conclude in Section 6.

2. PROBLEM STATEMENT AND RELATION TO THE ISO STANDARD

Numerous works refer to the notion of audibility to characterize a wide variety of sounds including alarms. Most of the time, the term audibility is used as a synonym of detectability which refers to the probability of a sound being detected under specific conditions. However, the specific meaning and criteria used to assess this concept may vary across studies depending on their specific goals and methodologies. Besides, the study of sound perception – especially for acoustic alarms – is not restricted to levels at which the sounds are just detected and can involve concepts different from audibility but for which the boundaries with audibility are narrow and not well established. In this article, to be in line with the ISO 7731 international standard, we will use the term audibility as a property of what is “clearly audible”. While we acknowledge that this definition is vague and goes beyond the single question of detectability, it is intended to capture the vocabulary of the standard. The purpose of this section is to introduce the different concepts related to or interacting with the perception of the audibility of acoustic alarms in noise based on the literature.

An acoustic alarm is expected to possess acoustic characteristics that allow for perceptibility and appropriate response by individuals located within the designated reception area [11]. This was first defined by Wilkins as the “effectiveness” of an acoustic alarm [33]. According to Wilkins, the perception of an alarm involves three components: audibility, attention demand, and recognition. Audibility determines if the sound can be heard amid background noise. Attention demand relates to the ability of the sound to attract attention and be consciously perceived when unexpected. Finally, recognition requires the sound to be distinguishable from other sounds and convey the meaning of danger. In the ISO 7731 standard, the idea of “effectiveness” of a danger signal is referred to as “reliable recognition”. Like Wilkins, the standard expresses criteria of effectiveness, namely audibility, distinctiveness, unambiguity and independence from source movement. Here, audibility means that the danger signal has to be “clearly audible”, and that the effective masked threshold must be “distinctly exceeded”. Distinctiveness requires that the signal be designed to stand out from all other sounds in the reception area including any other signals. Eventually, the signal must have an unambiguous meaning, and its characteristics must be recognizable no matter the potential movement of the source.

From the above, the standard mainly considers audibility as a masking issue. In order to meet the audibility criterion, the alarms are required to “distinctly” exceed the masked thresholds. However, there is no clear guidance on the optimal amount by which the alarms should surpass the masked thresholds. Previous studies have suggested that alarms should be 12 to 25 dB above the masked thresholds [16,34–36], but this range is too broad to be practically useful in most situations. Therefore, it is difficult to establish a more precise criterion solely based on masking that would be widely accepted. This indicates that other processes operating at supraliminal levels may also play a role in the perception of an alarm in noise as “clearly audible” and should be taken into account. In particular, it is likely that distinctiveness, as it is defined above, has a significant impact on the perception of an alarm as “clearly audible”. In this regard, recent works on auditory salience have highlighted the fact that sound events with divergent acoustic properties from those of the surrounding environment are more perceptually prominent, therefore easier to hear [37,38]. Nonetheless, while salience may influence audibility, it represents a distinct concept that focuses on the properties of sound eliciting involuntary attention, independently of top-down factors. Measuring salience therefore requires purely passive listening, contrary to audibility, which entails active listening to assess the “clearly audible” aspect of the sounds. In that respect, salience refers more to attention demand than to audibility *per se* [39]. This observation somehow reflects the fuzzy boundaries that exist between the different components of alarm effectiveness. In their study, Laroche *et al.* measure reaction thresholds [16],

defined as the levels at which the alarms elicit a response, such as turning towards the source, or moving away from the danger zone. This definition, while resembling the broader notion of alarm efficiency, emphasizes the crucial point that alarms must exceed a certain level to be effective, aligning with the ISO 7731 requirement for alarms to be "clearly audible". While it does not assume accurate source recognition, it is still based on the idea that the sound must indicate danger or urgency to trigger a reaction, which is only achievable if the alarm exhibits sufficient distinctiveness. This understanding of audibility supports the idea that the optimal level for an alarm to be perceived as "clearly audible" is modulated by factors beyond detectability, such as distinctiveness and attention demand.

In the present work, the choice was made to measure audibility in accordance with the terms of the ISO 7731 international standard. We assess the detectability as well as the "clearly audible" aspect of auditory alarms in occupational environments. Through data analysis, we compare our results to existing audibility criteria while also shedding light on the overlaps and disparities with detectability. Additionally, as there is currently no established linkage between the ISO 7731 standard requirements and recent research on auditory salience, we undertake a preliminary effort to connect audibility and salience, investigating potential relationships between these two notions. In addition to traditional psychoacoustical techniques, we introduce an automatic deep learning-based method to evaluate the audibility of acoustic alarms. This approach avoids reliance on a predefined explicit audibility criterion, which could potentially be subject to debate. Instead, we leverage a model that learns from examples of subjective audibility evaluations to produce outputs that emulate human judgments.

3. EXPERIMENTAL EVALUATION OF AUDIBILITY

3.1. Experimental design

3.1.1. Participants

The experiment involved the participation of 20 volunteers, aged from 20 to 50. All of them had normal-hearing according to the International Bureau of Audiophonology criteria, with an average tone loss of less than 20 dB HL across the frequencies 500, 1000, 2000 and 4000 Hz on both ears. The participants were compensated for the time spent on the tests.

3.1.2. Stimuli and material

Fifteen alarms and ten recordings of noisy work environments (backgrounds) were used to create the stimuli. The alarms and background sounds were mainly sourced from public platforms, namely Freesound [40] and BigSoundBank [41]. Additionally, we received some files from authors of published studies [42,43], and a few others were from personal recordings. Both the alarms and noisy backgrounds were evenly split into five categories based on their associated environmental contexts. A total of 30 alarm-background pairs were made by associating acoustic alarms and background noises within each contextual category[1]. To create the stimuli, monophonic sound clips were generated by adding an alarm to its background noise using a pseudo-random temporal onset, with boundary values set to avoid extreme temporal onset locations. Each sound clip was 5.5-second long and sampled at a rate of 44.1 kHz. Within the 5.5 s of a given sound clip, the alarm was played once. This single alarm occurrence could involve several bursts for intermittent alarms such as reverse alarms, but the overall alarm duration was limited to 1.8 s, which was the duration of the longest alarm. In order to avoid any clicking effect arising from a sudden volume change at the start or end of the stimuli, 20 ms raised-cosine onset and offset ramps were applied to the clips. The stimuli were output through a Babyface Pro soundcard (RME, Germany) and presented over DT 770 Pro circumaural headphones (Beyerdynamic, Germany) calibrated with an AEC101 artificial ear and a Model 824 sonometer (Larson Davis, USA). To manage stimulus presentation and participant responses, a custom interface was created using Matlab App Designer.

[1] Information regarding the contextual categories and alarm-background pairs, as well as spectrograms of the sounds, are provided in the supplementary material accompanying this paper

3.1.3.Procedure

The experiment consisted of two tasks: the first task aimed to measure detection thresholds; the second was an audibility assessment. Both tasks evaluated the 30 alarm-background pairs at two levels of background noise (60 and 80 dBA) using the method of constant stimuli. The value of 80 dBA was chosen in alignment with common practices observed in studies on workplace sound environments. As one of our objectives was to investigate the impact of ambient noise level on detection and audibility, we employed two distinct noise levels. We wanted to reflect ecological conditions while ensuring participant safety by avoiding exposure to excessively loud stimuli. Consequently, for the second noise level, we preferred the value of 60 dBA to levels above 80 dBA, which could have been hazardous.

By definition, at detection threshold, an alarm is minimally audible. For an alarm to be considered "clearly audible", its level must be much greater than detection threshold. Therefore, the study of audibility involves a scale of alarm levels different from that of detectability. This was accounted for by using a distinct range of SNRs for each task, varying alarm levels to reach a full coverage of the psychometric function domains.

The detection task followed a two-interval two-alternative forced choice (2I-2AFC) design. During each trial, participants were presented with two consecutive intervals separated by a 500 ms pause. Both intervals were generated using the same background noise, but only one of them contained the alarm to be detected. After the two intervals were presented in a random order, participants were required to indicate which interval contained the alarm. The 30 noise-alarm pairs were all presented at six different SNRs, namely – 30, – 22.5, – 17.5, – 12.5, – 7.5 and 0 dB. To ensure robustness of the results, each condition was repeated three times for each participant.

To assess audibility, participants were instructed to listen to the overall auditory environment without specifically focusing on attempting to detect the alarm[2]. The evaluation consisted in a straightforward Yes-No task. Each trial consisted of a single presentation of the stimulus, followed by the question "*Was the alarm clearly audible?*". The binary assessment of audibility using the term "clearly audible" was directly derived from the ISO 7731 standard. Despite the ambiguity of this expression, our choice was motivated by the necessity to offer a clearer understanding of this concept through our results and to enable comparison with the recommendations of the standard. Each alarm-background pair was presented three times at six different SNRs: – 25, – 15, – 10, – 5, 0 and 10 dB. This range is broad and extends down to very low SNR values. Even though an SNR of – 25 dB would probably not be encountered in practice, it was included in our experiment to cover the lower end of the psychometric function domain.

To minimize order effects, the task and stimulus presentation sequences were randomized for each participant. Considering the combinations of alarm-background pairs, SNRs, and noise levels, the listeners were presented with 360 different stimuli per task. As each stimulus was presented three times, the cumulative number of trials for each participant amounted to a total of 2160 trials. In order to avoid auditory fatigue among participants, the experiment was structured into discrete sessions. Each session was limited to a maximum duration of 2 hours. Within each session, the activity was subdivided into 3 to 4 blocks, with each block spanning approximately 25 minutes and separated with 5-minute breaks.

3.2. Results

3.2.1.Psychometric functions: general observations

For each participant, the output of each task is represented as a probability score, determined for a specific alarm-background pair at fixed noise level and SNR. In the detection part of the experiment, this score represents the correct response rate, which is the probability of the participant providing the correct answer in a 2I-2AFC trial. It is bounded between 0.5 and 1, where 0.5 indicates chance performance or random guessing between the two intervals, and 1 signifies perfect accuracy. Alternatively, for the audibility assessment, the score ranges from 0 to 1 and denotes the probability that the participant considers the alarm as "clearly audible".

Cumulative Gaussian sigmoids were fitted to these perceptual data using *psignifit* (version 4) toolbox [44]. The shape of these individual psychometric functions is determined by the four fitted parameters, which are the inflection point m, the slope (or width) w, and the upper and lower asymptotes λ and γ. The general form of the psychometric functions is expressed in Equation 1, with x representing the SNR and $F(x; m, w)$ a cumulative Gaussian distribution function.

[2] This instruction aimed to encourage participants to consider the surrounding sound context and prevent them from placing excessive emphasis on the alarm, which could potentially lead to distorted audibility judgments.

$$\psi(x; m, w, \lambda, \gamma) = \gamma + (1 - \gamma - \lambda)F(x; m, w) \quad (1)$$

Figure 1 shows the mean psychometric curves, grouped by noise level for both tasks. As we are not interested in individual curves, we rather represent the mean psychometric curves across listeners. These curves were derived by computing the average of the 20 individual curves for each alarm-background pair, and then averaging the resulting curves across all pairs. In order to visualize the variability across the various sound conditions, the representation includes the corresponding standard errors across the set of 30 alarm-background combinations. Notably, both detectability and audibility are influenced by the noise level, albeit with opposite effects. Specifically, in the detection task, the average psychometric function in the background noise level of 60 dBA is consistently positioned above that in the noise level of 80 dBA across the entire range of SNR. This result shows that higher noise levels negatively impact alarm detection, making it more challenging. This effect is likely attributed to frequency masking. In higher noise conditions, the auditory filters are broadened, leading to increased frequency masking and consequently decreased detectability. Conversely, for the audibility assessment, the order between the two curves is inverted. This indicates that, at supra-threshold levels, higher noise levels are associated with increased audibility for a given SNR. A possible explanation for this phenomenon is that audibility may primarily depend on loudness. Beyond the masked threshold, the loudness growth steepens in higher noise levels due to the elevated absolute alarm levels at a constant SNR. This, in turn, could result in higher audibility scores. Additionally, we observe that the audibility psychometric functions are superimposed in the low SNR region up to the inflection point. This superimposition indicates that the effect of noise level on audibility is not apparent when the alarm level is not high enough for reliable detection. In other words, the effect of noise level becomes evident in the region where studying audibility is of particular interest.

From an SNR of 7.5 dB, the correct response rate for the detection task is at a plateau value of nearly 100%, regardless of the noise level. Similarly, at the same SNR, the mean audibility scores are high for both noise levels of 60 and 80 dBA (90% and 95%, respectively). This observation supports the idea that a SNR of 15 dB, as stated in the ISO 7731 standard, may not always be necessary to ensure reliable detection and a strong level of reported audibility.

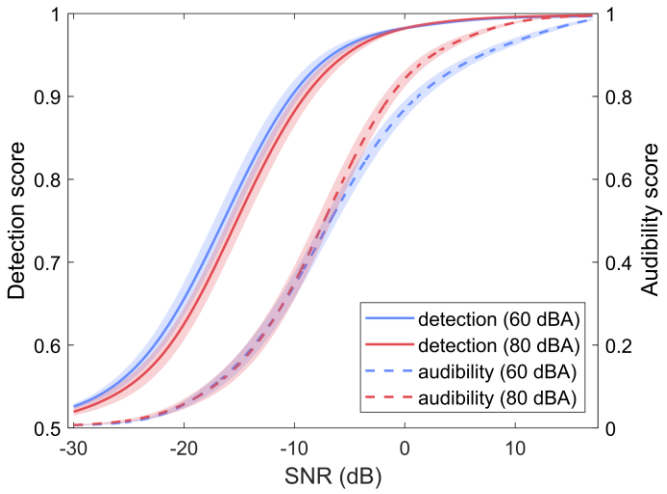

Figure 1: Mean across-participants psychometric curves, averaged across alarm-background pairs for both tasks, grouped by noise level. Standard errors across the 30 alarm-background pairs are represented in shaded areas. Left axis, plain line: Correct response rate as a function of SNR. The 2I-2AFC task leads to a scale ranging from 0.5 to 1. Right axis, dashed line: "Clearly audible" rate as a function of SNR. The Yes-No task leads to a scale ranging from 0 to 1.

To further analyze our results and confirm our observations, we conducted a series of statistical tests. We performed Bayesian ANOVAs, due to the flexibility and interpretability offered by the Bayesian framework in handling uncertainty and incorporating prior knowledge, facilitating robust and insightful inferences from the observed data. The first Bayesian ANOVA was applied to the detection threshold, defined as the value m of the SNR at the inflection point of the psychometric function, with the participant as a random variable. It revealed significant effects of the noise level (BF = 2.3×10^4), of the alarm-background pair (BF = ∞), and of their interaction (BF = 124.5). Regarding audibility, we chose not to study the inflection point, as an audibility score of around 50% is too weak and lacks practical interest. Instead, we considered the audibility score reported at an SNR

of 7.5 dB. To compensate for floor and ceiling effects, all scores were first transformed from percentages into rationalized arcsine units (RAU) before the analyses [45]. The second Bayesian ANOVA, applied to the audibility score with the participant as a random variable showed significant effects of the noise level ($BF = 1.1 \times 10^{13}$), of the alarm-background pair ($BF = 2.0 \times 10^{14}$), and of their interaction ($BF = 63.4$). Overall, these results provide confirmatory evidence that the background noise level, noise environment, and alarm type significantly influence the perception of acoustic alarms in terms of both detectability and audibility.

3.2.2.Comparison with existing criteria

The ISO 7731 standard sets two major requirements to adjust alarms at levels considered "clearly audible". First, alarm levels should be greater than 65 dBA. Second, they should meet at least one of the three following criteria: (Method A, Section 5.2.2.1) "*the difference between the two A-weighted sound–pressure levels of the signal and the ambient noise shall be greater than 15 dB*", (Method B, Section 5.2.3.1) "*the sound–pressure level of the signal in one or more octave-bands shall exceed the effective masked threshold by at least 10 dB in the octave-band under consideration*", (Method C, Section 5.2.3.2) "*the sound–pressure level of the signal in one or more 1/3 octave-bands shall exceed the effective masked threshold by at least 13 dB in the 1/3 octave-band under consideration*". For these two last methods (i.e. B and C), the "effective masked threshold" is computed from the octave-band or the third-octave band spectrum of the ambient noise using a simplified model of masking provided by the standard.

To assess the relevance of these recommendations, we compared them to the measured psychoacoustic data of our experiment. The upper panel of Figure 2 represents the SNR values computed according to the three criteria of the standard for each alarm-background pair. Method B and Method C yield close recommendations, reliably lower and significantly distant from the fixed SNR criterion of minimum 15 dB provided by Method A. Additionally, the SNRs corresponding to a measured average audibility score of 85% for the two noise levels are also plotted on the upper panel of Figure 2. They demonstrate that Method B and Method C of the standard not only effectively predict the relative differences in audibility between the tested alarm-background pairs but also consistently ensure an audibility score greater than 85%, for both noise levels in most cases.

The lower panel of Figure 2 also presents the recommendations of the three criteria of the ISO 7731 standard, on a new scale corresponding to the measured audibility score. As the measured audibility score is dependent on the ambient noise level, each criterion now appears twice per alarm-background pair. The chart also includes boxes that represent the range of 12 to 25 dB above the measured detection threshold, as recommended by the literature [16,34–36]. This representation combines information regarding both detectability and audibility, allowing for an evaluation of the standard criteria with regard to the actual measured detection thresholds and audibility scores. We observe that the lowest end of the range 12 – 25 dB above detection threshold often yields rather low audibility scores, particularly when the noise level is low (i.e. 60 dBA). This observation supports our reflection on the limited practical applicability of this range. Besides, comparing the positions of the boxes and the markers representing Method A, we notice that the fixed, over 15 dB SNR criterion, despite ensuring a 100% audibility score, appears to lead to excessive alarm levels for almost all alarm-background pairs. This result suggests that the criterion advocated by Method A is excessively conservative and not aligned with the other two criteria from the same standard. By comparison, a fixed SNR of 7.5 dB would result in lower alarm levels, falling almost exclusively within the 12 – 25 dB above threshold range, while maintaining good audibility. As evidenced by the variations in audibility scores observed among distinct alarm-background pairs, the choice of a fixed criterion relying on a global SNR may not be the most suitable. However, our findings indicate that if such a criterion were to be applied in accordance with the recommendations of Method A, it would be advisable to set it at a lower SNR value. In that respect, the proposed lower fixed SNR criterion of 7.5 dB appears to be equally effective and offers a more balanced compromise to take into account the risk of overexposure for workers. By adopting a lower alternative criterion, a better balance could be struck between ensuring adequate audibility and avoiding unnecessary alarm levels, resulting in a safer and more practical implementation of the standard.

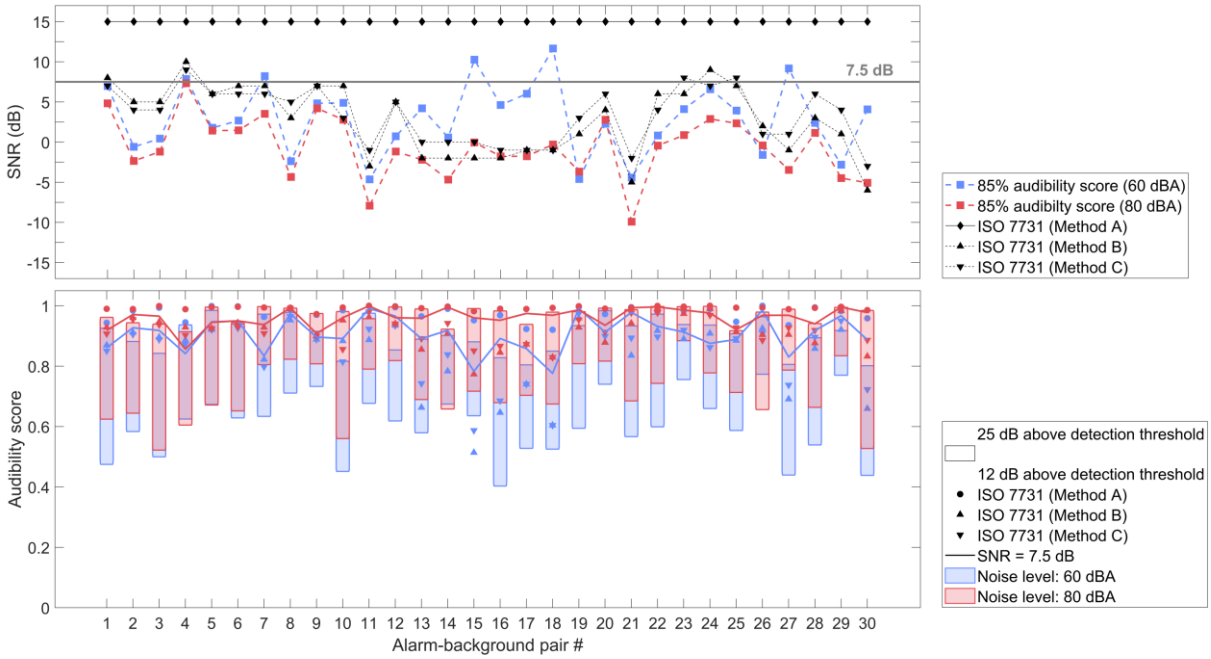

Figure 2: Adjustment of alarm levels based on ISO 7731 (Methods A, B and C) and psychoacoustic data, for each alarm-background pair, in the two noise levels (60 and 80 dBA). Upper panel: SNR as recommended by the ISO 7731 criteria and SNR leading to an averaged measured audibility score of 0.85. Lower panel: Measured audibility scores associated with all three ISO 7731 requirements along with vertical rectangular boxes representing audibility scores in the range 12-25 dB above detection threshold.

3.2.3. Understanding audibility: influencing factors

Given the absence of a precise meaning associated with the expression “clearly audible”, most of the perceptually grounded criteria depend on the detection thresholds. This reliance is evident both in the standard, with the use of the effective masked threshold, and more generally in the literature, where the 12 – 25 dB range above the detection threshold is often recommended for optimal audibility [16,34–36]. It is based on the assumption that the audibility at supraliminal levels can be inferred from the detection thresholds. However, while a clear relationship exists between detectability and audibility, as levels close to or below detection threshold are naturally associated with low audibility scores, this relationship becomes less direct at higher levels. Figure 3 illustrates the correlation between the detection threshold and two different parameters for each alarm-background pair: the inflection point m of the average audibility psychometric curve (left panel) and the SNR at which the average audibility score reaches 0.8 (right panel). The left panel shows at first sight a quasi-linear relationship between the inflection points of audibility and detection curves, with an adjusted-R^2 of 0.66. This finding suggests that the detection threshold indeed determines the level from which the notion of audibility becomes relevant. However, the scatter plot on the right panel highlights that the detection threshold is not a reliable predictor of audibility at higher levels. In the present case, the detection thresholds and the levels corresponding to an audibility score of 0.8 are poorly correlated (adjusted-R^2 = 0.33). This indicates that, despite an existing relationship between detectability and audibility at supraliminal levels, detection thresholds alone cannot fully explain audibility or serve as a basis for establishing optimal level recommendations.

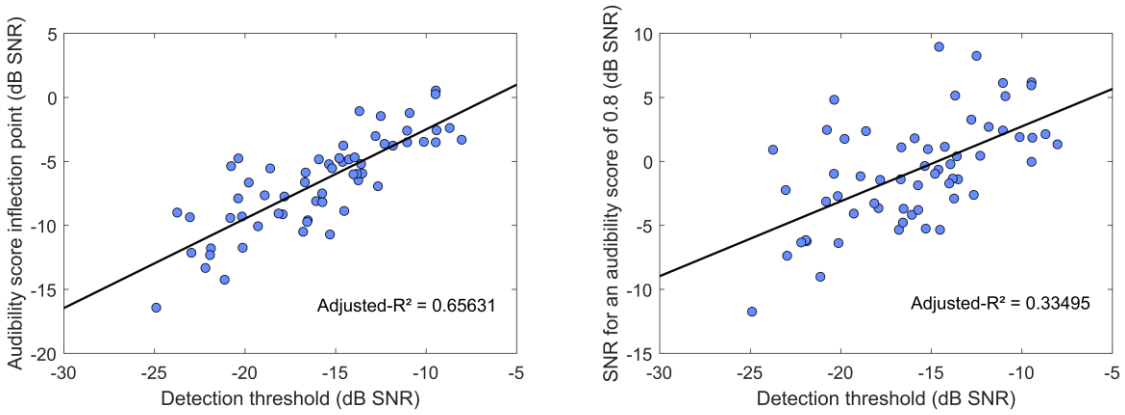

Figure 3: Scatter plots of audibility against detectability for each combination of alarm, background and noise level.
Left panel: Inflection point of the average audibility psychometric curve as a function of the corresponding detection threshold.
Right panel: SNR leading to an average audibility score of 0.8 as a function of the corresponding detection threshold.

In order to identify the other factors influencing the audibility of acoustic alarms in occupational environments, we carried an acoustic analysis of the stimuli used in our experiment. Our aim was to determine whether the audibility of an auditory alarm could be partially explained by the divergence in certain acoustic features compared to the surrounding acoustic environment. For each of the 360 listening conditions, we extracted nine features for both the alarm-background mix and the background alone, using a random alarm temporal onset. These features, namely spectral irregularity, spectral flatness, brightness, bandwidth, spectral modulations (scale), pitch, harmonicity, temporal modulations (rate), and loudness, were computed in accordance with the definitions and methods presented by Huang and Elhilali [38].

To ensure fair comparisons and remove the impact of different scales and distributions, a z-score normalization was applied across the features for each sound clip. Subsequently, the features were time-averaged over the period during which the alarm was supposed to be present. By time averaging the features, we obtained a representative measure of their behavior within the stimuli over this specific period, considering both the alarm-background mix and the background alone. The measure of the divergence in the acoustic features of the environment caused by the alarm was derived by computing the difference between the features of the alarm-background mix and those of the background alone. We collected these feature difference vectors for all listening conditions and then concatenated them into a single array.

To identify the acoustic features that contribute significantly to the audibility of the alarms, we performed a partial least squares logistic regression (PLSR) following the methodology employed by Thévenet *et al.* [46]. PLSR is a powerful multivariate statistical technique that combines aspects of principal component analysis and multiple regression. It is well-suited for handling situations involving high-dimensional predictor variables and potential multicollinearity among them. Unlike traditional regression methods that treat each predictor independently, PLSR identifies latent variables that capture shared information between the predictors and the response. Furthermore, it differs from classical principal component analysis by focusing on maximizing covariance between the predictors and the response variable, rather than just maximizing the variance in the predictors. This property makes PLSR particularly valuable for predictive modeling. The analysis was conducted using the *plsRglm* (version 1.5.1) package [47], employing the nine feature differences as predictor variables and the corresponding audibility scores as the response variable.

Following a preliminary cross-validation, we opted to retain only one component in the PLSR model. Figure 4 displays the results of the PLSR, represented by standardized regression coefficients and bias-corrected and accelerated bootstrap intervals. These intervals were obtained using the balanced bootstrap method with 1000 resampling iterations. Statistically significant coefficients are those whose bootstrap distributions lie either above or below zero. As anticipated, our analysis confirms that loudness serves as a robust predictor of audibility. This supports our interpretation that audibility is predominantly and positively influenced by the loudness of the alarms. Furthermore, the results reveal that variations in loudness, scale, brightness, pitch and harmonicity significantly contribute to predicting the audibility score. Remarkably, these findings closely align with the main predictors of auditory salience identified by Huang and Elhilali [38]. The strong agreement between our results and those of Huang and Elhilali supports the idea that the audibility of acoustic alarms is intricately related to auditory salience. Specifically, an alarm is more likely to be considered "clearly audible" if it induces variations in one or several of the aforementioned acoustic features in the acoustic environment, rendering it more attention-grabbing on the same occasion.

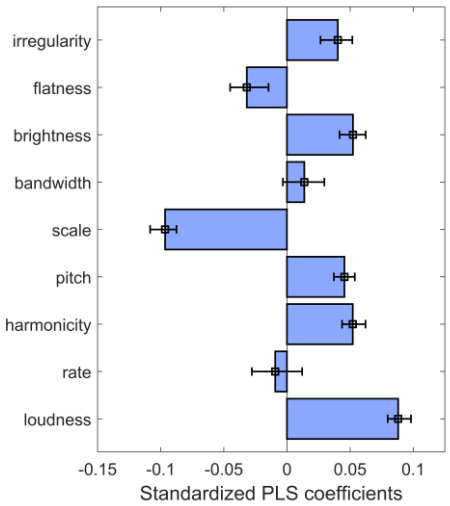

Figure 4: Coefficients of the first PLSR component. Acoustic features whose coefficients have bootstrap distributions above or below zero are considered statistically significant predictors.

3.3. Discussion

In high noise level conditions, the audibility of the alarms was found to be greater compared to lower noise levels. This is consistent with the findings of Żera and Nagórski [12]. This outcome underscores a divergence between audibility and detectability, as the latter becomes more difficult in the presence of higher noise levels. Additionally, while our results evidenced a link between detectability and audibility, we observed that detection thresholds were not a reliable predictor of audibility scores. This was particularly evident at higher alarm levels, which are closer those encountered in real-world scenarios. This finding emphasizes the importance of considering the perception of audibility at supraliminal levels as a distinct factor from detectability when characterizing acoustic alarms.

Consistent with earlier research [15,16,18], our experiment corroborated that the fixed criterion established by Method A of the ISO 7731 standard systematically leads to unnecessarily high alarm levels. On the other hand, although more complex, Method B and Method C, which are based on octave and third-octave band measurements are more suitable for achieving optimal audibility while avoiding reaching excessive alarm levels. The proposed global SNR criterion of minimum 15 dB already results in excessive alarm levels at the two noise levels tested in this study (60 and 80 dBA), and does not account for the noise level dependency of audibility. As a result, it may prove unnecessarily dangerous and even cause hearing impairments, especially in higher levels of noise. In this regard, while expressing reservations about the use of a fixed global SNR criterion, we advise setting this criterion to a lower SNR value. As such, our data demonstrated that a 7.5 dB SNR value is sufficient to maintain good audibility, making it a viable substitute for the current 15 dB criterion. An alternative approach could be to establish an adaptive criterion that would be adjusted based on the ambient noise level, thereby mitigating the risks associated with excessively loud alarm levels while ensuring reliable audibility.

Eventually, our acoustical analysis of the stimuli showed that the acoustic features associated with auditory salience were also those who best predicted audibility. This result suggests that the most distinctive sounds tend to be perceived as the most audible. For comparison purposes, our study focused only on the same features as explored in the work of Huang and Elhilali [38]. However, other features, such as roughness which also contributes to salience [48], could be added to get an even better picture.

4. DEEP LEARNING-BASED AUDIBILITY PREDICTION

4.1. Problem definition

In accordance with the ISO 7731 standard, which stipulates that alarms must be "clearly audible", we frame the evaluation of audibility as a binary classification task in which we classify acoustic alarms as either "clearly audible" or not within different work environments. To achieve this, we implement a system that makes binary audibility predictions. Given an input sound clip containing an alarm played against an occupational background noise, the system produces a binary estimate $\hat{y}$ of whether the alarm is "clearly audible" ($\hat{y} = 1$) or not ($\hat{y} = 0$).

Our proposed approach is data-driven and leverages deep learning methodologies. By embracing a supervised learning framework, we can train a neural network to assess the "clearly audible" attribute of alarms with no need for explicit criteria, as the network learns through exposure to alarm examples paired with human perceptual assessments of audibility. To accommodate the data requirements of deep learning, this approach necessitates the collection of a substantial dataset[3]. This dataset is divided and used across two stages. In the first stage, referred to as development, the model is trained to extract relevant patterns and features from the data. The second stage, known as evaluation, consists in assessing the performance of the trained model on a separate and independent subset[4].

4.2. Dataset

The dataset comprises a collection of sound clips generated by mixing recordings of occupational noises (backgrounds) with acoustic alarms, using a random temporal alarm onset. Each clip is 5.5-second long and contains a single alarm whose duration varies between 0.2 and 1.8 seconds. The acoustic data is accompanied by perceptual annotations, acquired through a listening test involving normal-hearing participants. Specifically, the listeners were presented with the clips and subsequently queried regarding the audibility of the alarms, with the question: "*Was the alarm clearly audible?*".

The dataset is divided in two subsets: one for development and another for evaluation. These subsets serve different purposes and consequently have distinct constraints. A prior study [32] demonstrated that, while reliability and interpretability are crucial for evaluation data, using a lighter annotation for development data does not significantly affect the model performance. As a result, the annotation procedure differs for each subset. For the evaluation data, a well-controlled annotation procedure is essential to yield responses suitable for extensive analysis, similar to a standard psychoacoustic experiment. In contrast, for development data, the emphasis lies more on the richness of the sound corpus rather than the purity of the annotations. Therefore, a more flexible annotation procedure is desirable as it allows for coverage of a broader range of listening conditions in less time.

In an ideal scenario, the collection of such a dataset would necessitate independent groups of participants for annotating the development and evaluation data. However, due to challenges in recruiting participants for the listening tests, adopting this approach would have resulted in an insufficient number of annotations for either subset. As a practical compromise, we proposed the involvement of some individuals in the annotation of both development and evaluation sets. For methodological clarity, we categorized them into three distinct pools denoted as A, B, and C. The number of annotators and the involvement of each pool in the annotation process are shown in Table 1. The potential implications of sharing common annotators between the development and evaluation sets may be explored in future studies.

For each subset, we present its contents and provide a comprehensive description of the employed procedure for annotation. In contrast to the conventional practice of starting with development data, we intentionally present

[3] The dataset containing both the acoustic data and metadata along with the perceptual annotations will be uploaded and publicly available at: https://zenodo.org/doi/10.5281/zenodo.8417086

[4] The code for neural network model development and evaluation as well as the weights of the models analyzed in the paper will be made accessible at: https://github.com/effajr/predicting_alarm_audibility

the evaluation data first. This choice aims to better connect with the preceding section on the psychoacoustic experiment.

	Size	Development	Evaluation
Pool A	12		×
Pool B	8	×	×
Pool C	2	×	

Table 1: Pools of participants involved in the dataset annotation.

4.2.1.Evaluation data

We use the sound clips and subjective responses from the experiment presented in Section 3 as the evaluation data. The subset consists of 360 monophonic sound clips, which corresponds to the 30 alarm-background pairs mixed at six different SNRs, and across two levels of noise as determined in the psychoacoustic experiment. The annotations for the evaluation data are derived from the responses provided by the 20 participants. In the experimental procedure, each clip was presented to the subjects three times, resulting in a total of 60 binary annotations per example in the evaluation set. A binary value of 1 indicates that the alarm was reported as "clearly audible" by the participant, whereas a 0 signifies that the alarm was not perceived as such. Among the twenty annotators, twelve exclusively annotated the evaluation data (**pool A**), while the remaining eight (**pool B**) also contributed to the annotation of development data.

4.2.2.Development data

The development set is composed of 2000 monophonic sound clips, generated using a combined total of 70 alarms and 52 backgrounds. All sounds used for development were sourced exclusively from the Freesound library [40]. In order to maintain a strict separation between development and evaluation sets, care was taken to ensure that the alarms and backgrounds used for development were distinct from those present in the evaluation set. Additionally, in the consideration of Freesound being a collaborative library, we selected development alarms and backgrounds from various Freesound users, distinct from those in the evaluation set. This aimed to prevent biasing the evaluation set with sounds potentially recorded under identical conditions as some of the development data.

The distinction between development and evaluation sounds implies that they were not the exact same signals. However, some of them roughly shared spectral or temporal attributes with sounds of the evaluation set. Notable examples of these shared characteristics included a pulsed temporal structure or a complex harmonic frequency content for alarms, and the presence of broadband factory noise for background sounds. The selection criteria for sounds of the development set were less restrictive compared to those of the evaluation set. Consequently, not all alarms within the development data can be unequivocally categorized into specific types. We provide a global overview of alarm characteristics in the supplementary material accompanying this paper. For illustrative purposes, we also show examples of spectrograms of the development sounds.

The clips were generated through a random pairing process, where an alarm and a background noise were randomly drawn from our pool of sounds. The background noise level was randomly set at either 60 or 80 dBA, and the mixing was made at a randomly selected integer SNR ranging from –30 to +15 dB. The choice of the noise level values was guided by the same motivations as in the psychoacoustic experiment. The setup used to present the stimuli was the same as described in the psychoacoustic experiment. As the annotation procedure was intended to be faster than for evaluation, each clip within the development set was presented to the listeners only once (as opposed to the standard three times for evaluation data). A total of 10 normal-hearing individuals participated in the annotation of the development clips. This group consisted of the eight annotators of (**pool B**), along with two additional annotators (**pool C**). As a result, each sample in the development set received 10 binary annotations.

4.2.3.Summary

Table 2 succinctly recapitulates the key aspects of the distinct procedures employed in the generation and annotation of development and evaluation sound clips. Additionally, Table 3 provides a summary of the divergences in the content of the two subsets following the data collection process.

	Development	**Evaluation**
Alarm and background pairing	Random	In predefined list
Selection of the SNR	Random in ⟦– 30 ; 15⟧ dB	List of 6 values – 25, – 15, – 10, – 5, 0 and 10 dB
Selection of the background noise level	Random in {60 ; 80} dBA	List of 2 values 60 and 80 dBA
Number of presentations for each clip	1	3

Table 2: Summary of the procedures used to generate and annotate sound clips for development and evaluation sets.

	Development	**Evaluation**
Number of sound clips	2000	360
Number of different alarms sounds	70	15
Number of different background sounds	52	10
Annotators	10	20
Total number of annotations per clip	10	60

Table 3: Differences between the contents of the development and evaluation sets.

4.3. Labels and evaluation metrics

4.3.1.Labeling strategies

Measuring the model ability to correctly classify positive (i.e. “clearly audible”) and negative (i.e. not “clearly audible”) samples relies on the prior definition of the expected class associated with each sample. This expected class is typically denoted as the label or ground truth. However, in our specific problem, inferring the “truth” is challenging due to the subjective nature of perceiving an alarm as “clearly audible”. As such, each sample in the dataset comes with multiple annotations, which may not all be identical. Consequently, the question arises of how to derive single binary labels from these diverse annotations. In our study, we employed three distinct strategies to label the data. These strategies are elucidated below.

Average Psychometric Function: The evaluation data are the most comprehensive in our dataset, as each sample has undergone three annotations by each annotator. In addition, the listening experiment encompasses multiple SNRs, enabling the fitting of psychometric functions. The first labeling approach consists in fitting the individual psychometric curves of all annotators and subsequently averaging them across the annotators for each combination of alarm, background, and noise level. Then, for each sample within the evaluation set, we extract the corresponding value from the average psychometric function at the given SNR. This continuous value is eventually transformed into a binary label by applying a binarization threshold, which is set to 0.5 by default. We term this strategy the Average Psychometric Function (APF) strategy. This approach maximizes the utilization of all annotations collected for each sample, drawing upon psychometric functions that are widely used and easily interpretable in the field of psychoacoustics. It is used as the main labeling method to evaluate the model on evaluation data.

Majority Voting: The annotation procedure employed for the development data precludes the fitting of psychometric functions. Consequently, an alternative labeling strategy is required, applicable to both the development and evaluation sets, allowing for meaningful comparisons. We propose majority voting (MV), which is a widely used and straightforward labeling method, particularly suitable for crowdsourced data [49,50]. It relies on binary individual annotator responses, in contrast with the continuous individual values of APF. For development data, the binary responses provided by the 10 annotators are aggregated for each sample, taking the

majority opinion as the final label. As for the evaluation data, where each annotator has provided three binary responses per sample, we average and binarize these individual responses. Following that, we aggregate the binarized responses from all annotators, again using the majority opinion.

Random Drawing: Lastly, the third labeling method is exclusively employed during development. It consists in randomly drawing, for each example in the development set, one of the 10 available annotations and use it as the actual label. This method simulates a worst-case scenario where we must learn a model with a limited number of annotations. This limitation arises due to the time-consuming nature of the annotation process, which may lead to the temptation of employing a single annotator per development example to accelerate data collection. It is worth noting that the binary response given by a single annotator is a noisy estimate of the consensus, potentially resulting in some mislabeled samples. However, it has been observed that these noisy labels act as a form of regularization, preventing overfitting to the training data rather than compromising the overall classification performance [32]. This method is referred to as the random drawing (RD) strategy in the remainder of the paper.

4.3.2.Metrics

In order to assess the classification performance, we employ two widely used metrics: the area under the receiver operating characteristic curve (AUROC) and the F1-score. The AUROC quantifies the overall performance of a binary classifier. It measures the ability of the model to discriminate the positive and negative instances. It is determined through the integration of the receiver operating characteristic (ROC) curve, which depicts the relationship between the model true positive rate and false positive rate. This integration summarizes the model overall classification performance in a single metric ranging from 0 to 1. The F1-score is a measure of the model performance, taking into account both precision and recall. It balances the trade-off between correctly identifying positive samples (precision) and capturing all positive samples (recall). The F1-score ranges from 0 to 1, with 1 indicating perfect precision and recall.

4.3.3.Human baseline performance

To establish a benchmark for model performance comparison, we suggest evaluating the metrics against the performance of an average human. This approach will serve as a baseline. To achieve this, we calculate the evaluation metrics for each annotator in the dataset, while employing the remaining annotators to derive the reference labels on each iteration. Subsequently, we compute the mean and the standard error across all annotators for both metrics.

We first calculate the human baseline performance using MV labels to enable comparison between the development and evaluation sets. The results, reported in Table 4, demonstrate consistently high performance in both subsets. The similarity in scale between development and evaluation performance suggests that one subset should not be significantly more challenging to predict than the other. However, we do notice a slight discrepancy between the two subsets, indicating that, on average, there is less agreement between a single annotator and a reference annotator group for the evaluation data.

	Development	**Evaluation**
AUROC	87.68 ± 1.71	84.53 ± 1.79
F1	87.70 ± 1.79	82.78 ± 2.82

Table 4: Performance metrics computed for the human baseline on development and evaluation sets.

Second, we proceed to compare the human baseline performance on the evaluation data, computed with two different label types: MV labels and APF labels. As presented in Table 5, the labeling strategy acts differently on AUROC and F1-score. While the F1-score remains consistent regardless of the labeling strategy, the AUROC significantly increases when APF labels are used. This disparity arises from the inherent differences in how these metrics are computed. Both metrics require binary reference labels, but they impose different constraints on the individual responses (outputs) of the annotator whose performance is being assessed.

The F1-score requires binary outputs, as it relies on precision and recall, which are computed from binary predictions obtained by applying a binarization threshold to the continuous outputs. Conversely, the AUROC can handle continuous outputs since it examines the classifier ability to distinguish between classes across various discrimination thresholds. Consequently, for both MV and APF labeling methods, the F1-score is computed with

binarized individual annotator responses. The only difference lies in the accuracy of individual responses, which are binarized from the full psychometric function, rather than relying just on three Yes-No trials. In contrast, the AUROC is computed with continuous individual outputs for APF labels, against binary individual responses for MV labels. As a result, the AUROC takes advantage of the entire range of continuous response probabilities provided by the psychometric function and is not affected by the choice of the binarization threshold applied to the individual annotator responses.

This last observation underlines the significance of using APF labels to evaluate the model on evaluation data on two counts. First, it leverages comprehensive knowledge of the psychometric functions of the annotators, thereby elevating the human baseline score in terms of AUROC and enhancing the interpretability of this metric. Second, it provides justification for utilizing both AUROC and F1-score metrics. While the F1-score directly measures the binary classification performance, the AUROC facilitates the comparison of the classifier continuous output with the psychometric function of an average human.

	MV	APF
AUROC	84.53 ± 1.79	97.01 ± 0.49
F1	82.78 ± 2.82	83.48 ± 2.93

Table 5: Performance metrics computed for the human baseline on evaluation data, using MV and APF labels.

4.4. System

Our system uses 5.5-second sound clips to generate an estimate, at the clip-level, of whether the alarm within the sound clip is “clearly audible” or not. The process consists of two distinct stages: feature extraction and classification. In the first stage, spectro-temporal representations of the sound clips are extracted. These representations are then fed into a classifier in the second stage, which generates audibility predictions for each clip. The general structure of the system is depicted in Figure 5.

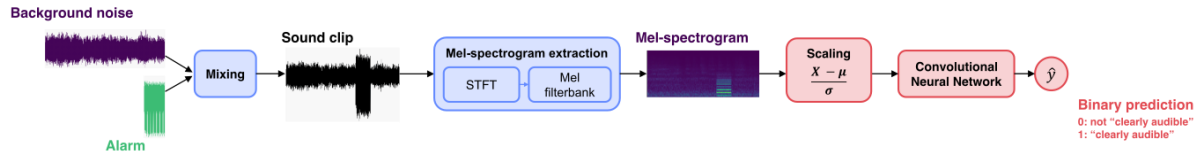

Figure 5: Illustration of the different stages of the system.

4.4.1.Input acoustic features

The input acoustic features used in our study are mel-spectrograms, which find extensive application in machine listening. To generate the mel-spectrogram for a combination of an alarm and background noise, we compute a short-time Fourier transform (STFT) using a Hamming window with a segment size of 1024 and a hop size of 512. Subsequently, we apply a mel-filterbank over the resulting spectrogram to convert it into 64 mel-spaced frequency bins, covering the range from 20 Hz to 22.05 kHz. With only 64 frequency bins, these mel-spectrograms offer more compact representations than conventional magnitude spectrograms, making them more computationally efficient. Additionally, they are based on the Mel scale, which makes them more aligned with human perception.

4.4.2.Model architecture

The proposed classifier is a CNN whose architecture is inspired by models used in rare sound event detection [51,52]. A similar architecture has demonstrated ability to reach high classification performance in two preliminary studies [31,32].

The model is composed of four convolutional layers, each having a different number of filters: 32, 64, 64, and 128 filters per layer, respectively. These filters are essentially small grids used to scan the input data. They

have a receptive field size of 3 by 3, meaning they analyze a 3-by-3 section of the input mel-spectrograms at a time.

Following each convolutional layer, Rectified Linear Unit (ReLU) activations [53] are applied, and max pooling is performed along the frequency axis. ReLU is a mathematical function that introduces non-linearity into the model, thereby facilitating the capture of complex patterns and features within the data. Max pooling is a downsampling technique that retains the maximum value from a set of values within a certain region. The pooling operations have different strides for each layer, which control how the pooling regions overlap or skip. Specifically, the strides are (1, 4), (1, 4), (1, 2), and (1, 2) for the respective layers. This operation reduces the dimensions of the data, thus decreasing computational complexity and memory usage across layers while preserving essential features.

The activation outputs from the final convolutional layer are then fed into a fully connected layer that processes each time frame separately, followed with a sigmoid activation function. This step transforms the data into a one-dimensional vector representing frame-level posteriors, where each value in the vector ranges between 0 and 1 due to the sigmoid activation.

To obtain a clip-level score, an L_p aggregation with $p=5$ is applied over the time axis on this vector. This process aggregates the information across time to make a decision at the level of the entire sound clip. The resulting score is a continuous value that aims to be as close to 1 as possible when the alarm is considered "clearly audible" and close to 0 when it is not. By default, we set the discrimination threshold to 0.5, meaning that the value is simply rounded to yield a binary output.

4.4.3.Development procedure

In order to prevent overfitting, the model is not trained using the entire development set. Instead, a portion of the development data is reserved for continuous monitoring of the model performance on data that the model has not learned from. This practice is commonly referred to as validation. In our work, we randomly allocated 20% of the development data for validation purposes, while the remaining 80% was exclusively designated for training. This random allocation was performed once and kept fixed throughout the study.

In accordance with the commonly recommended practice in deep learning approaches, especially for accelerating the convergence of the training process [54], we scale the data to zero mean and unit variance along mel frequency bins prior to feeding it into the model. The standardization coefficients are computed on the training samples. The same standardization coefficients are used to scale the whole dataset, ensuring uniform treatment of both development and evaluation data.

The CNN is trained with backpropagation, a fundamental deep learning method for adjusting model parameters to minimize prediction errors. The training samples are fed into the model in mini-batches of size 16. Within each mini-batch, we quantify the prediction error of the model by the means of a binary cross-entropy loss function. We then calculate the gradients of the loss function with respect to the internal parameters of the model. These gradients indicate the direction and magnitude of the adjustments required to minimize the loss. Using Adam optimizer [55], we leverage the gradient information to iteratively update the parameters, progressively improving model predictions. We set the learning rate, which determines the step size during weight update, to 0.0001.

To prevent the model from overfitting the training data, we fix a 0.0001 weight decay within the Adam optimizer. Weight decay imposes a penalty on the model parameters, discouraging them from growing too large during training. Additionally, we apply dropout with a rate of 0.25. Dropout randomly deactivates a fraction of neurons in the network during training, preventing the model from overly relying on any individual neuron or group of neurons.

The model is trained for 250 epochs, with each epoch representing the complete processing of all mini-batches of training data through the model. Throughout training, the model performance on the validation data is evaluated after each epoch, using RD labels. The final parameters of the models are retained from the epoch with the highest validation accuracy for subsequent evaluations and predictions.

4.5. Model performance

To capture the variance and uncertainty in the training process resulting from weight initialization and data sampling, we perform 10 runs of model training, each with different random initialization and data sampling. Before each training run of the network, a random draw is performed to determine the training RD labels.

In order to guarantee that the evaluation of the model performance on the evaluation set is only based on its capacity to generalize to unseen data, we employ separate and non-overlapping groups of annotators for

development and evaluation. For this purpose, we derive the APF labels for the evaluation set exclusively from the responses provided by the 12 annotators of (**pool A**). Meanwhile, our development procedure relies on RD labels sourced from the eight annotators of (**pool B**) and the two annotators of (**pool C**). This way, we aim to mitigate any potential annotator biases that may have been learned during development and ensure a fair and reliable evaluation.

We report both the model and human performance on the evaluation set with mean value and standard error across the 10 runs in Table 6. As we can observe, the model performance closely approximates that of humans, albeit with a slight deficit in comparison to the human baseline. The disparity between model and human baseline is much more pronounced and significant for the AUROC metric than it is for the F1-score, suggesting comparable binary classification performance in terms of precision and recall. This observation is further supported by the AUROC value of the model, which aligns closely with that of the human baseline when evaluated on binary MV labels.

	Model	Human baseline
AUROC	85.84 ± 0.75	97.28 ± 0.48
F1	76.03 ± 0.57	81.50 ± 3.86

Table 6: Performance on the evaluation set for the model and the human baseline.

5. COMPARATIVE ANALYSIS AND DISCUSSION

The evaluation of the system using the two classification metrics has provided valuable insights into the achievable performance using a low-complexity deep-learning model for predicting alarm audibility. In this section, we expand upon this initial analysis by conducting a more comprehensive examination of the model performance, drawing comparisons with the human baseline.

5.1. Audibility criterion and binary classification performance

First, our specific focus centers on the binarization threshold employed to generate the APF labels within the evaluation set. For the model evaluation presented in Table 6, the evaluation label binarization threshold was set at a default value of 0.5. This choice, guided by pragmatic technical factors, corresponds to a rounded mean audibility score for each sample and ensures a balanced evaluation set. Additionally, this threshold is closer to a majority voting approach, as it defines samples as positive if that have a 50% or higher probability of being considered as “clearly audible”, on average, by the participants. Nevertheless, it may not be the most suitable in terms of practical applicability of the model. While the model holds potential for alarm level configuration, the evaluation label binarization threshold, governing the audibility score at which an alarm becomes labelled as “clearly audible”, should be selected with care. Opting for a threshold of 0.5 adopts a rather permissive criterion, as all samples with an audibility score above 50% would be considered as “clearly audible”. This liberal approach might prove inadequate for ensuring reliable audibility if the model was to be employed in setting alarm levels.

We conducted a study of the effects resulting from different binarization thresholds applied to obtain the evaluation labels. We compare the effects on both the model and the human baseline performance. To evaluate the model, we analyzed the outputs obtained from the 10 runs, as presented in Section 4, by comparing them against the APF labels subjected to different evaluation label binarization thresholds. The evaluation was conducted based on precision, recall and F1-score. The results are presented in Figure 6.

In Figure 6(a), the precision demonstrates a monotonic decline with increasing evaluation label binarization thresholds, signaling an increase in false positives. This trend is consistent for both the model and the baseline. From the perspective of the model, this outcome can be attributed to the fact that, as the evaluation label binarization threshold increases, the number of positive samples in the evaluation set decreases. Since the model discrimination threshold remains fixed, its outputs do not change with the binarization threshold of the labels. Consequently, samples that were correctly classified as positive with a low binarization threshold gradually become false positives as the threshold increases. In contrast, the interpretation of decreasing precision in the human baseline performance differs. Since the binarization threshold is identical for both individual annotators’ outputs and reference labels, the diminishing precision with increasing binarization threshold reflects a distinct mechanism. This shows that, on average, a reference group of listeners adopts a more conservative stance compared to an individual listener when assessing alarm audibility. In essence, as we deviate from an audibility

score of 0.5 toward higher values, an individual listener's audibility score tends to overestimate that of a collective listener group. This discrepancy leads to an increased number of false positives in the human baseline. Furthermore, our analysis reveals that the model consistently falls short of the baseline precision across the entire range of evaluation label binarization thresholds.

The recall, depicted in Figure 6(b), shows a different evolution. As the evaluation label binarization threshold increases, more samples in the evaluation set are labeled as negative. This results in a higher recall for the model, as it correctly identifies a larger proportion of positive samples. In contrast, the baseline recall remains steady throughout the range of evaluation label binarization thresholds. Regarding the comparison between the model and the baseline, the baseline initially outperforms the model up to an evaluation label binarization threshold of approximately 0.6. Beyond this threshold, the model recall continues to increase, while the baseline's remains constant. Consequently, the model gradually surpasses the baseline in recall as the evaluation label binarization threshold increases.

The F1-score, which is the harmonic mean between precision and recall, is represented in Figure 6(c). Consistent with the results presented in Table 6, the model is outperformed by the baseline when an evaluation label binarization threshold of 0.5 is used. Between threshold values of 0.5 and 0.6, the model F1-score remains constant, while the baseline performance slightly declines. At a binarization threshold of 0.6, both the model and baseline achieve equivalent F1-scores, and from there, both show a gradual decrease until they reach a value of 0 for a binarization threshold of 1. The model performance approaching the baseline for higher evaluation label binarization threshold values is highly promising. This observation indicates that the model performs well under more stringent audibility criteria, crucial for ensuring reliable assessments, as they represent more prudent operating conditions for setting alarm levels.

These comparable trends in metrics for the model and the human baseline suggest a substantial similarity in their behavior. The model performance closely resembles that of an average normal-hearing listener in a Yes-No listening task, which is consistent with its learning on binary labels using a single normal-hearing listener per example. However, it is important to note that a single listener cannot accurately predict the response of a reference group of listeners, highlighting the necessity of involving multiple participants in psychoacoustic experiments. This limitation becomes more evident when the audibility criterion defined by the evaluation label binarization threshold is increased, resulting in a notable collapse in precision for both the baseline and the model. Reduced precision signifies an increased occurrence of false positives within positive predictions. This indicates a highly undesired scenario, where alarms are classified as “clearly audible” when they are not, which could potentially compromise user safety.

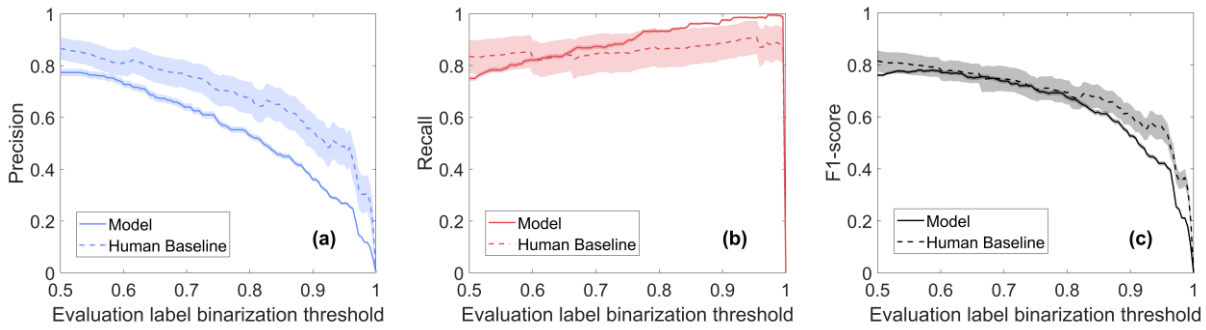

Figure 6: Precision, recall and F1-score for model and human baseline, computed with different evaluation label binarization thresholds.

5.2. Psychoacoustical perspective on model evaluation

The previous diagnosis, considering precision and recall, offers insights into the model and human baseline performance on the specific binary classification problem formulated in this paper. However, these metrics are limited since they are computed for a given discrimination threshold, and therefore only provide a partial view of the system. In particular, the decrease in precision observed in Figure 6(a) could be simply attributed to a non-optimal discrimination threshold that would cause wrong predictions.

As the model binary predictions rely on continuous outputs that are subjected to a discrimination threshold, we decide to focus on the continuous output of the model before applying any threshold. Figure 7 depicts the mean

outputs of the model, along with the mean values of the average psychometric function, both plotted against SNR across all evaluation clips. Before analyzing the curves, it should be pointed out that the human listeners and the model are not evaluated in the same exact way. The human psychometric function is influenced by multiple factors. Such factors may include personal experience or familiarity with a specific type of alarm or sound environment, but also the sounds heard during the listening test itself [56]. In our data, subjects-related factors were mitigated by averaging the responses within a representative group of listeners. Regarding the continuous aspect of human learning, it leads to a psychometric function that is potentially influenced by the experimental design. For instance, it is likely that hearing an alarm at a very high SNR after lower SNR values (or vice versa) might have had an effect on the participants' responses. We tried to limit this effect as much as possible by randomizing the presentation order of the multiple conditions during the listening experiment. The way the model produces outputs is different. While the model output is also influenced by what it has seen in the past, its parameters are fixed at inference time. This means that the model does not change after being presented with evaluation clips [57]. Evaluating the model does not have any impact on future predictions. This difference with human should be kept in mind when comparing the model continuous output to a human psychometric function.

With that being said, noteworthy parallels do exist in the responses of model and human. One shared aspect is the susceptibility to order effects. Although such effects are not related to the evaluation data for the model, the training can be influenced by the sequential arrangement of the training samples. This is comparable to the phenomenon described with the human responses. To address this issue, we employed sample shuffling during the training process, just as we randomized the presentation order in human experiments. Furthermore, an additional point of convergence between human and model responses is their shared reliance on temporal context. The results of our psychoacoustic experiment showed that human assessment of audibility is influenced by how distinctive the alarm is from its surrounding acoustic environment. Human therefore uses what is heard before and after an alarm to evaluate its audibility. Likewise, the model uses temporal context to produce an output. First, the CNN filters have a receptive field that spans over multiple time frames of the input mel-spectrograms. This feature enables the model to capture local temporal context to produce frame level estimates of audibility. Second, the clip-level model output is obtained through the aggregation of these frame-level estimates. In essence, the final model estimate of audibility incorporates information from the entire 5.5-second long sound clip, including parts after and before the alarm. These similarities between the model output and human responses offer scope for a meaningful comparative analysis.

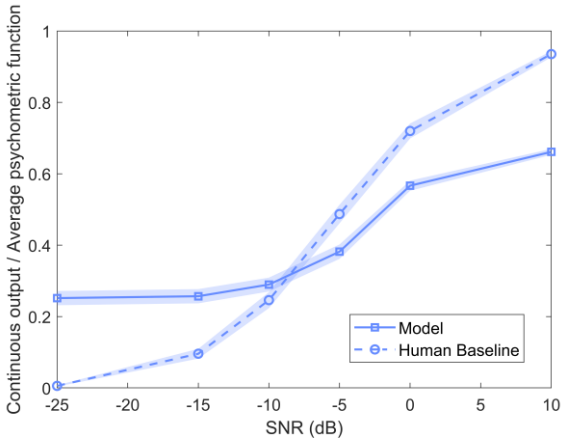

Figure 7: Average human baseline psychometric curve and model continuous output values over all the evaluation clips

Figure 7 illustrates the model and human baseline behavior as a function of the SNR. Interestingly, while the model is only trained to generate outputs that approach 0 or 1, we observe that the model output evolution with alarm signal strength exhibits similarities to the shape of a psychometric function. However, the two curves show distinct horizontal position, width and dynamic range, making difficult to establish a direct relationship between the model continuous output and a human psychometric value. As a result, while it would make sense choosing a specific discrimination threshold that is directly based on the value of psychometric function for the human baseline, selecting a discrimination threshold for the model is not as straightforward. A comprehensive analysis of the system should therefore consider the plurality of possible discrimination thresholds, rather than being restricted to a single operating point.

The model and human baseline ability to discriminate between positive and negative samples based on their respective continuous "psychometric outputs" can be visualized with ROC curves. ROC curves represent the trade-

off between true positive rate and false positive rate across different discrimination threshold values. Each point of the ROC curve is obtained by setting a discrimination threshold along the "psychometric curve", where alarms are classified as either "clearly audible" when they fall above the threshold or not "clearly audible" when they fall below it, and subsequently comparing the predictions with the actual evaluation labels. The upper right corner of the ROC curve represents the lowest possible discrimination threshold, classifying all samples as positive, resulting in a true positive rate of 1 and a false positive rate of 1. Conversely, the lower left corner of the ROC curve corresponds to the highest possible discrimination threshold, leading to a true positive rate of 0 and a false positive rate of 0 as all samples are classified as negative.

Figure 8 shows ROC curves of the model (left panel) and human baseline (right panel) for different evaluation label binarization thresholds. This allows us to visualize the model performance for different values of the binarization threshold used to define the labels, independently of the choice of a discrimination threshold. As we can observe in Figure 8, the human baseline ROC curves are unaffected by the choice of a different evaluation label binarization threshold. This is likely due to the similarity in shape between the psychometric curve of an individual and the average curve of a reference group. Contrastively, a closer look at the model ROC curves shows that when the evaluation label binarization threshold is higher, the model achieves higher true positive rates at lower false positive rates. However, even though, in proportion, the model is able to correctly identify more positive samples while predicting fewer false positives, it must be said that the absolute number of positive samples in the dataset considerably decreases when the binarization is increased. Specifically, the number of positive samples in the dataset changes from 148 for a binarization threshold of 0.5 to 53 for a binarization threshold of 0.9. This indicates that the model performs very well at discriminating "easy" positive samples, i.e. samples where the alarm is very likely to be considered by a group of human listeners as "clearly audible". On the other hand, it is less able to identify all positive samples without making false positive predictions when the samples are defined as positive from a lower average audibility score. In both cases, however, model performance remains similar in scale to that of the human baseline, which is a satisfactory result.

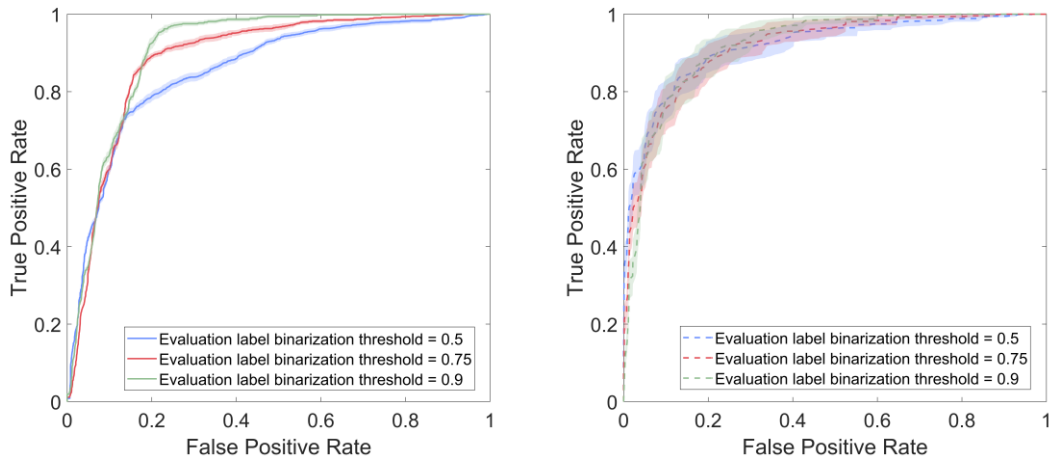

Figure 8: ROC curves of the model (left panel) and human baseline (right panel) for three different test label binarization thresholds (0.5, 0.75, and 0.9).

These results highlight the model capacity to predict alarm audibility consistent with human perception. The model continuous output can be leveraged to discriminate between "clearly audible" and not "clearly audible" alarms for various audibility criteria, provided the selection of an appropriate operating mode. Such operating mode should be selected depending on the specific application framework of the model, including tolerance to false positives or false negatives for instance. Finally, a binary response may not be the most appropriate way to characterize human auditory perception. The analysis using the ROC curves shows that, even if the model was trained on a binary classification task, it produces a continuous output that could be exploited in a way that better reflects perception, which is often more nuanced and complex. Further developments should explore strategies to align the model outputs more closely with human psychometric functions, particularly focusing on the high-probability region of the curves. The formulation of safe alarm level recommendations should rely on this specific region. It is therefore imperative that the model output emulates human behavior in this critical area as closely as possible. In particular, a safe application of the model requires being able to properly select a discrimination threshold that minimizes the number of false positive predictions while maintaining a high overall performance

level. This would enable enhancing the potential of the model to provide well-informed recommendations concerning alarm levels by avoiding positively classifying alarms that are not “clearly audible”.

6. CONCLUSION

In this study, we presented an experiment that provided subjective evaluations of the audibility of acoustic alarms in occupational environments. Following the terms of the ISO 7731 standard, we assessed the “clearly audible” aspect of these alarms, and additionally measured their detection thresholds. Our analysis of these data has underscored the true importance of assessing audibility at supraliminal levels as a distinctive property that goes beyond direct and exclusive correlation with detectability.

By contrasting our findings with established standards and commonly employed criteria, we have formulated general guidelines to ensure a reliable audibility of acoustic alarms, thereby corroborating existing criticisms that the requirements of ISO 7731 are not always necessary and may lead to excessively high alarm levels. In particular, we proposed revising the current SNR criterion of minimum 15 dB to a lower value, such as 7.5 dB. An alternative option could be to carry additional investigations to implement an adaptive criterion that adjusts according to the ambient noise level, which would account for the reduced need for high alarm levels in environments with higher noise levels. As stated in the standard, such audibility criteria are designed for normal-hearing people with no hearing protection. It is crucial to carefully adapt these criteria in situations where individuals with hearing impairments or personal hearing protection are present.

Furthermore, through our acoustic feature analysis, we have identified some amount of correlation between audibility and salience. This analysis has shed light on the fundamental acoustic attributes that drive audibility and has highlighted the pivotal role of distinctiveness in enhancing the perceived audibility of alarms.

Audibility is contingent upon an array of multiple factors, whose individual and cumulative effects remain incompletely understood and quantified, making it difficult to model using a purely psychophysical parametric approach. As a solution, we developed a model based on data-driven deep learning techniques to assess the audibility of acoustic alarms in contexts where subjective testing may not be employed. We described the collection of an extensive dataset of perceptually annotated sound clips mixing acoustic alarms with occupational noise, necessary to learn and evaluate the model. Our approach was shown to achieve close-to-human performance at classifying alarms as “clearly audible” or not “clearly audible”.

A more detailed analysis determined the most favorable operating conditions of the model in its most straightforward application as a binary classifier, evaluating the model performance using various audibility criteria to label the evaluation data. It showed that the agreement between model and human performance was stronger when the minimum audibility score required to label alarms as “clearly audible” was high. However, both the model and human performance collapsed when this minimum audibility score was set to extremely high values.

Eventually, a closer examination of the model outputs evidenced similarities between model and human responses, offering a more intuitive interpretation of the ROC curves from a psychophysical point of view. This analysis, encompassing all possible discrimination thresholds of the model, offered a broader view on model performance. It revealed that the model is actually able to perform well at discriminating “clearly audible” alarms, even better when high audibility scores are required to label alarms as “clearly audible”, provided a proper configuration of its decision stage. However, despite the existing similarities between the model and human responses, our approach is open to development to align more closely the model behavior with that of humans for more reliable model predictions.

In conclusion, our findings demonstrated the great potential of a deep learning model to solve the problem of setting appropriate alarm levels in work environments. At this point, we have a binary classifier capable of discriminating between “clearly audible” and not “clearly audible” alarms that demonstrates strong correlation with human responses. This model could serve as a basis for formulating alarm level recommendations. At the current stage of development, this could be done by adjusting alarm levels based on the SNRs at which the model predictions turn from negative to positive. Additional research is needed to refine this approach. For instance, the development of a decision-making framework that leverages the continuous output of the classifier, as opposed to the fixed threshold currently in use, could be beneficial. Alternatively, one may consider training a deep-learning model on another task, such as regression to predict psychometric functions. However, training a model on psychometric functions would require new and more extensive development data, annotated through a more demanding and time-consuming procedure.

ACKNOWLEDGEMENTS

The authors extend their sincere gratitude to Dr. Mounya Elhilali for generously sharing the code used for acoustic feature analysis, as well as Baptiste Bouvier and Valerian Fraisse, who contributed some of the stimuli used in the psychoacoustic experiment.

This research has been conducted within the framework of the Labex CeLyA (Lyon Center of Acoustics ANR-10-LABX-60).

REFERENCES

[1] Feder K, Michaud D, McNamee J, Fitzpatrick E, Davies H, Leroux T. Prevalence of Hazardous Occupational Noise Exposure, Hearing Loss, and Hearing Protection Usage Among a Representative Sample of Working Canadians. Journal of Occupational and Environmental Medicine 2017;59.

[2] Green DR, Masterson EA, Themann CL. Prevalence of hearing protection device non-use among noise-exposed US workers in 2007 and 2014. American Journal of Industrial Medicine 2021;64:1002–17. https://doi.org/10.1002/ajim.23291.

[3] Le TN, Straatman LV, Lea J, Westerberg B. Current insights in noise-induced hearing loss: a literature review of the underlying mechanism, pathophysiology, asymmetry, and management options. Journal of Otolaryngology - Head & Neck Surgery 2017;46:41. https://doi.org/10.1186/s40463-017-0219-x.

[4] Chen K-H, Su S-B, Chen K-T. An overview of occupational noise-induced hearing loss among workers: epidemiology, pathogenesis, and preventive measures. Environmental Health and Preventive Medicine 2020;25:65. https://doi.org/10.1186/s12199-020-00906-0.

[5] Basner M, Babisch W, Davis A, Brink M, Clark C, Janssen S, et al. Auditory and non-auditory effects of noise on health. The Lancet 2014;383:1325–32. https://doi.org/10.1016/S0140-6736(13)61613-X.

[6] Themann CL, Masterson EA. Occupational noise exposure: A review of its effects, epidemiology, and impact with recommendations for reducing its burden. The Journal of the Acoustical Society of America 2019;146:3879–905. https://doi.org/10.1121/1.5134465.

[7] Wilkins PA, Acton WI. Noise and Accidents — A Review. The Annals of Occupational Hygiene 1982;25:249–60. https://doi.org/10.1093/annhyg/25.3.249.

[8] Dias A, Cordeiro R. Fraction of work-related accidents attributable to occupational noise in the city of Botucatu, São Paulo, Brazil. Noise Health 2008;10:69–73. https://doi.org/10.4103/1463-1741.44344.

[9] Deshaies P, Martin R, Belzile D, Fortier P, Laroche C, Leroux T, et al. Noise as an explanatory factor in work-related fatality reports. Noise Health 2015;17:294–9. https://doi.org/10.4103/1463-1741.165050.

[10] Dzhambov A, Dimitrova D. Occupational Noise Exposure and the Risk for Work-Related Injury: A Systematic Review and Meta-analysis. Annals of Work Exposures and Health 2017;61:1037–53. https://doi.org/10.1093/annweh/wxx078.

[11] ISO 7731 — Ergonomics — Danger signals for public and work areas — Auditory danger signals 2003.

[12] Żera J, Nagórski A. Preferred Levels of Auditory Danger Signals. International Journal of Occupational Safety and Ergonomics 2000;6:111–7. https://doi.org/10.1080/10803548.2000.11105112.

[13] Dolan TG, Rainey JE. Audibility of Train Horns in Passenger Vehicles. Hum Factors 2005;47:613–29. https://doi.org/10.1518/001872005774859999.

[14] International Organization for Standardization (ISO). ISO 9533 — Earth-moving machinery — Machine-mounted audible travel alarms and forward horns — Test methods and performance criteria 2010.

[15] Laroche C, Vaillancourt V, Giguère C, Ellaham N, Gagnon C, Laflamme P, et al. Detection of Reverse Alarms In Noisy Worplaces. Sound And Vibration. International Congress. 22nd 2015. (ICSV 22), vol. 4, Florence, Italy: International Institute of Acoustics and Vibration (IIAV); 2015, p. 3292–9.

[16] Laroche C, Giguère C, Vaillancourt V, Roy K, Pageot L-P, Nélisse H, et al. Detection and reaction thresholds for reverse alarms in noise with and without passive hearing protection. International Journal of Audiology 2018;57:S51–60. https://doi.org/10.1080/14992027.2017.1400188.

[17] Vaillancourt V, Nélisse H, Laroche C, Giguère C, Boutin J, Laferrière P. Comparison of sound propagation and perception of three types of backup alarms with regards to worker safety. Noise Health 2013;15:420–36.

[18] Schell-Majoor L, Rennies J, Ewert SD, Kollmeier B. Application of psychophysical models for audibility prediction of technical signals in real-world background noise. Applied Acoustics 2015;88:44–51. https://doi.org/10.1016/j.apacoust.2014.08.001.

[19] Dau T, Püschel D, Kohlrausch A. A quantitative model of the ‘‘effective’’ signal processing in the auditory system. II. Simulations and measurements. The Journal of the Acoustical Society of America 1996;99:3623–31. https://doi.org/10.1121/1.414960.

[20] Glasberg BR, Moore BCJ. Development and Evaluation of a Model for Predicting the Audibility of Time-Varying Sounds in the Presence of Background Sound. Journal of the Audio Engineering Society 2005;53:906–18.

[21] Jepsen ML, Ewert SD, Dau T. A computational model of human auditory signal processing and perception. The Journal of the Acoustical Society of America 2008;124:422–38. https://doi.org/10.1121/1.2924135.

[22] Abeßer J. A Review of Deep Learning Based Methods for Acoustic Scene Classification. Applied Sciences 2020;10. https://doi.org/10.3390/app10062020.

[23] Xia X, Togneri R, Sohel F, Zhao Y, Huang D. A Survey: Neural Network-Based Deep Learning for Acoustic Event Detection. Circuits, Systems, and Signal Processing 2019;38:3433–53. https://doi.org/10.1007/s00034-019-01094-1.

[24] Grumiaux P-A, Kitić S, Girin L, Guérin A. A survey of sound source localization with deep learning methods. The Journal of the Acoustical Society of America 2022;152:107–51. https://doi.org/10.1121/10.0011809.
[25] Dohi K, Imoto K, Harada N, Niizumi D, Koizumi Y, Nishida T, et al. Description and Discussion on DCASE 2022 Challenge Task 2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques. Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), Nancy, France: 2022, p. 31–5.
[26] Huang N, Slaney M, Elhilali M. Connecting Deep Neural Networks to Physical, Perceptual, and Electrophysiological Auditory Signals. Frontiers in Neuroscience 2018;12.
[27] Wang L, Liu H, Zhang X, Zhao S, Guo L, Han J, et al. Exploring Hierarchical Auditory Representation via a Neural Encoding Model. Frontiers in Neuroscience 2022;16.
[28] Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 2018;98:630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044.
[29] Francl A, McDermott JH. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nature Human Behaviour 2022;6:111–33. https://doi.org/10.1038/s41562-021-01244-z.
[30] Saddler MR, Gonzalez R, McDermott JH. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. Nature Communications 2021;12:7278. https://doi.org/10.1038/s41467-021-27366-6.
[31] Effa F, Serizel R, Arz J-P, Grimault N. Convolutional Neural Network for Audibility Assessment of Acoustic Alarms. Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), Nancy, France: 2022, p. 36–40.
[32] Effa F, Serizel R, Arz J-P, Grimault N. Lightweight Annotation and Class Weight Training for Automatic Estimation of Alarm Audibility in Noise. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece: 2023.
[33] Wilkins PA. Assessing the Effectiveness of Auditory Warnings. British Journal of Audiology 1981;15:263–74. https://doi.org/10.3109/03005368109081448.
[34] Patterson RoyD, Mayfield TF, Broadbent DE, Baddeley AD, Reason J. Auditory warning sounds in the work environment. Philosophical Transactions of the Royal Society of London B, Biological Sciences 1997;327:485–92. https://doi.org/10.1098/rstb.1990.0091.
[35] Patterson RoyD, Milroy R. Existing and recommended levels for the auditory warnings on civil aircraft. Cambridge: MRC Applied Psychology Unit; 1979.
[36] Zheng Y, Giguère C, Laroche C, Sabourin C, Gagné A, Elyea M. A Psychoacoustical Model for Specifying the Level and Spectrum of Acoustic Warning Signals in the Workplace. Journal of Occupational and Environmental Hygiene 2007;4:87–98. https://doi.org/10.1080/15459620601115768.
[37] Kaya EM, Elhilali M. Investigating bottom-up auditory attention. Frontiers in Human Neuroscience 2014;8.
[38] Huang N, Elhilali M. Auditory salience using natural soundscapes. The Journal of the Acoustical Society of America 2017;141:2163–76. https://doi.org/10.1121/1.4979055.
[39] Kothinti SR, Huang N, Elhilali M. Auditory salience using natural scenes: An online study. The Journal of the Acoustical Society of America 2021;150:2952–66. https://doi.org/10.1121/10.0006750.
[40] Font F, Roma G, Serra X. Freesound technical demo. Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain: Association for Computing Machinery; 2013, p. 411–2. https://doi.org/10.1145/2502081.2502245.
[41] Sardin J. BigSoundBank by Joseph Sardin 2005. https://www. bigsoundbank.com.
[42] Lemaître G. Étude perceptive de nouveaux avertisseurs sonores automobiles. Université du Maine, 2004.
[43] Fraisse V, Nicolas E, Schütz N, Ribeiro C, Misdariis N. Évaluer l'impact d'installations sonores sur la perception du paysage sonore urbain : cas d'étude d'une place publique parisienne, 2022.
[44] Schütt HH, Harmeling S, Macke JH, Wichmann FA. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. Vision Research 2016;122:105–23. https://doi.org/10.1016/j.visres.2016.02.002.
[45] Studebaker Gerald A. A "Rationalized" Arcsine Transform. Journal of Speech, Language, and Hearing Research 1985;28:455–62. https://doi.org/10.1044/jshr.2803.455.
[46] Thévenet J, Papet L, Coureaud G, Boyer N, Levréro F, Grimault N, et al. Crocodile perception of distress in hominid baby cries. Proceedings of the Royal Society B: Biological Sciences 2023;290:20230201. https://doi.org/10.1098/rspb.2023.0201.
[47] Bertrand F, Myriam M, Meyer N. PlsRglm : Régression PLS et modèles linéaires généralisés sous R. 42èmes Journées de Statistique 2010.
[48] Bouvier B, Susini P, Marquis-Favre C, Misdariis N. Revealing the stimulus-driven component of attention through modulations of auditory salience by timbre attributes. Scientific Reports 2023;13:6842. https://doi.org/10.1038/s41598-023-33496-2.

[49] Yu H, Li J, Wu Z, Xu H, Zhu L. Two-step learning for crowdsourcing data classification. Multimedia Tools and Applications 2022;81:34401–16. https://doi.org/10.1007/s11042-022-12793-4.
[50] Martín-Morató I, Mesaros A. Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2023;31:902–14. https://doi.org/10.1109/TASLP.2022.3233468.
[51] Çakir E, Virtanen T. Convolutional Recurrent Neural Networks for Rare Sound Event Detection. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany: 2017, p. 27–31.
[52] Turpault N, Serizel R. Training Sound Event Detection on a Heterogeneous Dataset. Proceedings of the 5th Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), Tokyo, Japan: 2020, p. 200–4.
[53] Agarap AFM. Deep Learning using Rectified Linear Units (ReLU). arXiv Preprint arXiv:180308375 2018.
[54] LeCun YA, Bottou L, Orr GB, Müller K-R. Efficient BackProp. In: Montavon G, Orr GB, Müller K-R, editors. Neural Networks: Tricks of the Trade: Second Edition, Berlin, Heidelberg: Springer Berlin Heidelberg; 2012, p. 9–48. https://doi.org/10.1007/978-3-642-35289-8_3.
[55] Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings, San Diego, CA, USA: 2015.
[56] Gelfand SA. Hearing: An Introduction to Psychological and Physiological Acoustics, Sixth Edition. CRC Press; 2017.
[57] Aggarwal CC. Neural Networks and Deep Learning: A Textbook. 1st ed. Springer Publishing Company, Incorporated; 2018.