MACHINE LEARNING: PREDICTING AUDIO QUALITY FOR HIGH-SPL SMARTPHONE RECORDINGS

Philippe Guelen, Dan Zhao, Pietro Terra Pizutti Dos Santos, Arthur Drouadene and Justin Bacle

pguelen@dxomark.com, dzhao@dxomark.com

http://corp.dxomark.com
24-26 Quai Alphonse le Gallo
92100 Boulogne-Billancourt, France

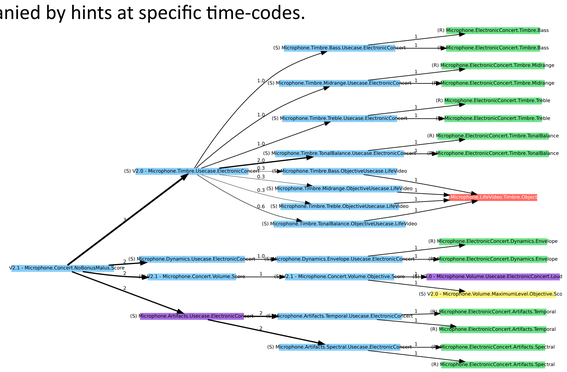
Abstract

In this paper, we explore a machine learning approach to evaluate audio quality for high sound pressure level (SPL) smartphone recordings. Our study is based on perceptual evaluations conducted by technical experts on eight audio sub-attributes (tonal balance, treble, midrange, bass, dynamics, temporal artifacts, spectral artifacts, and other artifacts) of audio quality for 121 smartphones released from 2019 to 2021. To address this task, we propose a Convolutional Neural Network (CNN) model, which proves to be a simple yet effective choice. We employ a pre-augmentation technique to enhance the training dataset size, creating a comprehensive dataset comprising recording spectrograms and corresponding perceptual evaluation scores. Our findings indicate that while the CNN model has certain limitations, it demonstrates promising capabilities in predicting evaluation scores, particularly in aspects of tonal balance, bass, and spectral artifact assessment.

Concert test protocol

The goal is to assess the audio quality of smartphones while recording musical content in concert scenarios. The Concert Use Case is performed in a custom anechoic box, where a loudspeaker is calibrated to play musical content at high-SPL (1kHz sine wave at 115 dBA, 0.3 meters). The recorded content is a combination of two audio musical clips: What's Golden by Jurassic 5 (Hip-Hop) and Hunter by Björk (Electronic). The output, for each tested smartphone, is one audio file.

The sub-attributes evaluated in our case are Tonal balance, Treble, Midrange, Bass, Dynamics (Envelope), Temporal Artifact, Spectral Artifact, and Other Artifacts. They are heavily based on audio attributes of the sound wheel described in the ITU-R 2399. Our approach to perceptual evaluation aligns with pre-established guidelines, specially curated for selected audio tracks and accompanied by hints at specific time-codes.



Concert use case scoring. Green boxes are ratings of each musical track extract.

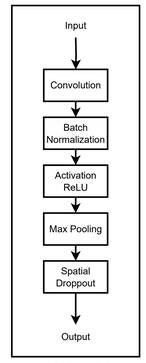
References

- Volk, C. P., Nordby, J., Stegenborg-Andersen, T., and Zacharov, N., "Efficient data collection pipeline for audio machine learning of audio quality," in Audio Engineering Society Convention 150, Audio Engineering Society, 2021.
- BS, I., "2399-0,"Methods for selecting and describing attributes and terms, in the preparation of subjective tests,"," International Telecommunications Union, 2017.

Model definition and methodology

Our study aims to explore the potential of a CNN model in predicting ratings for perceptual audio evaluations of smartphone-recorded audio tracks. These tracks are evaluated on eight single elements (each sub-attribute). Consequently, our model is designed to produce eight continuous output values, one for each evaluation criterion. In this configuration, the overall architecture incorporates eight regression heads, each constituting a prediction network composed of four convolutional layers.

For each smartphone, we calculated the arithmetic mean of the perceptual evaluation scores from the two tracks (green boxes), resulting in the ground truth score for each of the eight distinct audio sub-attributes. Similarly, we will derive these scores from the CNN inferences for our analysis.



Convolutional layer structure

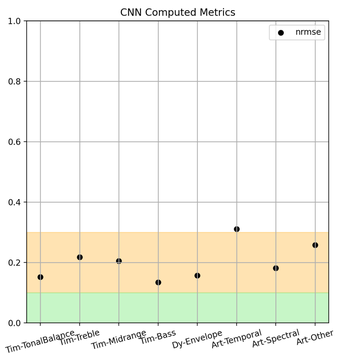
Dataset and training

The dataset covers 121 smartphones from diverse brands, geographic locations, quality tiers, and price ranges. All devices underwent evaluations with standardized conditions following the test protocol. The dataset comprises audio files and their corresponding ratings obtained from our high SPL recording use case scenario.

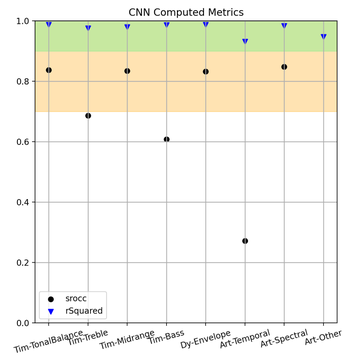
- 75% used for the training dataset (97 devices).
- Dataset pre-augmentation methods (random crop, whitening, repetition).
- 25% used for testing (24 devices).

Metrics results

We employ the normalized root mean squared error (NRMSE) as a metric to gauge the accuracy of our predictions concerning the data range. We also compute the Spearman Rank-Order Correlation Coefficient (SROCC) to assess the monotonic relationship between predicted and ground truth ratings. The ground truth values represent the scores derived from ratings provided by our sound engineers while the Prediction values are the scores computed using inferences provided by the ML model.

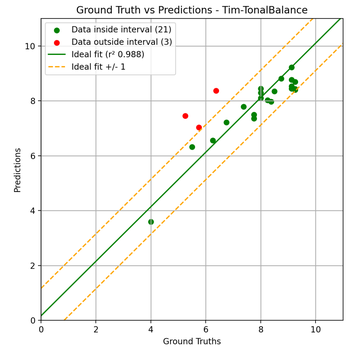


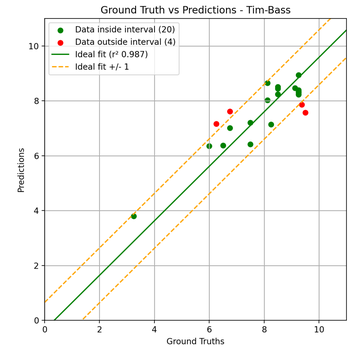
NRMSE results

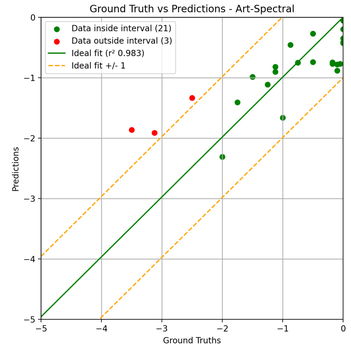


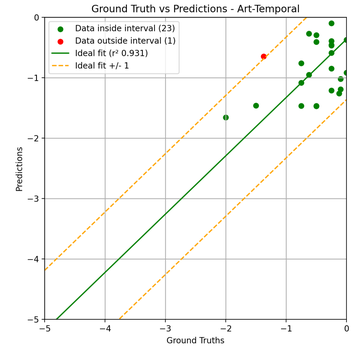
R^2 *and SROCC results*

Results by sub-attributes









Conclusion

CNN model has demonstrated satisfactory results in predicting perceptual ratings for smartphone recordings evaluations. Several improvements can be considered:

- Dataset size and pre-augmentation: Expanding the dataset and employing further data augmentation techniques, such as random signal cropping, to enhance the model's performance and generalization.
- Parameter Reduction: Consider reducing the complexity of the learning model by training it on individual regression heads for each attribute, rather than simultaneously on all eight attributes. This refinement could lead to more accurate predictions.
- Transfer Learning: employing pre-trained backbone models and fine-tuning them with our dataset solely or regression head training can leverage the advantages of large pre-existing datasets.
- Multi-Head Attention: Exploring advanced architectural choices like Multi-Head Attention may improve estimation quality while requiring less extensive training data, a valuable advantage for small datasets.

155th AES Convention - AES New York 2023