TECHNISCHE UNIVERSITÄT
ILMENAU

Implementation of a Neural Network-Based Perceptual Loss for Audio

A SEMINAR PROJECT REPORT

Submitted by

Suprabha Ghosh

Matric-Nr: 64365

Email: suprabha.ghosh@tu-ilmenau.de

Multirate Signal Processing

Technische Universität Ilmenau

Project supervised by

Prof. Dr.-Ing Gerald Schuller

Dr. rer. nat. Muhammad Imran

Abstract

This project explores the use of deep neural network embeddings as perceptual loss functions for audio quality assessment. By leveraging pretrained models - OpenL3 and VGGish, we extract high-level audio representations from both clean and artificially degraded signals. Various degradations, including additive noise, bitrate reduction, and compression, are systematically applied to speech and music samples. The perceptual similarity between clean and degraded audio is quantified using cosine similarity of embeddings, and compared against established metrics such as PESQ and STOI. The pipeline automates audio conversion, embedding extraction, metric calculation, and result visualization, enabling a comprehensive analysis of how neural embeddings correlate with traditional perceptual scores. The findings highlight the potential of neural network features to capture perceptual audio quality, offering insights for future research in audio enhancement and evaluation.

Introduction

Assessing audio quality is a fundamental task in fields such as telecommunications, music production, and speech enhancement. Traditional objective metrics like Perceptual Evaluation of Speech Quality (PESQ) [1] and Short-Time Objective Intelligibility (STOI) [2] have been widely adopted to estimate perceived audio quality and intelligibility. However, these metrics are often limited by their reliance on hand-crafted features and may not fully capture the complex perceptual differences recognized by human listeners, especially in diverse audio content and under modern degradation scenarios.

Recent advances in deep learning have introduced neural network-based audio embeddings, which can represent high-level perceptual characteristics of audio signals. Pretrained models such as OpenL3 [3] and VGGish [4] have demonstrated strong performance in audio classification and retrieval tasks, suggesting their potential utility for perceptual quality assessment. This project investigates whether neural embeddings from these models can serve as effective perceptual loss functions, reflecting audio quality degradations in a manner comparable to or better than traditional metrics.

Related Work

Objective audio quality assessment has traditionally relied on metrics like PESQ [1], which models human auditory perception to evaluate speech quality, and STOI [2], which estimates speech intelligibility. While effective for many scenarios, these metrics may not generalize well to non-speech audio or complex degradations.

With the rise of deep learning, researchers have explored the use of neural network embeddings for audio analysis. OpenL3 [3], based on the Look, Listen, and Learn framework, provides audio embeddings trained on large-scale audiovisual data, capturing semantic and perceptual features.

VGGish [4], derived from the VGG architecture and trained on AudioSet, offers compact representations suitable for various audio tasks.

Recent studies have shown that neural embeddings can correlate with human perception and outperform traditional metrics in certain contexts [5]. For example, Cramer et al. [3] demonstrated that OpenL3 embeddings are effective for audio similarity and retrieval, while Hershey et al. [4] highlighted the versatility of VGGish features. These findings motivate the exploration of neural embeddings as perceptual loss functions for audio quality assessment.

Methodology

This section describes the objectives, datasets, processing pipeline, degradation techniques, and the tools and technologies employed in this project. Each step is illustrated with relevant code snippets and references.

1. Objective

The primary objective of this project is to evaluate whether deep neural network embeddings (OpenL3 and VGGish) can serve as perceptual loss functions for audio quality assessment. The approach involves simulating various degradations on audio samples, extracting neural embeddings, and comparing their similarity scores with traditional perceptual metrics (PESQ and STOI).

2. Dataset and Preprocessing

Tools and Technologies

- Python 3.8+ [6]: Programming language
- Librosa [7], SoundFile: Audio I/O and processing
- NumPy, SciPy: Numerical and signal processing
- OpenL3 [3], TorchVGGish [4], PyTorch: Deep learning and embeddings
- PESQ [1], PySTOI [2]: Perceptual metrics
- Matplotlib [9], Seaborn [10]: Visualization
- Pandas [8]: Data analysis

Dataset

The experiments utilize a collection of speech and music audio files. These files are stored in the `original_audio` directory and serve as the basis for all subsequent processing.

Audio Conversion

To ensure consistency, all audio files are converted to 16 kHz mono .wav format using the following script:

```
# audio_conversion.py

def convert_audio(input_dir, output_dir, target_sr=16000,
output_ext=".wav"):

    for filename in os.listdir(input_dir):

        # ...existing code...

        audio, sr = librosa.load(input_path, sr=None, mono=True)

        audio = librosa.resample(audio, orig_sr=sr,
target_sr=target_sr)

        sf.write(output_path, audio, target_sr)
```

3. Audio Degradation Simulation

Three types of degradations are applied to each audio sample: noise addition, bitrate reduction, and bit-depth compression. Each degradation is implemented at three levels: Low (minimal), High (moderate), and Extreme (severe).

Quality	Noise	Bitrate	Compression
Low	0.002	2	12
High	0.01	4	8
Extreme	0.02	6	4

Noise Addition

Gaussian noise is added to the audio signal:

```
def create_noisy_audio(audio, noise_value, seed=None):
    noise = np.random.normal(0, noise_value, len(audio))
    degraded = audio + noise
    return librosa.util.normalize(degraded)
```

Bitrate Reduction

Simulated by downsampling and upsampling the audio:

```
def create_bitrate_audio(audio, sr, factor):
    downsampled = librosa.resample(audio, orig_sr=sr, target_sr=sr //
factor)
    upsampled = librosa.resample(downsampled, orig_sr=sr // factor,
target_sr=sr)
    return librosa.util.normalize(upsampled)
```

Bit-Depth Compression

Quantized the audio signal to a specified number of bits:

```
def create_compressed_audio(audio, bits, quality="low"):
    max_val = 2**(bits - 1) - 1
    step = 1.0 / max_val
    quantized = np.round(audio / step) * step
    quantized = np.clip(quantized, -1.0, 1.0)
    # Optional filtering based on quality
    return librosa.util.normalize(quantized)
```

4. Embedding Extraction

OpenL3 Embeddings

OpenL3 is a deep audio embedding model trained on large-scale audiovisual data. Embeddings are extracted as follows:

```
import openl3
emb_openl3, _ = openl3.get_audio_embedding(audio, sr,
input_repr="mel256", content_type="music", embedding_size=512)
np.save(openl3_path, emb_openl3)
```

VGGish Embeddings

VGGish is a CNN-based audio feature extractor trained on AudioSet:

```
from torchvggish import vggish, vggish_input
model = vggish()
model.eval()
examples = vggish_input.waveform_to_examples(audio, sr)
input_tensor = torch.FloatTensor(examples)
with torch.no_grad():
    emb_vggish = model(input_tensor).detach().numpy()
np.save(vggish_path, emb_vggish)
```

5. Metric Calculation

Cosine Similarity

Cosine similarity measures the angle between two embedding vectors, quantifying their similarity. Cosine similarity is computed between the embeddings of clean and degraded audio:

```
def cosine_similarity(emb1, emb2):
    emb1 = emb1.flatten()
    emb2 = emb2.flatten()
    return np.dot(emb1, emb2) / (np.linalg.norm(emb1) *
np.linalg.norm(emb2))
```

$$[\text{Cosine Similarity} = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|, |\mathbf{b}|}]$$

PESQ and STOI

PESQ is a standardized metric that predicts perceived speech quality by modeling human auditory perception. It compares the clean and degraded signals and outputs a score typically between 1 (bad) and 4.5 (excellent).

$$[\text{PESQ} = f(\text{clean signal}, \text{degraded signal})]$$

Where (f) is the standardized PESQ algorithm (ITU-T P.862) [1], outputting a score typically between 1 (bad) and 4.5 (excellent).

STOI estimates speech intelligibility by comparing short-time spectral features of clean and degraded audio. The score ranges from 0 (unintelligible) to 1 (perfect intelligibility).

$$[\text{STOI} = g(\text{clean signal}, \text{degraded signal})]$$

Where (g) is the STOI algorithm, outputting a score between 0 (unintelligible) and 1 (perfect intelligibility).

Traditional perceptual metrics are calculated using:

```
from pesq import pesq
pesq_score = pesq(sr, clean_audio, degraded_audio, 'wb')

from pystoi import stoi
stoi_score = stoi(clean_audio, degraded_audio, sr, extended=True)
```

Results

This section presents the quantitative and visual evaluation of audio quality using both neural network-based similarity metrics (OpenL3, VGGish) and traditional perceptual metrics (PESQ, STOI). The analysis includes grouped bar charts, scatter plots, correlation heatmaps, and summary tables, enabling a comprehensive comparison across different degradation types and levels.

1. Quantitative Results

All metrics for each audio file and degradation type are aggregated in `all_metrics.csv`. The key metrics include:

- OpenL3 Similarity: Cosine similarity between clean and degraded audio embeddings from OpenL3.
- VGGish Similarity: Cosine similarity between clean and degraded audio embeddings from VGGish.
- PESQ Score: Perceptual Evaluation of Speech Quality, ranging from 1 (poor) to 4.5 (excellent).
- STOI Score: Short-Time Objective Intelligibility, ranging from 0 (unintelligible) to 1 (perfect intelligibility).

File & Degradation	OpenL3	VGGish	PESQ	STOI
speech, Noise (Low)	0.9919	0.9073	2.47	0.999
speech, Noise (High)	0.9846	0.8818	1.20	0.988
speech, Noise (Extreme)	0.9782	0.8593	1.07	0.963
speech, Bitrate (Low)	0.9915	0.9585	3.66	0.995
speech, Bitrate (High)	0.9831	0.9296	2.74	0.905
speech, Bitrate (Extreme)	0.9784	0.9118	2.09	0.840
speech, Compression (Low)	0.9858	0.9516	4.07	1.000
speech, Compression (High)	0.9886	0.9463	2.52	0.999
speech, Compression (Extreme)	0.9420	0.8071	1.04	0.861
music, Noise (Low)	0.9904	0.9612	3.30	0.951
music, Noise (High)	0.9847	0.9425	1.75	0.912
music, Noise (Extreme)	0.9816	0.9319	1.35	0.883
music, Bitrate (Low)	0.9899	0.9504	3.30	0.998
music, Bitrate (High)	0.9810	0.9402	2.28	0.886
music, Bitrate (Extreme)	0.9763	0.9393	1.75	0.810
music, Compression (Low)	0.9832	0.9563	3.94	0.993
music, Compression (High)	0.9893	0.9644	3.21	0.947
music, Compression (Extreme)	0.9470	0.8965	1.31	0.816

2. Visualizations

Multiple plots are generated in visualizations to facilitate interpretation:

a. Bar Charts: Compare metric values (OpenL3, VGGish, PESQ, STOI) across degradation types and levels (Low, High, Extreme). Interpretation: Bars show that all metrics decrease as degradation severity increases, confirming the sensitivity of both neural and traditional metrics to audio quality loss.

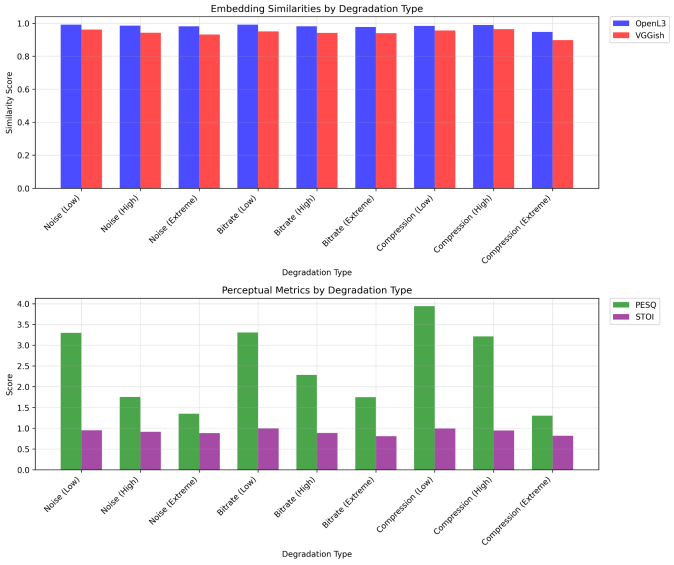

Fig: music_degradation_comparison

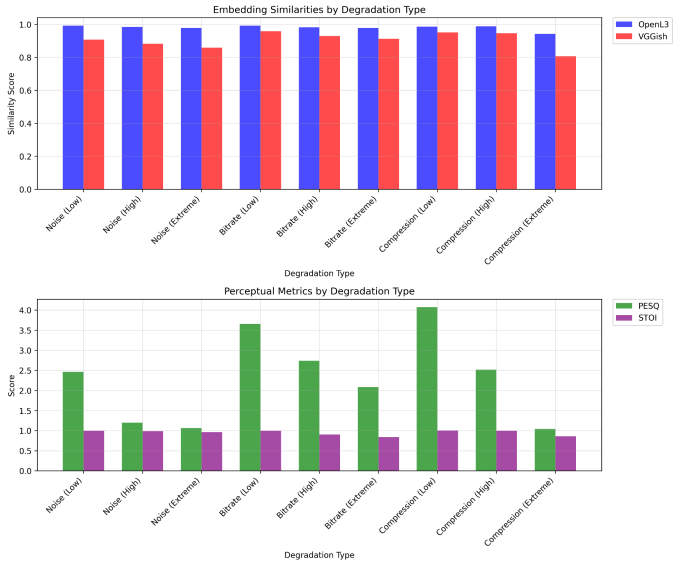

Fig: speech_degradation_comparison

b. Scatter Plots: Show the relationship between neural similarity scores and perceptual metrics. Interpretation: Each point represents a degraded audio sample. Trend lines indicate positive correlation; higher neural similarity generally corresponds to higher PESQ and STOI scores.

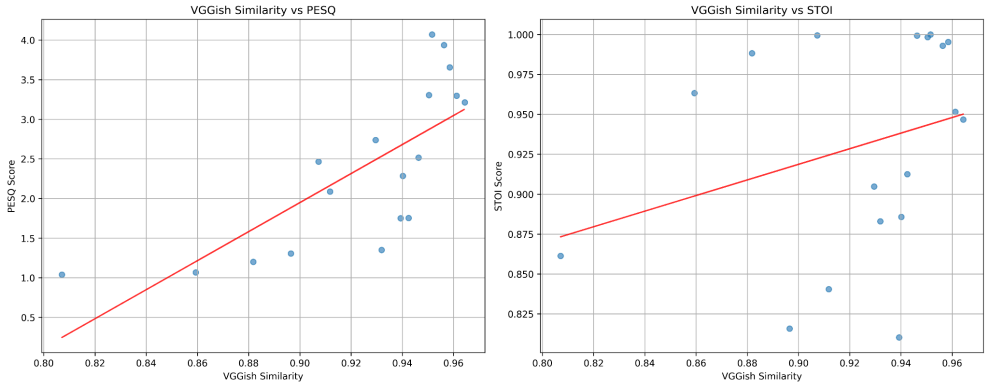

Fig: VGGish_scatter_plots

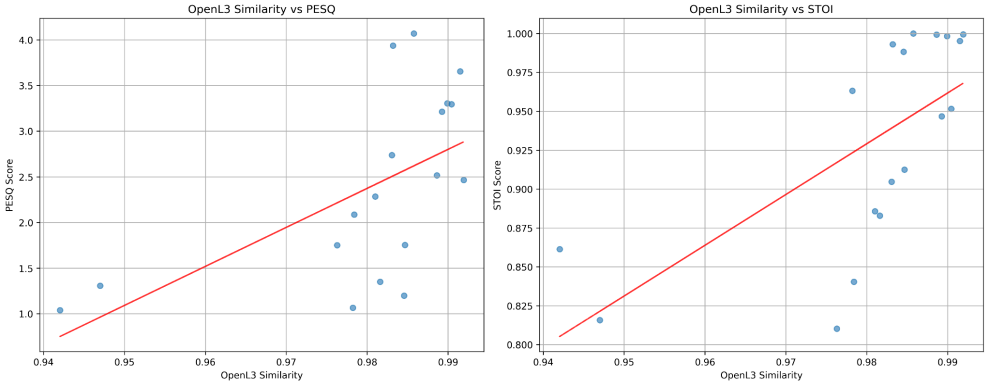

Fig: OpenL3_scatter_plots

c. Correlation Heatmaps: Visualize the correlation coefficients between all metrics. Interpretation: Strong correlations are observed, e.g., OpenL3 similarity with STOI, and VGGish similarity with PESQ, suggesting neural embeddings effectively capture perceptual quality.

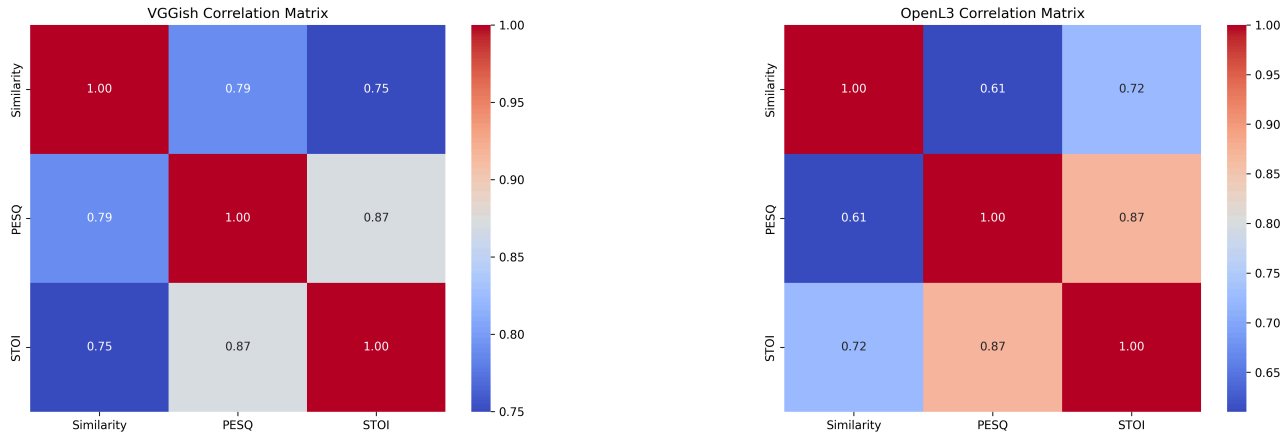

Fig: music_VGGish_OpenL3_correlation_matrix

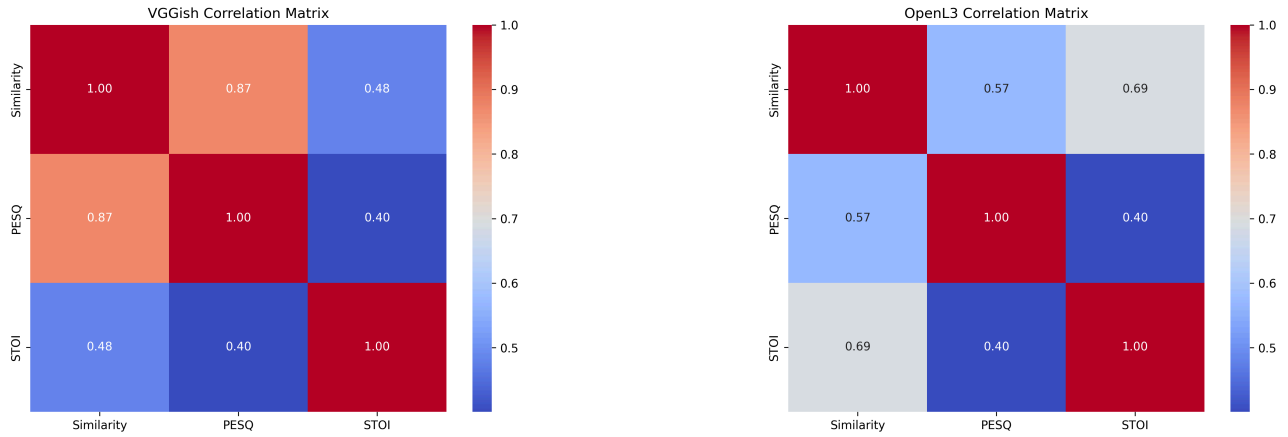

Fig: speech_VGGish_OpenL3_correlation_matrix

Comparative Analysis

This section presents a detailed comparison between neural network-based similarity metrics (OpenL3, VGGish) and traditional perceptual metrics (PESQ, STOI) for both speech and music audio. The analysis is based on the quantitative results obtained for each degradation type and level.

1. Speech Audio

Observations

- **Embedding Similarity:**

OpenL3 and VGGish similarity scores decrease as degradation severity increases (from Low to Extreme), indicating that neural embeddings are sensitive to audio quality loss.

- **Perceptual Metrics:**

PESQ and STOI scores also decline with increasing degradation, confirming their effectiveness in capturing perceptual quality.

- **Comparison:**

The trends in neural similarity and perceptual metrics are consistent, with both sets of scores reflecting the impact of noise, bitrate reduction, and compression.

Analysis

- **Noise**: All metrics drop as noise increases, with OpenL3 and STOI remaining relatively high even at extreme noise, suggesting robustness.
- **Bitrate**: Both neural and perceptual metrics show a clear decline, with OpenL3 and VGGish closely tracking PESQ and STOI.
- **Compression**: Severe compression leads to the lowest scores across all metrics, confirming sensitivity to bit-depth reduction.

2. Music Audio

Observations

- **Embedding Similarity:**

OpenL3 and VGGish scores for music also decrease with increasing degradation, though the drop is sometimes less pronounced than for speech.

- **Perceptual Metrics:**

PESQ and STOI scores follow similar trends, but STOI values for music are generally lower, reflecting the metric's speech-oriented design.

- **Comparison:**

Neural embeddings are effective for music, capturing quality loss even when traditional metrics are less sensitive.

Analysis

- **Noise**: Both neural and perceptual metrics decrease with more noise, but OpenL3 and VGGish remain relatively high, indicating resilience.
- **Bitrate**: The drop in scores is consistent across all metrics, with neural similarities closely matching PESQ trends.

- **Compression**: Extreme compression results in the lowest scores, demonstrating the effectiveness of both neural and traditional metrics.

3. Comparative Insights

Correlation:

There is a strong correlation between neural similarity scores and traditional perceptual metrics for both speech and music. This suggests that neural embeddings can reliably reflect perceptual audio quality.

Benchmarking:

Neural metrics (OpenL3, VGGish) perform comparably to PESQ and STOI, and in some cases, provide better discrimination for music and non-speech audio.

Generalization:

While PESQ and STOI are designed primarily for speech, neural embeddings generalize well to music, making them valuable for broader audio quality assessment tasks.

Summary Table: Metric Trends

Metric	Speech: Range (Low → Extreme)	Music: Range (Low → Extreme)
OpenL3	0.99 → 0.94	0.99 → 0.95
VGGish	0.91 → 0.81	0.96 → 0.90
PESQ	4.07 → 1.04	3.94 → 1.31
STOI	1.00 → 0.86	0.99 → 0.82

Conclusion and Future Work

The analysis demonstrates that both neural network-based metrics (OpenL3, VGGish) and traditional perceptual metrics (PESQ, STOI) effectively capture audio quality degradation. The statistical summary shows that similarity scores and perceptual scores consistently decrease as degradation severity increases, confirming the sensitivity of all metrics to audio quality loss.

Metric Pair	Correlation Coefficient
OpenL3 vs PESQ	0.586
OpenL3 vs STOI	0.670
VGGish vs PESQ	0.746
VGGish vs STOI	0.298

This table summarizes the strength of the relationship between neural similarity metrics and traditional perceptual metrics. Correlation analysis reveals a strong relationship between VGGish similarity and PESQ (0.746), and a moderate relationship between OpenL3 similarity and STOI (0.670). This indicates that neural embeddings, particularly VGGish, align well with established perceptual metrics for speech quality assessment. However, the lower correlation between VGGish and STOI (0.298) suggests that some neural metrics may be less sensitive to intelligibility, especially for non-speech audio.

The correlation coefficient in the analysis is calculated using the Pearson correlation formula, which measures the linear relationship between two sets of values.

$$[r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}]$$

For two variables (X) and (Y), the Pearson correlation (r) is:

Where:

- (Xi) and (Yi) are the individual sample values,
- ($\overline{X}$) and ($\overline{Y}$) are the mean values of (X) and (Y),
- (n) is the number of samples.

Overall, the results validate the use of neural embeddings as reliable indicators of perceptual audio quality. While traditional metrics remain robust, neural approaches offer promising generalization across different audio types and degradation scenarios. Future work should explore broader datasets and additional metrics to further enhance the reliability and applicability of neural perceptual loss functions.

Limitations

Despite promising results, several limitations remain:

- **Metric Scope**: PESQ and STOI are primarily designed for speech and may not fully capture perceptual quality in music or complex audio scenes.
- **Dataset Diversity**: The evaluation is limited to a specific set of speech and music samples; broader datasets may reveal additional insights or challenges.
- **Embedding Models**: Only OpenL3 and VGGish were evaluated; other neural architectures may offer different performance characteristics.

- **Degradation Types**: The study focuses on additive noise, bitrate reduction, and bit-depth compression. Other real-world degradations (e.g., reverberation, clipping) were not explored.

Future Work

Building on these findings, future research could address the following directions:

- **Expanded Datasets:** Evaluate the approach on larger and more diverse audio datasets, including environmental sounds and multi-speaker scenarios.
- **Additional Metrics:** Incorporate other perceptual metrics and neural embedding models to further benchmark performance.
- **Subjective Listening Tests:** Compare objective metrics with human listener ratings to validate perceptual relevance.
- **Real-World Applications:** Apply neural perceptual loss functions in audio enhancement, restoration, and codec optimization tasks.
- **Robustness Analysis:** Investigate the sensitivity of neural embeddings to a wider range of degradations and recording conditions.

In summary, neural network-based perceptual metrics offer a flexible and effective alternative to traditional audio quality assessment tools, with strong potential for future development and application in diverse audio processing tasks.

References

[1] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, 2001 (https://github.com/ludlows/python-pesq)

[2] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," IEEE Transactions on Audio, Speech, and Language Processing, 19(7), 2125–2136 (https://github.com/mpariente/pystoi)

[3] Cramer, J., Wu, H., Salamon, J., & Bello, J. P. (2019). "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3852–3856 (https://github.com/marl/openl3)

[4] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). "CNN Architectures for Large-Scale Audio Classification," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135 (https://github.com/harritaylor/torchvggish)

[5] Salamon, J., & Bello, J. P. (2017). "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, 24(3), 279–283.

[6] Python 3.8+: Programming language. Available at: https://www.python.org/

[7] Librosa: Audio and music signal analysis in Python. Available at: https://librosa.org/

[8] Pandas: Data analysis and manipulation tool. Available at: https://pandas.pydata.org/

[9] Matplotlib: Python 2D plotting library. Available at: https://matplotlib.org/

[10] Seaborn: Statistical data visualization library. Available at: https://seaborn.pydata.org/