DOI: 10.61782/fa.2025.0352

FORUM ACUSTICUM EURONOISE 2025

ESTIMATION OF INTERPRETABLE NON-LINEAR SOUND QUALITY METRICS USING KOLMOGOROV-ARNOLD NETWORKS (KANS)

Thiago Lobato[1*] **Fabian Kamp**[1]

[1] HEAD acoustics GmbH, Germany

ABSTRACT

Sound Quality Metrics are widely used to evaluate the sounds of machines, environments, and essentially anything audible to people. Typically, combinations of (psycho-) acoustic parameters are used to establish such metrics by mapping the perceptive evaluation of sound and quantifying the perceived sound quality. These metrics can replace costly jury testing activities, which need to be conducted once to determine the metrics themselves. To employ them in product optimization tasks, engineers and designers must understand which properties of a sound must be modified to improve the metric results and hence, the perceived sound quality of the product. Thus, the metrics need to be interpretable, which may limit their accuracy and expressive power. Recently, Kolmogorov–Arnold Networks (KANs) were introduced, which can estimate analytical expressions of complex learned relationships, thereby providing an effective way to learn non-linear metrics with a differential formulation. This stands in contrast to classical symbolic regression methods, which must be learned in a discrete manner. This work investigates how well KANs can be used to learn non-linear relationships for Sound Quality Metrics, and compares the results both to fully interpretable linear equations and established black-box machine learning methods, such as Support Vector Machines and Gaussian Processes.

***Corresponding author**: *thiago.lobato@head-acoustics.com.*

Keywords: *Sound quality metrics, Psychoacoustics, non-linear models, machine learning, KANs*

1. INTRODUCTION

Sound quality metrics are essential tools for evaluating and optimizing auditory experiences across various domains, from automotive [1] to consumer electronics [2]. The standard approach is to estimate linear relationships based on psychoacoustic parameters [1, 3, 4]. While there are approaches that explore the use of black-box models [4, 5], the interpretability of the results remains a crucial factor when deciding which model to use. Recent advancements in machine learning, particularly the advent of Kolmogorov–Arnold Networks (KANs) [6], offer opportunities to examine complex non-linear dependencies while retaining interpretability.

This work begins by defining sound quality metrics in more detail, then provides the background on KANs and explains why they can maintain interpretability. Afterwards, we discuss the methodology used to evaluate the methods and present the results. One of our main contributions is to show that using linear models based on relevant analyses generates interpretable results comparable to (or occasionally even better than) some black-box models. Another is to demonstrate that KANs can provide a modest boost in quality without sacrificing interpretability, especially when a sufficient number of samples is available.

2. SOUND QUALITY METRICS

Models of sound quality metrics can be generally represented as follows:

$$f(\mathbf{X}) = \mathbf{y} \quad (1)$$

In this formulation, $f(\cdot)$ is a transformation, $\mathbf{X}$ are the predictors (often psychoacoustic metrics) of a set of sounds, and $\mathbf{y}$ represents aggregate statistics of listening test results, such as the average or median rating. Both annoyance and sound quality can be surveyed in such a listening test to obtain $\mathbf{y}$. Although $f(\cdot)$ can theoretically be any function, interpretability is highly advantageous: it allows users to directly identify which predictors are responsible for poor evaluations, thus enabling targeted and efficient troubleshooting.

The simplest approach to maintaining interpretability is to apply a linear model on the predictors, so that:

$$\mathbf{Xa} + \mathbf{b} = \mathbf{y}. \quad (2)$$

The coefficients in the vector $\mathbf{a}$ can be readily interpreted, and if a standardization is performed on $\mathbf{X}$ beforehand, they can be reasonably compared to each other. The main advantage of such a method is that a user can easily identify why a sound was evaluated badly, and therefore how it should modify it to achieve a desired metric target. For some cases, however, this linear relation may be too simple to capture the full complexity of the relationship, so using general black-box models may be advantageous [4]. Nevertheless, interpretable non-linear models do exist, and one modern alternative to obtain them is the KAN architecture [6], which is introduced in the following section.

3. KOLMOGOROV-ARNOLD NETWORKS (KANS)

KANs are a modern type of neural network architecture. Unlike conventional multi-layer perceptrons, they learn the activation function through spline interpolations while retaining the same node weightings (a sum operator) [6]. KANs can be described by the equation:

$$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right), \quad (3)$$

where x_p represents the predictors and Φ_q and $\phi_{q,p}$ are univariate activation functions, parameterized as splines. If one replaces those spline with a symbolic formula, one can derive a complete analytical equivalent of the KAN network, making it interpretable. Even if no exact symbolic match is found, a "best-analytical-(greedy)-approximation" can often provide sufficient accuracy while retaining a certain level of interpretability.

This process allows one to **find complex analytical non-linear relations without pre-knowledge of their form**. What this means is that if, let's say, there is a quadratic relation between one of the variables and the response, a KAN model may be able to find this quadratic relation by itself, so that the user does not need to actively identify those relations.

The complexity of the analytical formula in KANs is bounded by the number of nodes and layers. If this value is excessively large, then even an analytical formula may become intractable, diminishing interpretability. As an example, suppose an architecture with layer sizes $\{2, 1\}$ is defined. KANs could generate a simple non-linear function involving only two inputs such as:

$$y = tanh(x_1^2 + cos(x_2)) \quad (4)$$

whereas an architecture with $\{2, 128, 64, 32, 16, 1\}$ layers can become virtually as opaque as a black-box model. In this study, we restrict the network to a maximum of three layers and five nodes per layer, allowing for reasonably complex yet still interpretable non-linear behavior. We emphasize again that those functional relations are found by the model, so the only prior a user should give is the atomic elements allowed to fit the activation functions (ex: x^2 , $\sqrt{x}$, $cos(x)$, etc.).

4. METHODOLOGY

To evaluate the effectiveness of KANs for sound quality tasks and compare them to established models, we curated a benchmark from internal listening tests on sound quality ratings from diverse technical sounds. These included electric vehicles, hair trimmers, refrigerators, coffee machines, motors, and other technical devices. Overall, 19 different listening tests are available, each containing between 15 and 85 samples, providing a representative set of practical sound quality applications.

All tests used the SAE subjective rating scale for noise evaluation [7] which ranges from 1 (lowest sound quality) to 10 (highest sound quality). For the error evaluation, we considered the mean absolute error (MAE) between the predicted score and the participants' average rating for each sound. The predictors were selected based on expert assessments of causal relationships between sound quality and specific sound types and were confined to the analyses implemented in the software ArtemiS from HEAD acoustics [8]. Most predictors were standardized aggregations of psychoacoustic analyses, such as those defined in the

ECMA 418-2 standard [9], for example Sottek tonality. We also used special analyses within ArtemiS, such as the Relative Approach [10], and classical acoustic standards like ISO 532-1 [11], ISO 532-3 [12], and DIN 45692 [13].

For each test, we evaluated the following models: a fully interpretable linear regression, classical black-box models (Support Vector Regression (SVR), Gaussian Process, and boosted trees), and two variants of KANs-one without symbolic constraints and another with symbolic constraints (making it interpretable). To obtain unbiased error estimations, we used a nested cross-validation approach, illustrated in Figure 1. For N folds, each iteration employed one fold as the test set, and the remaining $N-1$ folds as the training set. Within each training set, an internal cross-validation procedure was used to optimize the hyperparameters, and then the final model was retrained on the entire training set. This process yielded an unbiased performance estimate. The KAN implementation was obtained from the Pykan repository [6], boosted trees were implemented via XGBoost [14], and the other black-box models were implemented with Scikit-learn [15]. The linear regression variant used was Elastic Net, also from Scikit-learn, which optimizes both L1 and L2 penalty losses. The hyperparameters for each model in each internal CV step were optimized with the Optuna library [16] using a budget of 300 iterations.

Additionally, we considered two extra reference scenarios: a *"best"* case, selecting the best model for each test (regardless of interpretability), and a *"best interp."* case, selecting the best interpretable model (either linear and KAN) for each test.

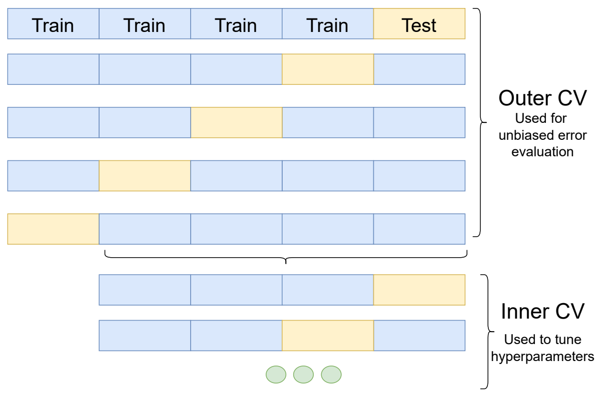

Figure 1. Outer cross-validation (CV) scheme used to evaluate each model and to choose the hyperparameters for each test fold.

5. RESULTS

5.1 Model evaluation

The results of the nested cross-validation evaluation for all models at the 19 listening test results are shown in Figure 2. The error is averaged for all sounds in each listening test.

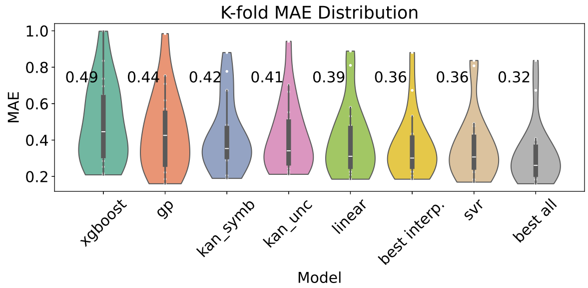

Figure 2. Outer CV MAE of each model over all listening tests.

All models achieve very low errors, with the best average performance coming from SVR, which is a black-box model. However, the difference between SVR and the linear model is small in absolute terms, suggesting that linear models already perform quite well. The KANs exhibit slightly higher errors than the linear models but still remain in a low range. This initially indicates that there may be no pressing need to resort to black-box models in typical sound-quality metric problems, particularly when the context is well understood.

However, the *"Best"* model shows a considerably lower error than the SVR, and the *"best interp."* approach outperforms the standalone linear model, suggesting that combining multiple models can provide a boost in overall performance, which is not particular a surprise. Interesting here is that now KANs represent a non-linear interpretable model.

Figure 3 illustrates the improvement (or lack thereof) of each method relative to the linear metric for individual tests (possible since we used the same folds for all models). While average differences may be small, some individual tests see an improvement of up to 0.2 in MAE for non-linear approaches, indicating that there are scenarios where more complex models can provide a clear benefit.

Figure 4 shows the "victory frequency" for each model, indicating how often a particular model outperforms another. For the black-box models, although SVR

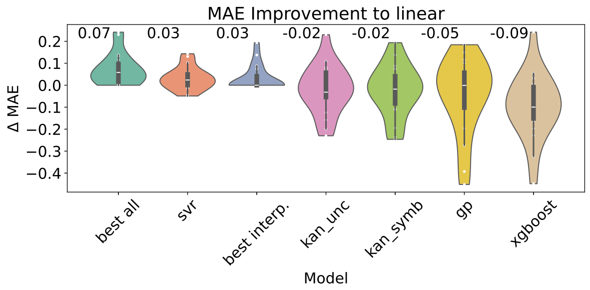

Figure 3. MAE improvement compared to the linear metric across individual test cases.

may produce the best results on average, there are many cases where Gaussian Process performs better, suggesting it can be a viable fallback if SVR underperforms (in contrast to linear). Comparing our two interpretable solutions (linear and KANs), KANs are better about 42% of the time, which aligns with the modestly superior results of *"best interp."*.

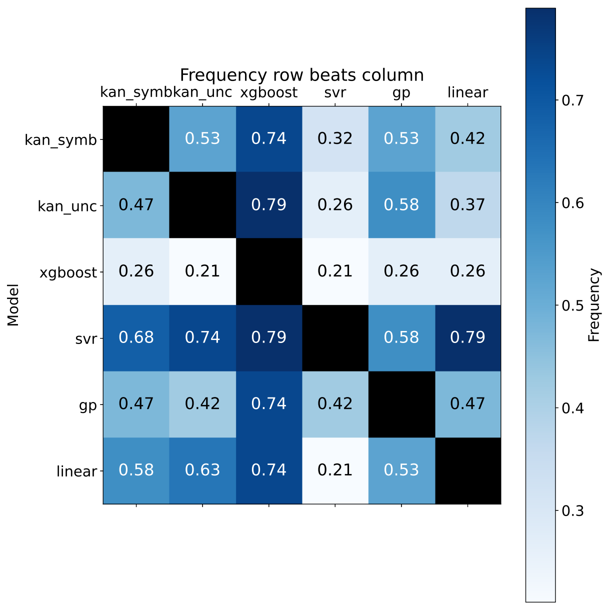

Figure 4. Frequency in which a given model outperforms another, computed over all tests.

It would be helpful to have a guide to when KANs may be better than linear models, and in fact, the better interpretable model correlates well with the number of samples available in the listening test. In Figure 5 we make a comparison **only of interpretable models** (linear or KAN) as a function of the number of samples in the listening tests. Linear models tend to prevail for smaller sample sizes (fewer than about 40 samples), whereas KANs often become more effective as the sample size increases. By contrast, when considering all models (including black-boxes), Figure 6 shows that black-box methods can outperform linear even in small-sample scenarios. Indeed, while linear models reliably produce small errors, they are hardly *the best* model.

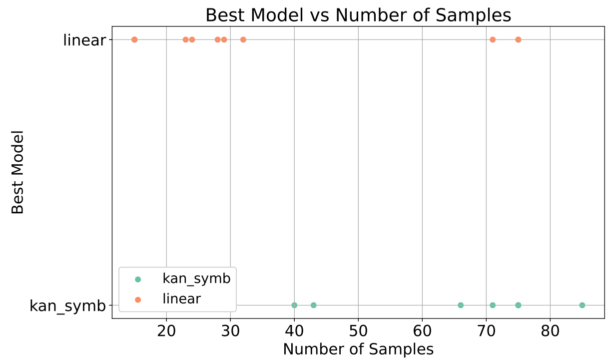

Figure 5. Which model is best with respect to the number of samples in the listening test. Here **only** interpretable models.

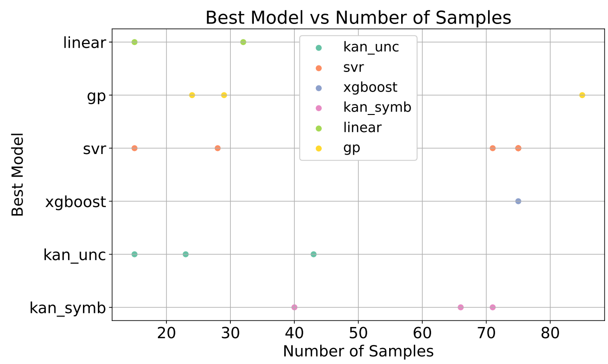

Figure 6. Which model is best with respect to the number of samples in the listening test. Here for **all** models.

5.2 KAN Interpretability

KANs not only allow for non-linear modeling but also yield interpretable symbolic formulas. Below, we provide two examples of the equations generated. First, a relatively simple one:

$$y = -0.83x_1 - 0.23x_3 - 0.57\cos(1.26x_2 - 9.62) \\ -0.23, \tag{5}$$

which is easy to interpret in terms of the roles of x_1, x_2 and x_3. To understand the advantages of interpretability, suppose we want to improve the quality of a sound but can only control x_2. If one starts with a value of $x_2 = 9.62/1.26$, we known that x_2 should increase to improve the metric, but only while $cos(1.26x_2 - 9.62)$ stays negative. This allows one to easily define a target without complex sensibility analysis. The second example is more complex (in which only two of 11 features were selected by the KAN):

$$\begin{aligned} y = -0.02(-x_1 - 0.44)^2 \\ -17.73\exp\big(-61.29(-x_1 - 0.51)^2 \\ +3.18\sin(0.23x_{11} - 0.93)\big) + 0.33. \end{aligned} \tag{6}$$

Although more challenging to interpret, it is still comprehensible, particularly because only two parameters are used, as opposed to 11 in the linear model. In this model, for example, x_1 is critical and should remain low to reduce the exponential term without favoring the first quadratic term too much. Interestingly, some KAN-derived models turned out to be essentially linear, confirming that in some scenarios linear models might be sufficient when based on relevant psychoacoustic analysis. While KANs are slightly less accurate than the best black-box methods, their interpretability can be essential for practical sound quality optimization.

6. CONCLUSIONS

We investigated the use of KANs for interpretable non-linear modeling of sound quality metrics by evaluating 19 different listening tests. Overall, we found that linear models are remarkably effective when using psychoacoustic parameters as predictors, often providing performance close to that of black-box methods. This is likely because psychoacoustic parameters themselves already capture significant non-linear aspects of human perception.

Nevertheless, some individual tests demonstrated a 0.2 MAE improvement with non-linear models (including interpretable ones like KANs), which supports a model diversity approach. KANs by their merit allow a "interpretable diversity approach", reducing error, on average, by 10% and, on individual cases, by up to 0.2 when combined with linear. Compared to linear models, KANs show particular benefits in higher-sample scenarios (around 40 or more samples), suggesting a practical rule-of-thumb for model selection.

Based on the results, we believe the main bottleneck in sound quality metric design seems to be not in the modeling itself, but rather in the feature selection part. Normally, an expert engineer needs to evaluate the sounds and find metrics with a causal relationship to the sound quality, which involves a great deal of experience. A way to automate this selection could be a good direction of future work.

7. REFERENCES

[1] D. Scouarnec and S. Bennouna, "Psychoacoustic assessment of automotive fan-systems," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 270, pp. 105–114, 2024. Presented at INTER-NOISE/NOISE-CON 2024.

[2] S. Atamer and M. E. Altinsoy, "Vacuum cleaner noise annoyance: An investigation of psychoacoustic parameters, effect of test methodology, and interaction effect between loudness and sharpness," *Applied Sciences*, vol. 13, no. 10, 2023.

[3] R. Sottek, "Sound quality evaluation of noises with spectro-temporal patterns," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 255, pp. 3927–3936, 2017. Presented at INTER-NOISE/NOISE-CON 2017.

[4] R. Sottek and T. Lobato, "AI-SQ Metrics: Artificial Intelligence in Sound Quality Metrics," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 261, pp. 3083–3091, 2020. Presented at INTER-NOISE/NOISE-CON 2020.

[5] M. Green and A. Torija Martinez, "A deep learning approach to predicting UAV sound affect," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 270, pp. 3540–3547, 2024. Presented at INTER-NOISE/NOISE-CON 2024.

[6] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark,

"Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[7] Light Vehicle Exterior Sound Level Standards Committee, "Subjective rating scale for evaluation of noise and ride comfort characteristics related to motor vehicle tires," may 2014.

[8] HEAD acoustics GmbH, "ArtemiS SUITE 16.5," 2024.

[9] ECMA, "Ecma 418-2: Methods for describing human perception based on the sottek hearing model psychoacoustic annoyance parameters," 2025.

[10] R. Sottek and K. Genuit, "Sound quality evaluation of fan noise based on advanced hearing-related parameters," *Noise Control Eng. J.*, vol. 57, no. 4, p. 384, 2009.

[11] "Iso 532-1: Acoustics — methods for calculating loudness. part 1: Zwicker method," 2017.

[12] "Iso 532-3: Acoustics — methods for calculating loudness. part 3: Moore-glasberg-schlittenlacher method," 2023.

[13] "Din 45692: Messtechnische simulation der hörempfindung schärfe," August 2009. Beuth Verlag GmbH, Berlin.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[15] F. e. a. Pedregosa, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, 2011.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.