CONVOLUTIONAL NEURAL NETWORK FOR AUDIBILITY ASSESSMENT OF ACOUSTIC ALARMS

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

CONVOLUTIONAL NEURAL NETWORK FOR AUDIBILITY ASSESSMENT OF ACOUSTIC ALARMS

François Effa[1,2,3]*, Romain Serizel*[3]*, Jean-Pierre Arz*[1]*, Nicolas Grimault*[2]*,*

[1] Institut National de Recherche et de Sécurité, F-54000 Nancy, France
[2] Centre de Recherche en Neurosciences de Lyon, CNRS, F-69500 Bron, France
[3] Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

ABSTRACT

In noisy workplaces, the audibility of acoustic alarms is essential to ensure worker safety. In practice, some criteria are required in international standards to make sure that the alarms are "clearly audible". However, the recommendations may lead to overly loud alarms, thereby exposing workers to unnecessary high sound levels, especially when ambient sound levels are high themselves. For this reason, it appears necessary to properly assess the audibility of alarms at design stage. Existing psychoacoustical methods rely on repeated subjective measurements at different sound levels and therefore require time-consuming procedures. In addition, they must be repeated each time the alarm or sound environment changes. To overcome this issue, we propose a data-driven approach to estimate the audibility of new alarm signals without having to test each new condition experimentally. In this study, a convolutional neural network model is trained to perform a binary classification task on short sound clips labeled with the outcomes of psychoacoustical experiments. We propose a proof of concept of this approach and analyze its performance depending on the data used at training and the temporal context used by the networks to predict the audibility of the alarm.

***Index Terms*—** Acoustic Scene Classification, Warning signals, Psychophysics, Machine learning

1. INTRODUCTION

The audibility of acoustic signals indicating a danger – for instance, the reverse alarm of a construction machine – is essential to ensure the safety of workers and to prevent the risk of accidents. The international standard dedicated to auditory danger signals for public and work areas requires for the alarms to be "clearly audible" [1]. In order to meet this requirement, several criteria related to the ambient noise levels are proposed. In particular an overall level criterion imposes a minimum signal-to-noise ratio (SNR) of 15 dB. This recommendation is questionable on two counts. First, the proposed criteria appear to lead in practice to excessive sound levels when the ambient noise levels are high. Żera and Nagórski have found that, when asking listeners to adjust the level of alarms so that they are judged to be clearly audible, the required SNR varies continuously from 15 to −2 dB as the noise level increases from 60 dB to 90 dB [2]. An SNR of 15 dB could therefore be too conservative or even harmful at high noise levels. Second, there is no scientific or formal definition of what "clearly audible" means. We can easily understand the difference between a sound that is simply detectable and another, qualified as clearly audible. The alarm must not only be perceptible but also "loud" enough to provide an effective warning of danger. This judgment can however vary greatly from one individual to another. It should also be noted that such consideration is dependent on the sound environment in which the alarm occurs.

In the presence of background noise, our ability to detect a given sound is reduced. It translates into an increase in the audibility threshold of the target sound, meaning that the level at which the sound is just audible is higher in noise than in silence [3, 4]. The noise in question is called *masker*, and we refer to the audibility threshold when a masker is present as the *masked threshold*. The masking mechanism is the basis for current alarm design methods. With the understanding that the alarm level must be well above the masked threshold to ensure reliable audibility, it has been suggested that alarms should exceed the masked threshold by 10 to 15 dB [5]. This recommendation is also included in the standard [1]. Masked thresholds can be measured experimentally [6], but such approach is not the most convenient, since a psychoacoustical experiment requires the involvement of multiple human subjects, and measuring an audibility threshold implies covering a range of different sound levels, which can be time-consuming. Furthermore, the measurements strongly depend on the acoustic properties of the sounds. Consequently, the experiment must be repeated for any new condition of interest such as a new alarm signal, or different ambient noise type or level. In response to this problem, some models have been developed to predict masked thresholds [7, 8]. However, these are based on the explicit estimation of the masked thresholds, and therefore only efficient in well-controlled conditions.

The main motivation of our work is to propose a model capable of accurately estimating the audibility of acoustic alarms in noise. We intend this model to be applicable to a large variety of sound environments, including fluctuating noises and different types of warning signals. In that regard, we suggest that a deep learning approach, which is data-driven, would be suitable. A few studies have already begun to pave the way for connecting psychophysics and deep learning methods, for instance, by implementing a psychophysically inspired methodology for model evaluation [9]. More recently, in the context of handwritten document transcription, a loss function was reformulated to account for perceptual data [10]. The authors used it to train a CNN to perform character recognition and obtained consistent and repeatable performance improvement on standard datasets. In the field of audition, recent works have replicated human perceptual judgments with a high level of likeness in word recognition, genre recognition [11] and sound localization in natural environments [12].

In this paper, we propose to adapt methods used in ambient sound analysis [13, 14] to the specific topic of audibility assessment of auditory alarms. To that end, we introduce a new dataset consist-

ing of sound clips made up of auditory warning signals embedded in noise labelled through psychoacoustical experiments. Based on this dataset, we train a model to reproduce human judgment regarding the audibility of the alarms.

2. METHOD

Ambient sound analysis is a broad field of research that encompasses several sub-areas such as Acoustic Scene Classification (ASC) [15], Sound Event Dectection (SED) [16, 17], or audio tagging [18]. Although our approach does not fully fit into any of these categories, it does share certain aspects with some of them. First, similarly to what is done in ASC and audio tagging, we want the model to produce file-level estimations. Besides, the analysis focuses on the alarms, which are temporally localized sound events. This is pretty similar to SED or tagging. Despite this, we are not directly interested in the ability of the model to detect the alarm. Yet, if the model can predict that an alarm is audible, it gives an indication that the model did initially detect it. The opposite, however, is not true. It is quite possible that the model predicts that an alarm is not audible although it has detected the alarm itself.

To assess whether a particular alarm can be considered clearly audible in a given environment, we propose to frame the problem as a binary classification task. The proposed method uses 5.5-second audio clips containing auditory warning signals embedded in background noises as input. After extracting acoustical features from these signals, we used a CNN to produce a binary estimate of whether the alarm present in the clip is clearly audible or not. More details about the audio clip generation and the annotation process are provided in the next section.

3. PSYCHOACOUSTICAL DATA COLLECTION

The need for relatively large amounts of data with a sufficient variability is a well-known constraint associated with deep learning. To develop our model with psychoacoustical data, no ready-to-use dataset was available from external sources. Therefore, we had to collect one. To do this, our choices were guided by two major considerations. First, we had to find a way to collect the maximum amount of data at the lowest time cost. Second, we still wanted to keep the possibility of explaining all or part of our results within the psychoacoustical framework. We decided to divide the psychoacoustical experiments in three parts. The first part is following within-subject design, closer to psychoacoustical procedures. The second and third parts follow lighter procedures and lower data collection time cost. In this section, we describe the psychoacoustical experiments that have been carried out to collect the dataset. The dataset in this paper is a preliminary version used for a proof of concept. The final dataset is expected to be larger.

3.1. Stimuli and material

Stimuli were made of short alarm sounds (between 0.243 and 1.763 s long) embedded in background noises. The alarms were mostly synthetic signals, but some of them were clean recordings. Backgrounds were field recordings, taken to be industry-related (factory, roadworks, construction) or captured in noisy public spaces. Both alarms and noises were mono signals collected from different sources, mainly the Freesound database [19], BigSoundBank [20], and to a lesser extent, a published set of medical alarms [21] or self-recorded railway warning signals [22].

In the experiments, 5.5-second clips were generated, each containing a background noise and a single alarm with a random onset temporal location. The level of the noise varied between 60 dBA and 80 dBA. The SNRs were all taken between −30 dB and +15 dB. The stimuli were played at a sample rate of 44.1 kHz using an RME Babyface Pro sound card, and presented over Beyerdynamic DT 770 Pro headphones calibrated with Larson Davis AEC101 artificial ear and Model 824 sonometer.

3.2. Procedure and datasets

Twelve volunteers aged from 18 to 43 and free of reported hearing problems took part in this study. They came for multiple sessions of one or two hours each. All the participants were compensated for the time spent on the experiments. To evaluate the audibility, the subjects were presented with a clip made up of an alarm embedded in a background noise. At the end of the presentation, they had to answer the question "*Was the alarm clearly audible?*" by simply clicking *Yes* or *No*. Three different experiments were carried out and most of the subjects took part in each of the three experiments. Each experiment served to collect a separate subset designed for a specific purpose.

The first set consists of 6 audio clips, each declined in 2 different noise levels (60 dBA and 80 dBA) and 10 SNRs linearly spaced between −30 and +15 dB, making a total of 120 signals. In each of the 6 clips, the alarm signals and backgrounds are different from the other clips. All these stimuli have been annotated once by the 10 participants. As a result, it contains psychoacoustical data that are quite close to what would have been obtained through a standard procedure, yet it is often recommended to make more than one repetition per subject [1]. This set is the most controlled set and is selected as the evaluation set. The data collection process on this subset will eventually allow for further comparison with more standard psycho-acoustic experiments. The labels (0 or 1) have been obtained by setting a 0.5 threshold on the proportion of "*Yes*" across participant answers. This subset is referred to as *subA* in the remainder of the paper.

The second set contains 1800 audio clips, made with the same noises and alarms as in the first set, except that the 6 formerly used alarm-noise combinations were avoided. As a consequence, there was a total of 30 possible alarm-noise combinations. Six different SNRs ([−25, −10, −5, 0, 5, 12.5] in dB) and six noise levels ([60, 64, 68, 72, 76, 80] in dBA) were used and uniformly distributed among the clips. Each of the 1800 clips had a unique alarm onset location. Ten participants were involved in this experiment. Each of them listened to 180 clips. The distribution of the clips among subjects was done in such a way that each subject listened to the same number of clips per SNR and per noise level. No repetition was made across subjects, meaning each stimulus has been annotated just once. This subset is referred to as *subB*.

The data eventually collected in these experiments will be used to form a large training subset. However, the third set introduced in this paper is also composed of 1800 different clips. This is motivated by the desire to keep comparable size between this set and *subB*. There are 70 alarms and 52 background noises in this set, all different from the 6 used in the first and second sets. Two noise levels were used (60 dBA and 80 dBA). There were 46 different SNRs ranging from −30 to +15 dB with a step of 1 dB. Ten subjects contributed to the annotation. Some of the conditions have been randomly repeated among subjects, making a total of around 11500 annotation points. For clips with a single annotation point,

the subject's answer was kept as the final annotation. For clips with multiple annotation points, the labels were derived by setting a 0.5 on the *Yes*-rate. We refer to this subset as *subC*.

Subsets *subB* and *subC* were used separately for development, using 1440 training clips and 360 validation clips, which corresponds to a 80%/20% ratio. The training/validation split was performed randomly and kept fixed for all the experiments.

4. EXPERIMENTAL SETUP

4.1. Acoustic features

The signals were sampled at 44.1 kHz. The features we used are mel-spectrograms with 64 coefficients. They were extracted using a 1024-sample short-time Fourier transform (STFT), with 50% overlap and a Hamming window.

We did explore the idea of using more perceptually relevant representations such as cochleagrams [12] or spectro-temporal excitation patterns [23]. However, preliminary experiments with these features did not provide any significant performance improvement in terms of accuracy on the development sets. Therefore, in this paper we use only mel-spectrograms as input features.

For the experiments, the input representations were standardized to zero mean and unit variance along mel frequency bins. The standardization coefficients were computed over the whole training set.

4.2. Convolutional Neural Network

The architecture of the CNN used in this paper is inspired by models used in SED [14] and Bird Audio Detection [13]. The model is composed of 4 convolutional layers with [32, 64, 64, 128] filters per layer. Each filter has a 3-by-3 receptive field. Each convolutional layer is followed by ReLU activations and max pooling along the frequency axis ([1, 4], [1, 4], [1, 2] and [1, 2], respectively). The activation outputs from the last convolutional layer are stacked along frequency axis [13]. Preliminary experiments with recurrent layers did not lead to significant improvement. Therefore they are not used in this paper. Instead, we directly operate L_p aggregation over the time axis on the stacked representations. L_p aggregation with $p = 2$ was preferred over max pooling which is not differentiable and may lead to instability [24]. In addition, it has been shown to be more robust to variations in the relative duration of the alarm compared to the clip length [25]. The aggregation layer is followed by the classification layer that has one single neuron with sigmoid activation. The neuron is intended to produce an activation which is close to 1 when the alarm present in the clip is clearly audible, and close to 0 when the alarm is not clearly audible.

For training, back-propagation was performed using a binary cross-entropy loss function and Adam optimizer [26] with a learning rate of 0.0001. To reduce overfitting, dropout was applied on the outputs of all the convolutional layers with a rate of 0.25 and regularization was employed by fixing a 0.0001 weight decay in Adam. The model was trained for a maximum of 250 epochs and the epoch giving the best accuracy on validation set was kept.

4.3. Model evaluation

For the experiments, the model was evaluated from the area under the receiver operating characteristic curve (AUC) and the F1-score. We trained the models with 10 randomized initializations. The metrics were computed on the outputs obtained with these 10 models. We report the mean and 95% confidence intervals of the metrics.

5. RESULTS AND DISCUSSION

In this section, we report the two series of experiments that have been conducted on the model. The first series of experiments focuses on the model performance depending on the data used at training. The second series of experiments assesses the potential effects of the temporal context used in the model to predict the audibility.

5.1. Impact of the training data

Our first series of experiments investigates the performance of the models trained on *subB* and *subC*. As described in Section 3.2, the clips in *subA* and *subB* were made with the same alarms and background noises. For this reason, we expect the model to perform better when it is trained on *subB* than on *subC*.

At first, the model was trained on *subB* or *subC* data while *subB* validation data were used to select the model. Table 1 shows the AUC and F1-score on development and test sets. As we can observe, performance on the test set is better when *subB* is used for training. There are two potential causes for this. The difference in performance can be due to the fact that the alarm signals and backgrounds in *subC* are different from those in *subB* and in the test set or to the fact the task addressed in *subC* is more difficult than in *subB* (or both). We evaluated models trained on *subC* and validated on either *subB* or *subC* to verify this second hypothesis (see Table 2). The results show a significant difference in performance on development set depending on whether the model was validated on *subB* or *subC*. The high development score when *subC* is used for validation suggests that the model can be fitted to *subC* data, which indicates that the task addressed in *subC* is in fact not more difficult than in *subB*. However, the performance on test set shows that the model gives better results on test data when the alarms and background noises have been seen during training. This raises the question whether collecting more training data with a larger set of alarms and backgrounds can help to compensate for this performance gap. If not, it would induce the need to see test alarms or test backgrounds or both during training. In practice, such a scenario would not be realistic. Despite the difference in performance, it should be noted that when training on different alarms and backgrounds, the model performance does not actually collapse.

Subset		Development score	Test score
subB	AUC	94.4 ± 0.3	95.3 ± 0.7
	F1	91.9 ± 0.3	89.2 ± 1.6
subC	AUC	78.9 ± 1.0	87.9 ± 1.9
	F1	79.1 ± 1.2	79.3 ± 2.4

Table 1: AUC and F1-scores on development and evaluation sets with 95% confidence intervals. *subB* is used for validation.

5.2. Impact of the clip duration

The alarms that are present in the different clips have variable lengths. Since the longest alarm is less than 1.8 s long, all the 5.5-second clips contain both portions of noise alone and the whole section where the alarm occurs. It is still to be determined whether the model bases its predictions on the entire alarm or on a specific region of the alarm. This region could be the onset, for instance.

Validation set		Development score	Test score
subB	AUC	78.9 ± 1.0	87.9 ± 1.9
	F1	79.1 ± 1.2	79.3 ± 2.4
subC	AUC	94.4 ± 0.4	88.3 ± 1.8
	F1	87.2 ± 0.4	79.5 ± 2.0

Table 2: Performance of the model trained on *subC* with either *subB* or *subC* used for validation.

Moreover, it is not sure whether it also relies on the information present in the parts of the signal where there is no alarm. In this second series of experiments, we are interested in observing how the model uses the temporal context to produce estimations of the audibility of the alarms.

The model was trained on *subB* since it resulted to better performance in the previous experiment. We varied the duration of the clips used to train the model. Four different durations were experimented: 5.5, 1.0, 0.5, and 0.1 seconds. For this, each input representation was shortened to the desired length around the alarm position. Every time, the model was tested on *subA* using all four clip lengths. The results are reported in Table 3.

As a first observation, when tested with the same clip length as the one used for training, the model shows relatively good performance. This result is true whatever the clip length. However, the model is only able to perform well for all test clip durations when it is trained on 5.5-second clips. This result suggests that the model needs temporal context at training time but not necessarily to make predictions at inference time. Finally, when 5.5-second clips are used for training, the performance of the model weakens slightly as the duration of the test clips is reduced, though it is still quite high. It would therefore be reasonable to train on long clips if the model is then to make predictions over shorter time periods. These observations are based on the AUC. The-F1 score shows some unexplained effects such as a lower value observed when the model is trained on 5.5-second clips and tested on 1-second clips. The investigation of these effects may require a more detailed analysis.

Training \ Test		0.1	0.5	1.0	5.5
0.1	AUC	88.6 ± 1.4	53.2 ± 11.1	51.8 ± 11.9	50.1 ± 7.5
	F1	81.8 ± 1.6	36.0 ± 19.5	32.9 ± 20.6	32.2 ± 21.0
0.5	AUC	53.3 ± 13.2	92.1 ± 1.5	55.5 ± 14.9	50.1 ± 6.7
	F1	44.4 ± 17.7	85.0 ± 1.3	41.9 ± 18.7	33.8 ± 20.2
1.0	AUC	43.9 ± 15.2	53.5 ± 13.5	89.5 ± 1.8	47.7 ± 6.0
	F1	36.7 ± 15.5	45.5 ± 14.7	81.6 ± 1.6	40.6 ± 14.3
5.5	AUC	88.8 ± 6.0	92.3 ± 1.9	93.9 ± 1.6	95.0 ± 1.0
	F1	75.3 ± 6.5	80.0 ± 5.7	75.7 ± 4.2	87.0 ± 1.2

Table 3: Performance on test set depending on the clip length used for training and evaluation.

5.3. The model's output as a psychometric value

As previously mentioned, psychoacoustical experiments are usually conducted in a repeated measures design. For example, in order to evaluate the audibility of an alarm through a *Yes-No* task, a common approach consists in presenting same clip once or several times to every participant. With such procedure, we can measure the proportion of *Yes* responses over all trials. Then by varying a given attribute of the stimulus, it is possible to establish a relationship between this specific attribute and the subjects' responses. Such a relationship is called a psychometric function.

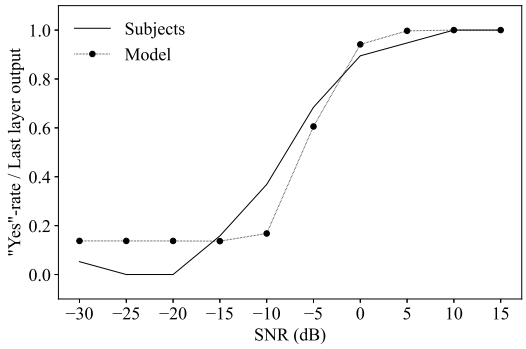

Figure 1: Psychometric function of a clip from *subA*. The rate of positive responses averaged across all participants as a function of the SNR is represented by the plain curve. The dotted curve with round markers shows the values taken by the last neuron.

For instance, consider the procedure described in Section 3.2 to collect data for *subA*. Different clips were generated from the 6 initial clips by varying the SNR and the noise level. By taking a given clip at a single ambient noise level, we can represent the evolution of the proportion of *Yes* responses as the SNR increases. This approach is quite different from what we do when we train a CNN to perform a binary classification task. Indeed, we use binary labels for training. This means that the model is trained to produce outputs as close to 1 as possible when the alarm in the clip is judged to be clearly audible and close to 0 when the alarm is not clearly audible. However, information such as an actual *Yes*-rate is totally absent from the data seen by the model. As a consequence, we do not necessarily expect a match between the output of the model and a psychometric function when varying the SNR of the alarm present in a clip. Yet, we did try to observe the output of the model when the inputs were the same clips of *subA* with different SNRs. The activation of the last neuron was found to roughly follow the evolution of what could be interpreted as a psychometric curve. This result opens up analytical perspectives for future studies. An example is shown in Figure 1.

6. CONCLUSION

In this paper, we proposed a proof of concept of a new approach to assess the audibility of acoustic alarms. We presented an experimental procedure that was specifically designed to collect a dataset with perceptual annotation. This dataset was used to develop a model that gave auspicious results on a binary classification task. Both the influence of the training data and the importance of the temporal context have been investigated. Our results showed that it is possible to predict the audibility of acoustic alarms in relative accordance with human perception, even if training was made on a dataset that was collected using a much lighter procedure than usual psychoacoustical tests. However, we are aware of the lack of a baseline to compare the results of the present work, and therefore plan to collect new perceptual data with different annotators whose "performance" will serve as a basis for comparison with the model. Lastly, the psychoacoustical experiments presented in this article are part of a broader experimental method that includes the numerical rating of the audibility and a detection task that have not been detailed here. The data collected on this occasion will be used in future developments.

7. REFERENCES

[1] I. O. for standardization (ISO), "Ergonomics — danger signals for public and work areas — auditory danger signals," 2008.

[2] J. Żera and A. Nagórski, "Preferred levels of auditory danger signals," *Int. J. Occup. Saf. Ergon.*, vol. 6:sup1, pp. 111–117, 2004.

[3] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin: Springer, 1999.

[4] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Leiden, Netherlands: Brill, 2013.

[5] R. Patterson, "Auditory warning sounds in the work environment," *Phil. Trans. R. Soc. Lond.*, vol. B 327, pp. 485–492, 1990.

[6] M. R. Leek, "Adaptive procedures in psychophysical research," *Perception & Psychophysics*, vol. 63, pp. 1279–1292, 2001.

[7] Y. Zheng, C. Giguère, C. Laroche, C. Sabourin, A. Gagné, and M. Elyea, "A Psychoacoustical Model for Specifying the Level and Spectrum of Acoustic Warning Signals in the Workplace," *Journal of Occupational and Environmental Hygiene*, vol. 4, no. 2, pp. 87–98, Jan. 2007.

[8] B. R. Glasberg and B. C. J. Moore, "Development and Evaluation of a Model for Predicting the Audibility of Time-Varying Sounds in the Presence of Background Sounds," *J. Audio Eng. Soc*, vol. 53, no. 10, pp. 906–918, 2005.

[9] B. RichardWebster, S. E. Anthony, and W. J. Scheirer, "Psyphy: A psychophysics driven evaluation framework for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2280–2286, 2019.

[10] S. Grieggs, B. Shen, G. Rauch, P. Li, J. Ma, D. Chiang, B. Price, and W. Scheirer, "Measuring human perception to improve handwritten document transcription," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[11] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644.e16, 2018.

[12] A. Francl and J. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nat. Hum. Behav.*, vol. 6, pp. 11–133, 2022.

[13] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1744–1748.

[14] N. Turpault and R. Serizel, "Training sound event detection in a heterogenous dataset," in *Proc. DCASE Workshop*, 2020.

[15] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Appl. Sci.*, vol. 10, no. 6, 2020.

[16] E. Çakir and T. Virtanen, "End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018.

[17] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "A survey: Neural network-based deep learning for acoustic event detection," *Circuits Syst. Signal Process.*, vol. 38, pp. 3433–3453, 2019.

[18] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Proc. DCASE Workshop*, 2019.

[19] The Freesound project, https://www.freesound.org.

[20] BigSoundBank by Joseph SARDIN, https://www.bigsoundbank.com.

[21] J. Atyeo and P. M. Sanderson, "Comparison of the identification and ease of use of two alarm sound sets by critical and acute care nurses with little or no music training: a laboratory study," *Anaesthesia*, vol. 70, no. 7, pp. 818–827, 2015.

[22] J.-P. Arz, N. Grimault, and O. ElSawaf, "Experimental assessment of the effect of wearing hearing protectors on the audibility of railway warning signals for normal hearing and hearing impaired listeners," *International Journal of Occupational Safety and Ergonomics*, 2021. [Online]. Available: https://doi.org/10.1080/10803548.2021.1991681

[23] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[24] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[25] N. Turpault, R. Serizel, and E. Vincent, "Analysis of weak labels for sound event tagging," 2021, working paper or preprint. [Online]. Available: https://hal.inria.fr/hal-03203692

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.