Lightweight Annotation and Class Weight Training for Automatic Estimation of Alarm Audibility in Noise

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

LIGHTWEIGHT ANNOTATION AND CLASS WEIGHT TRAINING FOR AUTOMATIC ESTIMATION OF ALARM AUDIBILITY IN NOISE

François Effa[*,†,§], *Romain Serizel*[†], *Jean-Pierre Arz*[*], *Nicolas Grimault*[§],

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

ABSTRACT

In an effort to improve occupational health and safety, we recently proposed an approach to assess the audibility of acoustic danger signals. It is based on the use of a binary classifier trained on perceptual data to predict the audibility of acoustic alarms in audio clips. In the present article, we first investigate the impact of label noise in the training data induced by a flexible annotation procedure on the model performance. We show that a lighter annotation procedure at training still allows for reaching close to human performance at test time. Besides, threshold selection is a crucial aspect in our application as it can have a direct impact on user safety. We thus explore class weight to train a model that allows for a more robust decision threshold selection, ensuring a low false positive rate.

Index Terms— Psychophysics, Machine Learning, Auditory alarms, Acoustic Scene Analysis, Noisy labels

1. INTRODUCTION

In many occupational settings, the safety of workers largely depends on the audibility of the acoustic alarms that are used to warn of potential accidents. The ISO 7731 international standard [1] specifies criteria for auditory danger signals in public and work areas to be "clearly audible". One of these imposes a 15 dB minimum signal-to-noise ratio (SNR) between the alarm and ambient noise levels. However, this criterion is considered as "rather conservative" and may lead to excessive alarm levels, especially in work environments where the ambient noise level is already high [2]. Under such conditions, workers are unnecessarily exposed to very high sound levels which can damage their auditory systems.

From an occupational health and disease prevention perspective, a balance must be struck to ensure that an alarm is audible without posing a risk to workers' hearing. This comes down to determining the level at which the alarm becomes just "clearly audible". In practice, the most reliable approach consists in evaluating the audibility of acoustic alarms through psychoacoustical experiments. Yet, such experiments have two major drawbacks. First, they require time-consuming procedures based on repeated measures design and involving several participants [3]. Second, the audibility of an alarm depends on various acoustic properties other than just the signal and noise levels [4]. For this reason, the experimental approach is stimulus-dependent and new tests have to be conducted whenever the listening condition changes, whether it is a new alarm signal or a different background noise. To overcome this issue, we proposed a data-driven method for predicting the audibility of acoustic alarms [5]. We first collected a dataset consisting of sound clips made with alarm signals mixed with background noises, annotated by several normal-hearing people in perceptual experiments. This dataset was then used to develop a convolutional neural network (CNN) model to perform a binary classification task in which the alarms present in the sound clips were classified as "clearly audible" or "not clearly audible". In the long term, the model could be used to assess the audibility of new auditory alarms without the need for specific perceptual evaluations.

To limit the time cost of data acquisition, an efficient solution is to present every sound clip only once to one of all possible annotators. However, such a procedure stands in contrast with conventional psychoacoustical methods that require every measure to be repeated, and could result in noisy labels. As label noise can be responsible for a decrease in classification performance [6], there is an interest in questioning the strategy adopted to collect the training data for our predictive model. Consequently, a first motivation for the present paper is to investigate the impact of the precision required in the dataset annotation procedure on model performance. On another note, the approach we developed in order to assess the audibility of acoustic alarms should be studied in terms of its applicability to the adjustment of alarm levels. Indeed, the choice of a decision threshold to consider an alarm as "clearly audible" must be oriented by the imperative need to limit false alarms as much as possible while avoiding excessive alarm levels. Therefore, the present work also focuses on proposing an optimal configuration of the model based on these considerations.

2. EXPERIMENTAL SETUP

2.1. Dataset

The dataset contains audio clips made of auditory alarms, lasting between 0.2 and 1.8 s, mixed with 5.5 s background noises from different work-related environments, such as factory or roadwork noises. The alarms and noises were 44.1 kHz mono WAV files mostly downloaded from public sources[1], namely Freesound [7], BigSoundBank [8], and medical alarms from a scientific publication [9]. Each clip is associated with a label, 0 or 1. The label 1 means that the alarm is clearly audible, while the label 0 means that it is not. These labels were obtained by presenting the sound clips to normal-hearing annotators before asking them *"Was the alarm clearly audible?"*. The material used for this stage was detailed in a previous article [5]. Every clip in the dataset has been annotated by several participants: 10 and 11 for development and test sets, respectively. The actual labels were then set to 1 when more than half of the annotations were positive, 0 otherwise.

In order to be able to compare our approach to standard psychoacoustical methods, the test set was collected in well defined and controlled listening conditions. It contains several presentations of the same stimuli with varying SNRs and ambient noise levels. To elaborate this test set, we picked 6 different alarm and background noise pairs. Each pair was used to create 20 versions of a same clip by varying the SNR (from −30 to +15 dB with a step of 5 dB) and using two different ambient noise levels (60 and 80 dBA), resulting in a total of 120 clips.

The development set was purposely prepared with less constrained listening conditions than the test set. It contains 2000 audio clips made by mixing an alarm with a background noise, both randomly chosen among a total of 70 and 52 different alarms and background noises, respectively. The ambient noise levels were 60 and 80 dBA and the SNRs were integers between −30 and +15 dB. For validation, a 20% subset was randomly split from the training data and kept fixed during all the experiments.

2.2. System

The system used in the following is a CNN with 4 convolutional layers, each followed by ReLU activations and frequency max pooling. The convolutional layers have [32, 64, 64, 128] filters and a 3-by-3 kernel size. The max-pooling is [1, 4], [1, 4], [1, 2] and [1, 2], respectively. An L_p aggregation, with $p = 2$, is operated over the time axis on the outputs of the last convolutional layer stacked along frequency axis. The L_p aggregation is followed by a single-neuron classification layer with sigmoid activation.

The model takes mel-spectrograms as inputs and is intended to produce binary estimates of the audibility of the alarms present in the sound clips. The features are extracted from the clips by computing a 1024-sample short-time Fourier transform (STFT) with a Hamming window and a hop size of 512. The spectrograms are then mapped into 64 mel-spaced frequency bins between 20 Hz and 22.05 kHz. For the experiments, the whole development set is used to compute standardization coefficients in order to standardize all the mel-spectrograms to zero mean and unit variance.

The model is trained with a binary cross-entropy loss function and Adam optimizer [10]. In Adam, the learning rate is set to 0.0001 and a weight decay of 0.0001 is employed to reduce overfitting. For the same purpose, we also apply dropout on the outputs of all convolutional layers with a probability of 0.25. Training is made over a maximum of 250 epochs and the best model is retained based on the accuracy on the validation set.

2.3. Metrics

To quantify the classification performance, we use the area under the receiver operating characteristic curve (AUC) and the F1-score. The metrics are computed on the outputs obtained with 10 models trained with random initializations. The scores are reported with average and 95% confidence intervals.

3. ANNOTATION PROCEDURE AND LABEL NOISE

3.1. Problem definition

Annotating an audio dataset manually is a pretty challenging task. In some cases, the annotation process can necessitate several expert annotators and still be prone to annotators biases [11, 12, 13]. When it comes to perceptual annotation, we are interested in obtaining "perceptually relevant" labels at the lowest possible time cost. In our problem, to lower annotation time cost for the training data, we propose to deviate from classical listening tests that require several participants, each one presented a number of times with the same stimuli [1]. Instead, each stimulus in the dataset could be annotated by only one person, randomly chosen among the participants sample group. However, by proceeding in this way, we do not have any information regarding the intra- and inter-individual variability in the answers for each example present in the dataset. This can lead to mislabelled data. Indeed, a perceptual evaluation provided by one person after a single presentation is not necessarily representative of the average of multiple evaluations of the same stimulus by a group of people with the same hearing status. This section aims to answer the question of the impact of these noisy labels in the training data on the learning and performance of the model. Since it is intended to produce predictions close to human judgements, we also compare the model to an average "human performance".

[1] Only one alarm was self-recorded.

3.2. "Human performance" baseline

As detailed in Section 2.1, the test set was annotated by 11 normal-hearing participants. Consequently, for each of the 120 clips in the test set, we have 11 individual binary labels. This allows us to define an average "human performance" by computing the evaluation metrics on the test set for each participant (using the other 10 participants to obtain the reference labels). We obtain an AUC of 91.7 ± 2.2 and a F1-score of 91.0 ± 2.5. It should however be noted that this performance may be slightly overestimated since it is obtained from the same data that were averaged to form the test set.

3.3. Experiments

Since the development set was annotated by 10 participants, we conduct two experiments to investigate the effect of varying the number of annotators on the model performance.

First experiment: We define N, the number of annotators required to annotate each example. For each clip in the development set, we randomly select N in the 10 possible annotators. The label of the clip is then set to 1 if more than half of the N annotators reported the alarm in the clip as clearly audible, and to 0 otherwise. This random draw is performed before each run and kept fixed for the whole training. This experiment is repeated for different values of N. As reported in Figure 1, there is no significant improvement in performance on the test set brought by increasing the number of annotators of the training data from 1 to 10. This result differs from what is observed on the development set. Indeed, performance appears to get better on the development data as we increase the value of N. This can be interpreted as evidence that the label noise induced by reducing the number of annotators acts as a form of regularization and does not actually deteriorate model performance. As a result, using a lighter annotation procedure than standard psychoacoustical tests is a viable alternative since it saves time without causing a drop in classification performance.

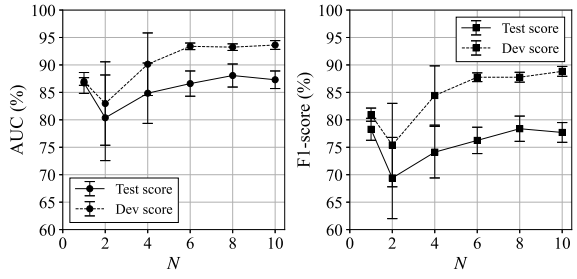

Fig. 1: Performance on development and test sets for the first experiment as a function of N.

Second experiment: We proceed similarly as in the first experiment, except that the random draw is performed at every epoch. The aim of this experiment is to determine whether labels from multiple annotators can be used as data augmentation. The results are presented in Figure 2. For $1 < N < 10$, there is a slight increase in the average test performance compared to the first experiment, but it is not significant. For $N = 1$, the average performance is equivalent to the first experiment but the variability is much higher. These results support the idea that annotating the training data using a single annotator per clip should be preferred over other procedures.

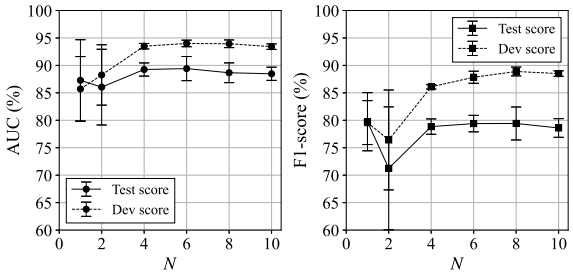

Fig. 2: Performance on development and test sets for the second experiment as a function of N.

Finally, we can compare the model to the "human performance" baseline described in Section 3.2. For now, the human is still outperforming the model, although the model's performance is quite strong. Further studies will be made to investigate the possibility of increasing model performance by training on larger datasets, or by trying different architectures. As for the F1-score in particular, improvements can already be achieved by optimal thresholding [14]. However, the choice of a threshold must follow rules that are determined by the conditions of application of the system. These aspects are discussed in the next section.

4. SELECTION OF A DECISION THRESHOLD

4.1. Problem definition

The target of our system is to predict at what level an alarm becomes "clearly audible" in a given sound environment. For a same clip presented at different alarm-to-ambient noise level ratios, the binary classifier is expected to produce an output that is close to 1 when the SNR is high enough for the alarm present in the clip to be "clearly audible", and close to 0 otherwise. This way, we can identify a minimal SNR value for which the alarm becomes "clearly audible". In practice, human perception does not have such hard thresholds. The alarm level or SNR is rather to be mapped to a probability of a given response (here, "clearly audible" or "not clearly audible"). The mapping function, known as *psychometric*, is never a step function [15]. In a previous article [5], we actually showed that the activation of the model's last neuron also

increases progressively with SNR, just like the probability of receiving a "clearly audible" response. With this in mind, we must now question the discrimination threshold employed in the model to consider an alarm as "clearly audible".

When it comes to using the model to make decisions regarding alarm levels, care must be taken in selecting the discrimination threshold. Indeed, for obvious safety reasons, we do not want any alarm to be missed. From the perspective of the model, it comes down to the necessity of minimizing the false positive rate (FPR). That being said, in the interest of preserving workers' health, we also want to prevent the model from recommending excessively high alarm levels. To satisfy these two requirements, we suggest setting the discrimination threshold of the model in such a way as to aim at a region of the psychometric curve that ensures good audibility of the alarm with a reasonable sound level. This section proposes class weight training before selecting a decision threshold as a solution to better meet the specifications of the model.

4.2. Method

As we aim to make sure that the alarms are clearly audible, we must target a region of the psychometric function where the probability of receiving a "clearly audible" response is high. We therefore prefer a $80-95\%$ probability to a 50% probability which is simply a majority vote. Consequently, to configure the system, we should aim for a good match between the shapes of the psychometric curve and the activation of the last layer of the model, particularly in the high probability regions. To this end, we propose to reformulate the loss function in order to give more importance to the positive labels. This is done by setting a weight to the positive class in the binary cross-entropy loss function [16]:

$$J_{wBCE} = -\frac{1}{M}\sum_{m=1}^{M}[w y_m \log(h_\theta(x_m)) + (1-y_m)\log(1-h_\theta(x_m))] \quad (1)$$

where M is the number of training examples, w the weight, h_θ the model with weights θ, and (x_m, y_m) the input feature and target label for training example m, respectively.

After training the model using the loss function formulated in Equation 1, we can select the discrimination threshold that minimizes the false positive rate.

4.3. Experimental results

This subsection presents an experiment whose purpose is not to select parameter values but to show trends. In order to configure the model, the following observations should be made at development stage on a validation set. However, since our dataset is still relatively small, we settle for a preliminary experiment showing interesting effects on the test set.

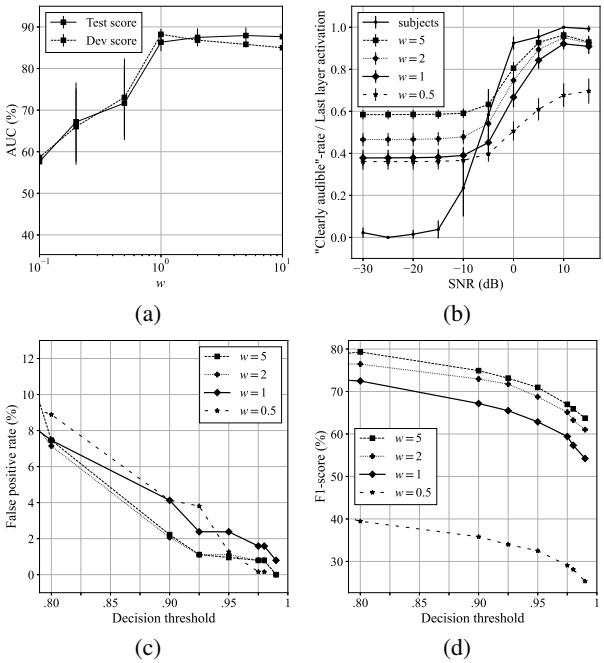

Fig. 3: Experimental results. (a) AUC score as a function of w. (b) Psychometric function and last layer activation as a function of SNR for different values of w. (c) FPR on test set versus decision threshold. (d) F1-score on test set versus decision threshold.

We train the model using different values of w. Figures 3a and 3b show that when $w > 1$, the AUC is maintained and the model output (before thresholding) reaches higher values for high SNRs. As a result, the distance between the model output and the subjects' psychometric function is reduced in this area. This allows for a better performance compared to the case $w = 1$ (unweighted loss function) when setting a high decision threshold. As depicted in Figures 3c and 3d, a same threshold gives lower FPR and higher F1-score for $w > 1$.

5. CONCLUSION

In this paper, we have showed that reducing the number of annotators for a single clip does not decrease the performance of the model. This aspect can help drastically reducing the annotation cost while maintaining performance at test time that are aligned with those obtained with standard psychoacoustical tests. We also explored a method to improve the robustness of the decision threshold selection in order to use the model to specify the level of acoustic alarms. Preliminary results suggest that a right formulation of the loss function during training can help improving model performance and reliability by lowering the false positive rate for a given decision threshold. However, for now, our model is still outperformed by the human baseline. Future developments will come to improve model performance.

6. REFERENCES

[1] ISO 7731 — International Organization 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] A. Taghipour, B. C. J. Moore, and B. Edler, "Masked threshold for noise bands masked by narrower bands of noise: Effects of masker bandwidth and center frequency," *J. Acoust. Soc. Am.*, vol. 139, pp. 2403–2406, 2016.

[4] L. Schell-Majoor, J. Rennies, S. D. Ewert, and B. Kollmeier, "Application of psychophysical models for audibility prediction of technical signals in real-world background noise," *Applied Acoustics*, vol. 88, pp. 44–51, 2015.

[5] F. Effa, R. Serizel, J.-P. Arz, and N. Grimault, "Convolutional neural network for audibility assessment of acoustic alarms," in *Proc. DCASE Workshop*, 2022.

[6] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.

[7] "The freesound project," https://www.freesound.org.

[8] "Bigsoundbank," https://www.bigsoundbank.com.

[9] 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.

[10] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

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

[12] 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.

[13] V. Morfi, Y. Bas, H. Pamuła, H. Glotin, and D. Stowell, "NIPS4Bplus: a richly annotated birdsong audio dataset," *PeerJ Computer Science*, vol. 5, pp. e223, Oct. 2019.

[14] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize f1 measure," in *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, 2014, pp. 225–239, Springer Berlin Heidelberg.

[15] F. A. Wichmann and F. Jäkel, *Steven's handbook of experimental psychology: Vol. 5. Methodology*, chapter Methods in Psychophysics. In Wixted, J. T. & Wagenmakers, E.-J. (Eds.), Wiley, 2018.

[16] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.