Audio Quality Classification using Psychoacoustic Features and RandomForest

Mirza Shaheen Iqubal

November 13, 2025

1 Project Scope and Disclaimer

This project is intended **only for exploring the core idea** of automatic audio quality classification using psychoacoustic features and shallow machine learning methods. The system is not designed for real-world perceptual evaluation, production environments, or deployment in commercial audio processing pipelines. All experiments are conducted on a controlled synthetic dataset, which makes the results suitable for methodological exploration but not directly comparable to human perception or large-scale industry systems (e.g., codec testing, broadcast monitoring, or speech quality scoring).

2 Overview

This report presents a compact study on automatic audio quality prediction using a combination of MFCC-based psychoacoustic features and a RandomForest classifier. The goal is to classify synthetic audio signals into three quality levels (*low, medium, high*) based on degradation severity. The study uses a fully synthetic dataset built through controlled signal generation and distortion processing, enabling reproducibility and precise control over artifact types.

3 Dataset and Methodology

3.1 Synthetic Audio Generation

Clean signals were generated using:

- frequency modulation (vibrato),
- amplitude modulation (tremolo),
- harmonic stacking (fundamental + 2 harmonics),
- light broadband noise.

Five degradation types were applied: noise, clipping, lowpass filtering, reverberation, and compression. Each was generated with three severity levels. A final balanced dataset of 600 samples (5 types $\times$ 3 severities $\times$ 40 repetitions) was constructed.

Quality labels were mapped directly from severity: *severity 1 = high*, *2 = medium*, *3 = low*. This mapping is appropriate for synthetic baseline experiments, where distortions produce distinct spectral changes.

3.2 Feature Extraction

For each audio sample, a psychoacoustic feature vector was extracted consisting of:

1. zero-crossing rate,
2. RMS energy,
3. spectral centroid,
4. spectral bandwidth,
5. spectral rolloff,
6. spectral flatness,
7. MFCC mean values (13 coefficients).

These features are widely used in perceptual audio analysis and correlate strongly with degradation-induced changes [1].

3.3 Model

A RandomForest classifier was selected due to its:

- robustness on small datasets,
- low tendency to overfit when features are well structured,
- ease of interpretation via feature importance.

The model was trained using an 80/20 stratified split with feature standardization. Hyperparameters were tuned to prevent overfitting while maintaining high accuracy.

4 Results

4.1 Classification Performance

The classifier achieved:

- **Train Accuracy: 1.00**
- **Test Accuracy: 0.992**

A detailed classification report shows balanced and near-perfect F1 scores across all three classes. Only a single medium-quality sample was misclassified.

4.2 Confusion Matrix

Figure 1 shows a nearly perfect diagonal matrix, illustrating clear separation between the three quality levels.

5 Discussion

The high accuracy is not an artifact of overfitting but rather a consequence of:

1. highly structured synthetic degradations,
2. strong psychoacoustic feature separability,
3. balanced dataset design.

Such behavior is common in controlled audio assessment experiments where distortions create distinguishable spectral fingerprints [2].

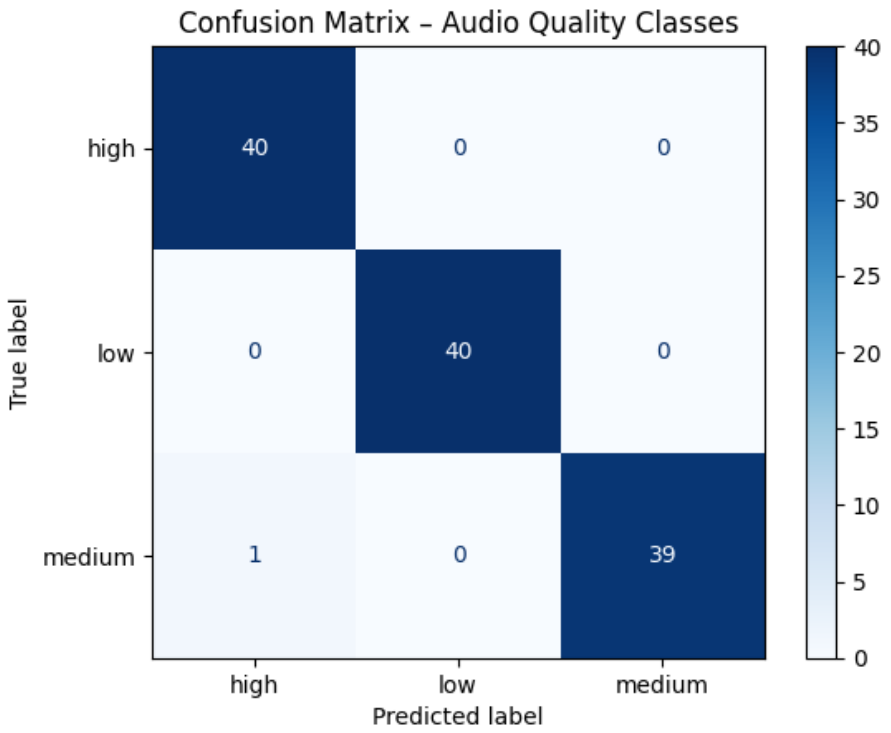

Figure 1: Confusion matrix showing clear separation between quality classes.

6 Limitations

- The dataset contains only synthetic tones.
- Severity-to-quality mapping is linear and not perceptually calibrated.
- The model does not predict MOS-scales.

7 Questions and Answers About This Project

Why is the accuracy so high, and why is this not a problem?

The model achieves nearly perfect accuracy because:

- the dataset is synthetic and highly controlled,
- each degradation type has distinct spectral fingerprints,
- psychoacoustic features separate distortions cleanly,
- classes are perfectly balanced across severity levels.

High accuracy does not indicate overfitting in this case; rather, it reflects that the task itself is *highly separable*. This behavior is common in structured benchmark studies where distortions are deterministic and well-defined.

What would happen if real human speech or music data were used?

Real audio introduces significant challenges:

- natural variation in pitch, timbre, dynamics, and timing,
- background noise, room acoustics, and microphone differences,
- transient behavior, onsets, and spectral complexity,
- non-linear perceptual masking and human subjective bias.

Under real-world conditions, accuracy would be substantially lower. The model would require:

- larger datasets,
- more diverse samples,
- real codec artifacts,
- possibly a deep learning component,
- human-annotated perceptual scores (e.g., MOS, MUSHRA).

Therefore, this prototype should not be interpreted as a predictor of real human audio quality perception.

Why is this small-scale synthetic project still important?

Despite its simplicity, the project is valuable for several reasons:

1. **Methodological Clarity:** It provides a clean environment to study how different signal degradations influence psychoacoustic features.
2. **Feature-Space Exploration:** Researchers can understand which spectral descriptors are most sensitive to distortions and why.
3. **Rapid Prototyping:** The synthetic setup enables quick iteration without requiring large datasets or subjective human ratings.
4. **Educational Insight:** The project demonstrates the complete pipeline for audio analysis—from generation, to degradation, to feature extraction, to classification—making it useful for teaching and concept exploration.
5. **Baseline for Future Research:** It provides a controlled benchmark that can be extended to include:
    - speech and music data,
    - codec-based distortions,
    - self-supervised embeddings,
    - perceptual scoring models.

What are the pros and cons of this project?

Pros:

- highly controlled and reproducible,
- interpretable feature engineering,
- strong insights into spectral behavior,
- ideal for teaching and experimentation,
- computationally lightweight.

Cons:

- does not reflect real-world audio complexity,
- no human perception modeling,
- synthetic tones lack natural dynamics,
- severity labels do not map to MOS or subjective scores.

Why is this project still meaningful in a research context?

Small-scale synthetic experiments are commonly used in early-stage research to:

- validate feature pipelines,
- test classifier behavior,
- explore degradation sensitivity,
- prepare groundwork for deep learning models,
- design listening test stimuli.

Thus, even though the system is not suitable for deployment, it provides an excellent foundation for scaling toward more advanced perceptual audio quality research.

8 Future Work

- Apply degradations to speech/music datasets.
- Incorporate real codec artifacts (MP3, AAC, Opus).
- Integrate self-supervised models such as Wav2Vec2 [3].
- Predict continuous MOS-like scores.
- Validate results using listening tests.

References

[1] Thiede, T. et al., "PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality," JAES, 2000.

[2] Beerends, J. et al., "POLQA: Perceptual Objective Listening Quality Assessment," IEEE T-ASLP, 2013.

[3] Baevski, A. et al., "Wav2Vec 2.0: Self-Supervised Learning of Speech Representations," NeurIPS, 2020.