forum acusticum 2026
8-12 September · Graz, Austria

PSYCHOACOUSTIC-BASED ACOUSTIC ANALYSIS RECOMMENDATION BASED ON THE SYMBIOSIS OF DOMAIN KNOWLEDGE AND AI MODELS.

Thiago Lobato[1], Andreas Herweg[1], Tim Kamper-Schley[1]

[1] *HEAD acoustics GmbH*

Abstract

Loudness, sharpness, tonality, roughness, fluctuation strength, relative approach, Fast Fourier Transform, modulation spectrum (and many others ...) The number of acoustic analyses can be daunting for inexperienced users. One consequence of this is that even when there may be a perfect match between an industrial problem and an acoustic solution, this link can be obfuscated by the lack of user expertise. We propose a new solution paradigm for this problem: we train an AI model able to perceive the prominent psychoacoustic characteristics of sounds and combine it with domain knowledge for improved analysis parameterization. This allows us to tailor analysis recommendations based on the symbiosis of how the sound is perceived by people together with expert domain knowledge. This is achieved by a psychoacoustic-based network pre-training coupled with a fine-tuning step using expert-generated listening test results on strongly diverse industrial sounds. The results of the model are then piped into a rule-based model extracted from expert knowledge. We show that the model can achieve good results in predicting expert ratings and can produce helpful recommendations. This is a first step in the direction of fully automated acoustic analysis recommendations, which we believe can empower a whole class of non-expert users.

Keywords: *Psychoacoustic, AI, recommendation system, hybrid method, sound classification*

1. Introduction

The importance of sound for the design of many industrial products continually increases. Interior vehicle noise, household appliances, information technology equipment, and consumer devices are all judged not only by their performance, but also by how they sound. The question "how it sounds", however, is far from trivial and often the realm of psychoacoustics [1], [2], [3], [4], [5]. Additionally, even outside perception, acoustic signals can be used to identify anomalies [6], [7] or classify sounds [8] with appropriate features, for example, the relative approach [9], [10]. In most cases, one can expect to almost always find an ideal acoustic analysis (or combination thereof) that is able to sufficiently analyze a given industrial problem. Not surprisingly, since the number of possible acoustic analyses and their parameters is very diverse [11]. While this rich toolbox enables very detailed assessments, it also creates a non-trivial decision problem in daily practice, namely which analysis to solve and how to parameterize it. Usually, expert users develop good intuition over years of project work, but if we desire to properly democratize acoustic analysis even for non-experts, this is not the solution.

On the positive side, machine learning is increasingly used as a flexible glue in acoustics between physical descriptors and perceptual judgments. The AI-SQ Metrics framework proposed by Sottek and Lobato[5] uses machine learning models to learn non-linear combinations of psychoacoustic and spectro-temporal parameters that better match listening test results than classical linear regression. Those relations can even be made interpretable by leveraging KANs [12]. Large foundational network models are also able to learn good general audio features [13], [14] for more complex problems, although with the disadvantage that they are not interpretable. However, we can still be confident that such models are able to learn reasonable perceptual characteristics of sounds.

In this work, we explore a psychoacoustic-based acoustic analysis recommendation framework that combines interpretable domain knowledge with black-box foundational models. For this, we fine-tuned a pre-trained neural network to encode the main psychoacoustic characteristics of sounds based on expert listening-

***Corresponding author**: andreas.herweg@head-acoustics.com.*

test data for a broad set of industrial sounds, so that the network can approximate expert judgments in a perceptually meaningful embedding space. This classification of "what prominent psychoacoustic properties are in the sound" can be exploited with domain knowledge (here, a rule-based approach) to propose ideal analyses from a set of known useful analyses. In this way, we can propose recommendations with "black-box accuracy" while retaining complete interpretability of results. We demonstrate that the resulting system can reproduce expert ratings with good accuracy while providing practically useful analysis suggestions for any user.

2. General Approach

Our main idea is divided into a couple of steps. First, we want to learn a classifier that, based on an audio signal, can tell us which psychoacoustic phenomena are present in the sound. We may also want to get a "general idea" of **where** in the frequency domain a phenomenon occurs (i.e. which frequency). This is useful for some analyses that are band-dependent. For those, we use neural surrogates such as those presented in [15]. Based on those properties, we select the appropriate analyses that would be able to properly evaluate them by leveraging expert know-how. A scheme of this process can be seen in Fig. 1.

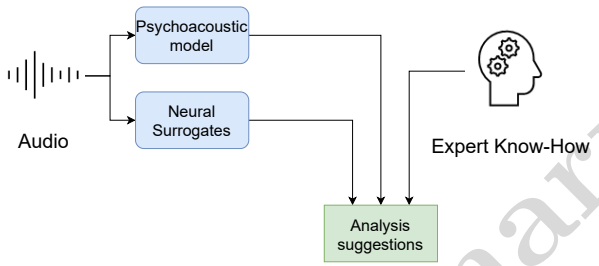

Figure 1: General description of our recommendation system.

3. Methodology

3.1. Psychoacoustic model

3.1.1. Data

We trained a model to predict whether a given parameter is relevant (to be analyzed) for a given sound. In theory, we could take any dataset, calculate psychoacoustic parameters on it, and then define a threshold for relevance for each one. The problem with this approach, however, is that it completely ignores the contextual information of the sounds. For example, a value of 0.5 Tu_{HMS} may be deemed too high/tonal for a toothbrush, while being too low for a flute sound. Thus, contextual information is important and requires expert judgment, which we use as training labels for our model. This also brings the point that we need a dataset for the context we want to evaluate. In this work, we focused on machine sounds, such as cars, hair trimmers, coffee machines, vacuum cleaners, and so on -mainly machines for which manufacturers may want to optimize sound quality. In total, we prepared an internal dataset of 400 highly curated real sounds. These were evaluated by seven expert acoustic engineers with binary questions for each analysis (relevant / not relevant) in a listening test (with calibrated reproduction systems from HEAD acoustics). The listening test was divided into four sections, each containing 100 sounds. In addition to the real data, we also trained the model with synthetically generated sounds that represent standard sounds for each analysis, such as modulated sine waves. This additional dataset serves to make sure the model can respond well also to synthetic (noiseless) data, which may be the case in applications such as sound design. The number of those sounds was 251, and they were used only for model training, not for the listening test. This number was selected as the maximum such that the accuracy on the real sounds did not decay by more than 1%.

3.1.2. Model

We fine-tuned a pre-trained model, namely the EfficientAT mn10_as [16], pre-trained on AudioSet [17]. The training was done in PyTorch as a multi-label problem using the AdamW optimizer. This model is relatively small and thus friendly to edge deployment, with both a low memory footprint and low latency. To investigate the accuracy of the model, we performed a 4-fold cross-validation (each fold corresponding to one listening experiment) and evaluated the accuracy of the out-of-fold predictions. As a baseline for the model, we compared it to an "ideal" psychoacoustic threshold for each analysis, fitted with a logistic regression based on the listening test data. Even though this disregards contextual information, we will see that it is still a very strong baseline. Additionally, we compared the model results to the expert evaluations by investigating how well one expert can predict the average recommendation of the others.

3.2. Neural Surrogates

We generated three neural surrogates, namely for tonality, roughness, and fluctuation strength, all based on the Sottek Hearing Model described in ECMA [3]. All were generated using the same process as in [15], and indeed, the tonality model is the same. For roughness and fluctuation strength, however, a larger block size is needed to obtain the fine-grained results necessary for the time-frequency analysis. We simply used a longer block size for those in a way that matches the ECMA standard [3]. We also modified the architecture by replacing the GRU layer with a causal transformer with multi-query attention. A detailed description of the surrogates is planned for a possible future work. The actual deployment of the model was done using ONNX [18].

3.3. Analyses

For the first version of our model, we focused on the following analyses:

- Main psychoacoustic analyses: loudness, tonal-

ity, roughness, sharpness, fluctuation strength, impulsiveness.

- Relative Approach.
- STFT (Short-Time Fourier Transform).
- Modulation spectrum.

Some of these analyses have parameters that we selected as a function of the previously calculated metrics. For example, let us say our neural models have shown that roughness is relevant for an audio signal, and the main frequency band of it is around 1 kHz. We may want to calculate its exact modulation frequency, from which we can calculate the modulation spectrum centered at 1 kHz and with parameters that emphasize roughness-relevant frequencies, namely approximately from 20 Hz to 100 Hz. As another example, we may get the information that we have impulsiveness in the signal but no tonality, so we can parameterize the Relative Approach analysis to focus on time-patterns. All these evaluations require domain knowledge that can be very elegantly combined with the neural surrogates, which have basically no perceptual latency.

4. Results

4.1. Psychoacoustic model

Results of the model compared with psychoacoustics and the average expert can be seen in Figures 2 and 3. We can verify that our psychoacoustic model has high accuracy for most analyses, which is above that of the (context-less) psychoacoustics, with the exception of impulsiveness. An improvement can be seen mainly for the fluctuation analysis, indicating that it contains a very strong context dependency. We also see that when we compare the results to the expert listeners from the listening test, the psychoacoustic model predictions are pretty good, being better than most of the individual expert listeners. These results indicate that our model works really well, and even though it is not perfect, it has equivalent performance both to experts as well as (contextual-less) psychoacoustics, while having a latency of less than 20 ms/s on an old Xeon W-2145 CPU. This latency is barely perceptible in regular use, even when a project with many files should be analyzed.

4.2. Recommendations

The recommendations are based fully on domain knowledge and thus are a ”white box” that can be modified as needed. However, to show the effect a good parameterization can have, we prepared two results for the following analyses: Relative Approach and Modulation Spectrum, both of which are notoriously challenging to parameterize. The comparison to the standard parameters for the modulation spectrum can be seen in Figures 4 and 5 for a fluctuation sound with a low modulation frequency. We see that the default analysis goes up to 750 Hz, which is just too high for this type of fluctuation, and thus it becomes very difficult to resolve the modulation. On the other hand,

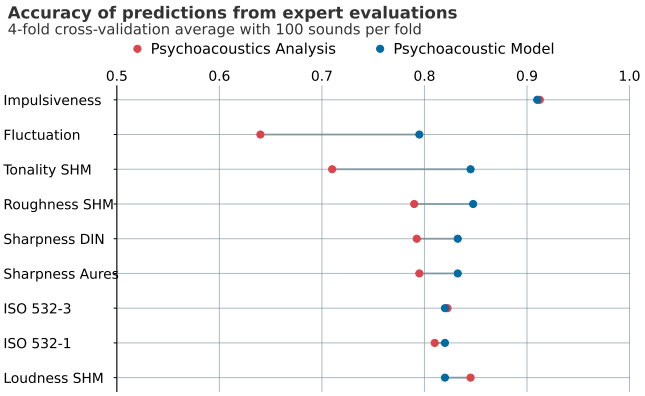

Figure 2: Comparison of the trained model with the ideal threshold of psychoacoustic analysis on out-of-fold data.

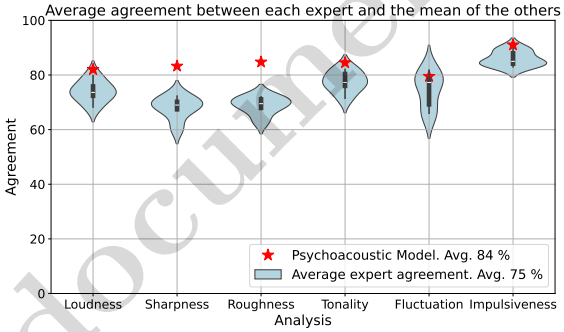

Figure 3: Comparison of the psychoacoustic model to the agreement of the individual listening-test experts with the average of the remaining experts.

the AI recommendation, based on our psychoacoustic model and neural surrogates, is able to pinpoint that we have a fluctuation and also in which band this fluctuation is prominent. Thus, we are able to see results with a considerably better resolution.

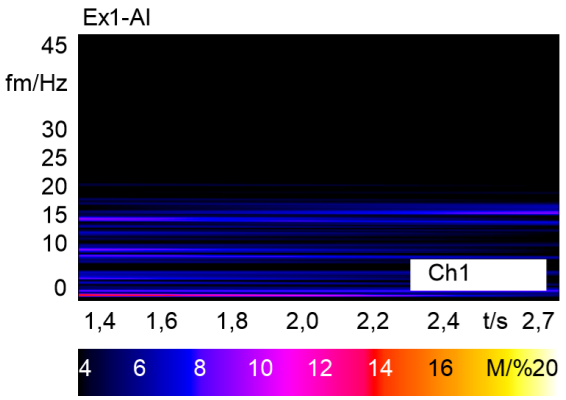

Figure 4: Example of the modulation spectrum of a fluctuating signal using parameters suggested by our framework (reduced time-axis due to analysis parameters).

In Figures 6 and 7, we see an example of using the Relative Approach on an impulsive sound. We see

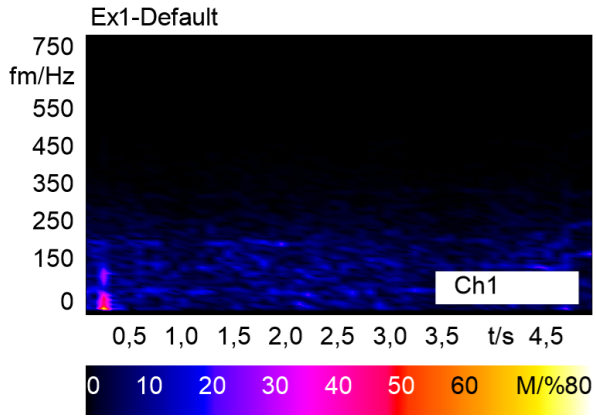

Figure 5: Example of the modulation spectrum of a fluctuating signal using default parameters.

that the suggested parameters allow a considerably less noisy result, especially at lower frequencies.

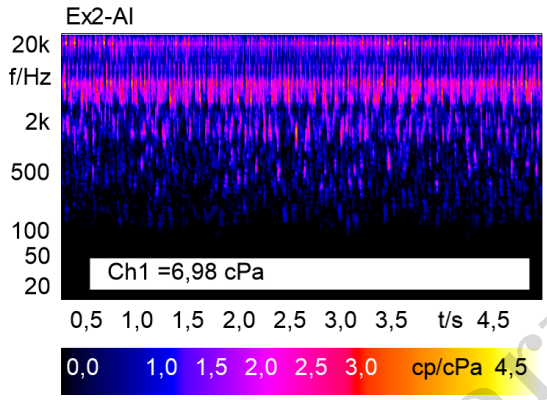

Figure 6: Example of the relative approach of an impulsive signal using parameters suggested by our framework.

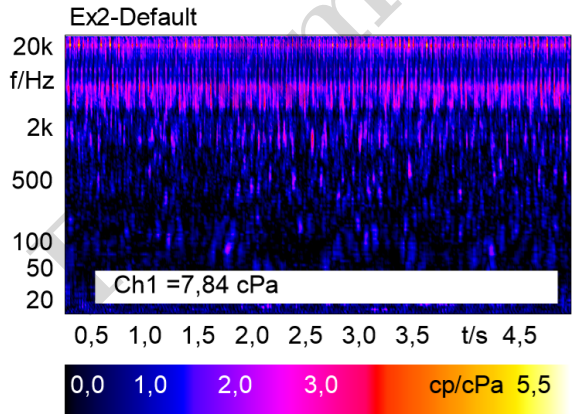

Figure 7: Example of the relative approach of an impulsive signal using default parameters.

These are improvements that users of such tools get "for free" in the sense that no expensive calculation is necessary.

5. Conclusion

We proposed a framework that combines neural network models and domain expertise to provide analysis recommendations in acoustic analysis. This was done by first classifying the prominent psychoacoustic characteristics of a sound, then finding the location of the most prominent events through neural surrogates, and, at last, combining this information with domain knowledge. Our framework was able to reproduce estimates from experts with very high accuracy (about 84%) while being equivalent to calculating individual psychoacoustic metrics; however, it does so with a minimal latency of only a couple of milliseconds. It was also shown how some results can be improved by ideal parameterization, and that our method is able to achieve this by leveraging domain knowledge. We note that our method still has some limitations in the sense that we may still expand it to more analyses and that our rules are focused mainly on stationary sounds. Still, we believe this work to be a fundamental stepping stone toward the goal of democratizing acoustic analysis for non-experts.

References

[1] H. Fastl, "Psychoacoustic basis of sound quality evaluation and sound engineering," *13th International Congress on Sound and Vibration 2006, ICSV 2006*, vol. 1, pp. 59–74, Jan. 2006.

[2] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Berlin, Heidelberg: Springer, 2007. DOI: 10.1007/978-3-540-68888-4

[3] Ecma International, "ECMA-418-2: Psychoacoustic metrics for ITT equipment – part 2: Models based on human perception," Ecma International, Geneva, Switzerland, Tech. Rep., 2025, 4th ed., June 2025.

[4] R. Sottek, T. Lobato, W. R. Bray, and A. Fiebig, "Contextual event-based sound quality metrics," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Chiba, Japan: Institute of Noise Control Engineering, 2023. DOI: 10.3397/in_2023_0222

[5] R. Sottek and T. H. G. Lobato, "AI-SQ Metrics: Artificial intelligence in sound quality metrics," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Seoul, Korea: Institute of Noise Control Engineering, 2020, pp. 3083–3091.

[6] H.-H. Wu, W.-C. Lin, A. Kumar, L. Bondi, S. Ghaffarzadegan, and J. P. Bello, "Towards Few-Shot Training-Free Anomaly Sound Detection," in *Interspeech 2025*, 2025, pp. 3384–3388. DOI: 10.21437/Interspeech.2025-467

[7] K. Dohi et al., "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and*

Events Workshop (DCASE), arXiv:2305.07828, Tampere, Finland, 2023, pp. 31–35.

[8] G. Mauer and W. Krebber, "Brake squeal and moan detection: Important requirements concerning mobile recording systems," in *Automobile Comfort Conference*, 2006.

[9] W. R. Bray, "Using the "relative approach" for direct measurement of patterns in noise situations," *Sound and Vibration*, pp. 12–16, Sep. 2004.

[10] K. Genuit, "Objective evaluation of acoustic quality based on a relative approach," in *Proceedings of Internoise '96*, Liverpool, UK, 1996.

[11] HEAD acoustics GmbH, *ArtemiS SUITE: Modular software platform for sound and vibration analysis*, Version 17, software for sound and vibration as well as psychoacoustic analysis, HEAD acoustics GmbH, Herzogenrath, Germany, 2025. [Online]. Available: `https://www.head-acoustics.com/products/analysis-software/artemis-suite`

[12] T. Lobato, R. Sottek, and M. Vorländer, "Estimation of interpretable non-linear sound quality metrics using kolmogorov-arnold networks (kans)," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, Malaga, Spain: European Acoustics Association, Jan. 2024.

[13] B. Elizalde, S. Deshmukh, and H. Wang, *Natural language supervision for general-purpose audio representations*, 2023. arXiv: `2309.05767 [cs.SD]`. [Online]. Available: `https://arxiv.org/abs/2309.05767`

[14] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575. DOI: `10.21437/Interspeech.2021-698`

[15] A. H. Thiago Lobato Julian Becker, "Fast calculation of the psychoacoustic metric tonality using a surrogate neural network model," in *DAGA 2024 Hannover*, 2024, pp. 1243–1246.

[16] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: `10.1109/ICASSP49357.2023.10096110`

[17] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, IEEE, 2017, pp. 776–780. DOI: `10.1109/ICASSP.2017.7952261` [Online]. Available: `https://doi.org/10.1109/ICASSP.2017.7952261`

[18] O. R. developers, *Onnx runtime*, `https://onnxruntime.ai/`, Version: x.y.z, 2021.