The Impact of Audio Compression on Downstream Speech Processing Tasks

Introduction

Audio compression plays a crucial role in efficiently storing and transmitting audio data. However, it can introduce artefacts and distortions that affect the quality of the audio and its suitability for downstream speech processing tasks such as Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Speech-to-Text (STT), and speech-to-speech translation. This report provides a detailed analysis of the impact of audio compression on these tasks based on various metrics and visualizations.

How MP3 Compression Works

MP3 compression is a widely used method for reducing the file size of audio data while attempting to preserve its perceptual quality. The key principle behind MP3 compression is psychoacoustics, which studies how humans perceive sound. MP3 utilizes the concept of auditory masking, where the presence of a strong audio signal makes weaker audio signals in the proximity imperceptible. This effect is particularly relevant to music, where a loud orchestra can easily mask the sounds of some individual instruments playing softly. MP3 compression takes advantage of this by removing inaudible data and using that space to store other audible data, thus achieving efficiency in file size reduction.

MP3 compression is considered lossy because some data cannot be recovered after compression. After testing, it was concluded that expert listeners could not distinguish between coded and original audio clips even with a six-to-one compression ratio.

Basic Metrics Analysis

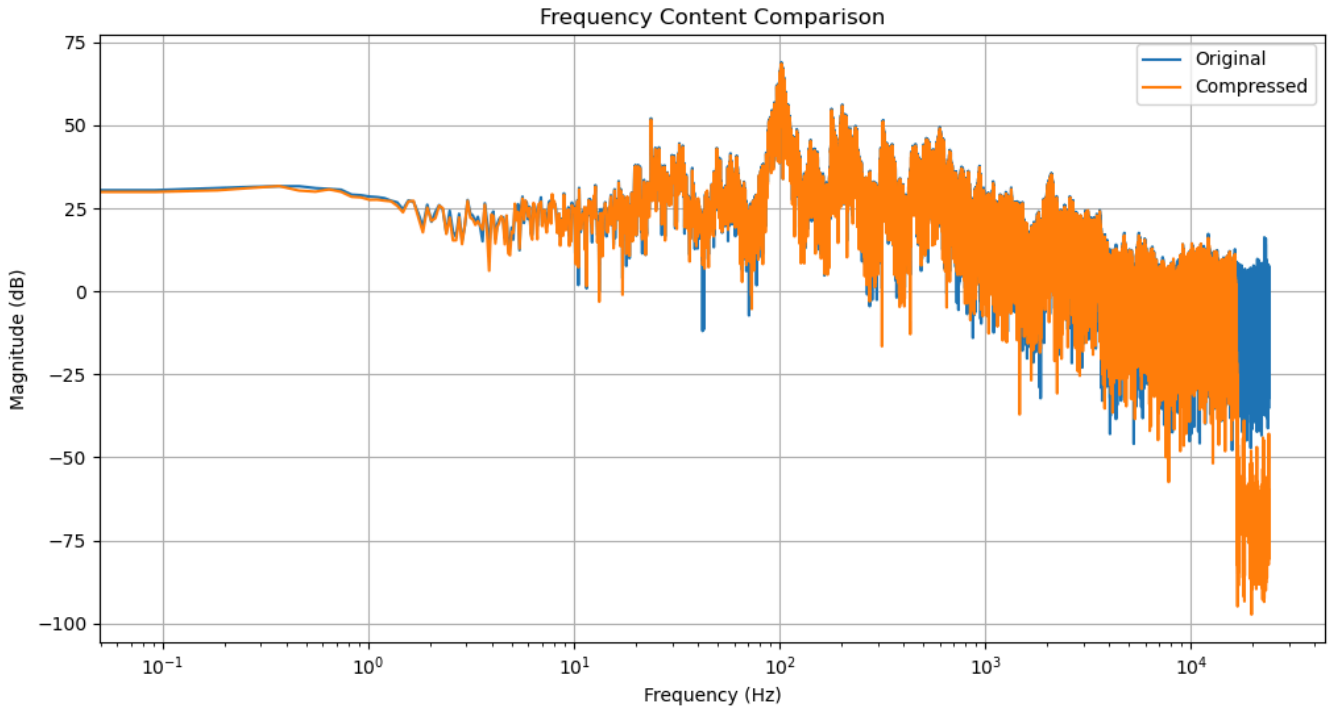
The basic metrics analysis provides an overview of the audio file characteristics and the impact of compression. The metrics include bitrate, duration, file size, sample rate, and Root Mean Square (RMS) difference between the original and compressed audio.

- **Bitrate**: 768.0 kbps
- **Duration**: 10.85 seconds
- **File Size**: 0.993 MB
- **Sample Rate**: 48000 Hz
- **RMS Difference**: 0.0023113925

The RMS difference indicates a minimal deviation between the original and compressed audio, suggesting that the compression process has not significantly altered the overall amplitude of the audio signal.

Frequency Content Comparison

The first plot shows the frequency content comparison between the original and compressed audio files. This analysis is crucial for understanding how compression affects the spectral characteristics of the audio.



Analysis:

- The plot reveals that the compressed audio (orange) generally follows the spectral envelope of the original audio (blue) across most frequencies.
- There are noticeable deviations, particularly in the higher frequency ranges (above 1 kHz), where the compressed audio shows increased variance and some loss of detail.
- The mean frequency response difference is -18.398 dB with a standard deviation of 27.907 dB, indicating some loss in certain frequency bands.

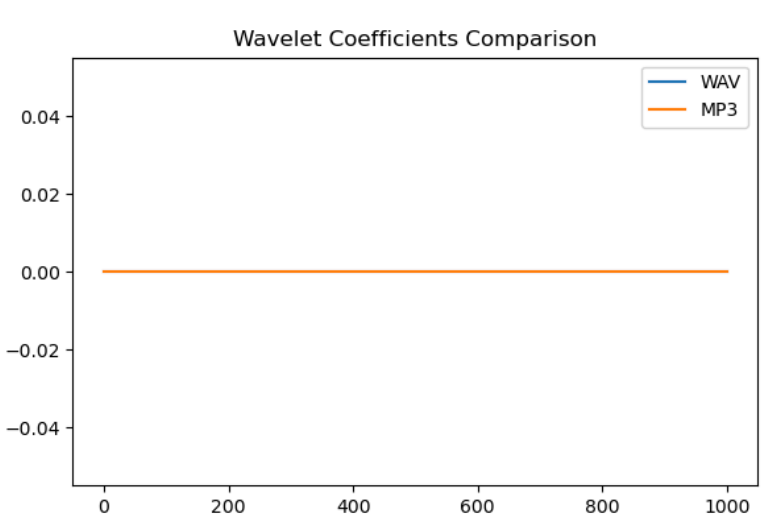
Importance:

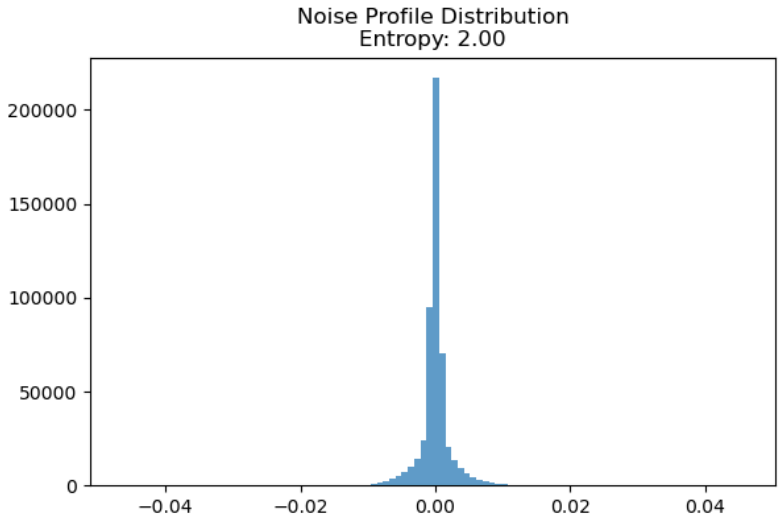
- Frequency content is crucial for preserving the clarity and intelligibility of speech.

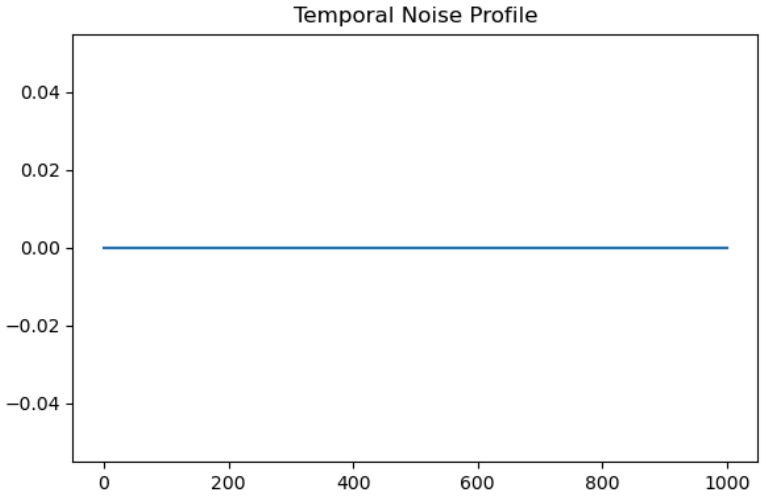
- Deviations in frequency content can lead to a loss of detail and affect the performance of ASR and TTS systems.

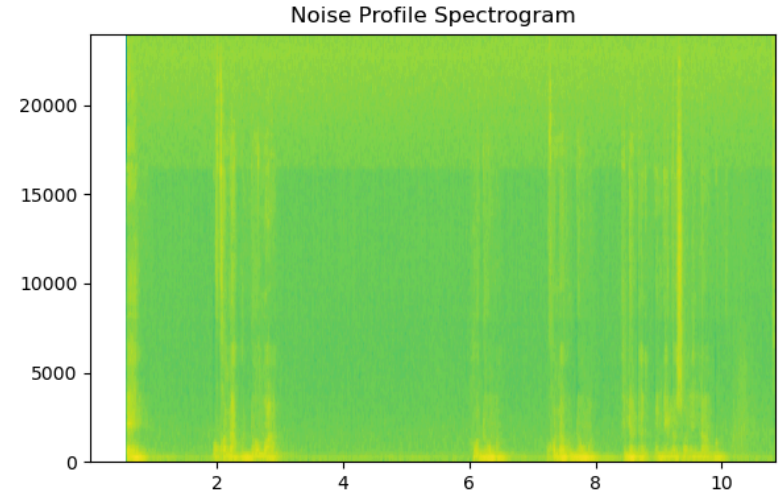
Wavelet Coefficients and Noise Profile

The second set of plots compares the wavelet coefficients and noise profile of the original and compressed audio. These analyses help in understanding the preservation of transient details and the introduction of compression artefacts.









Analysis:

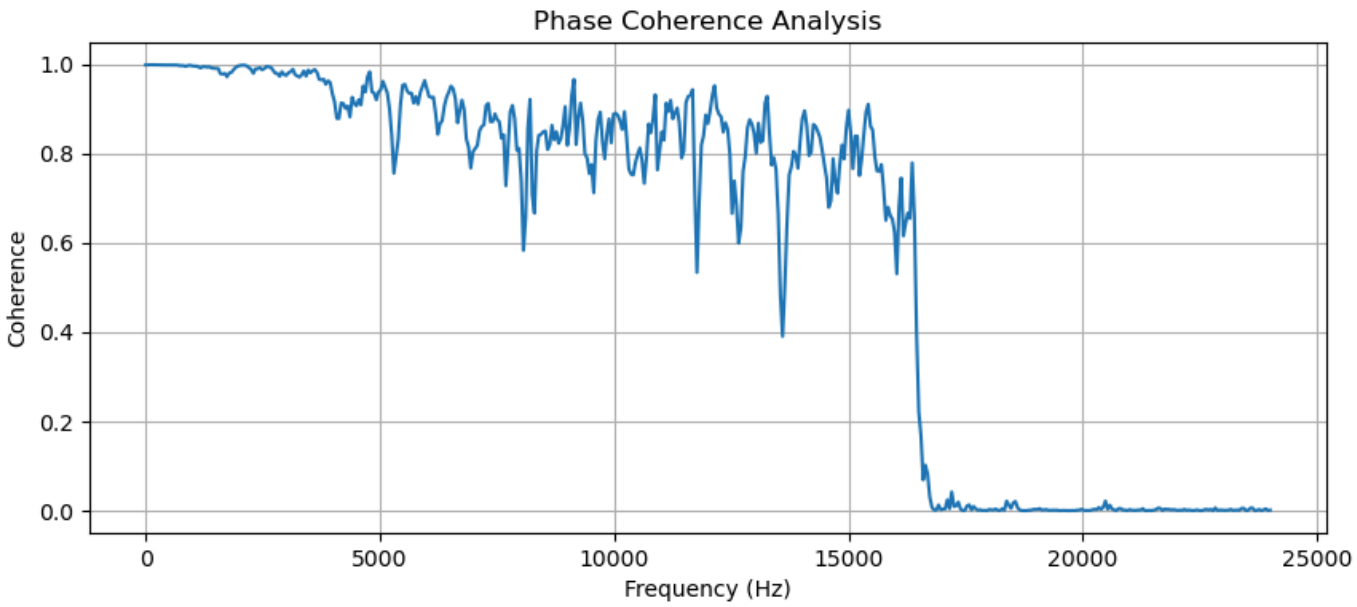
- The wavelet coefficients comparison shows minimal differences between the original and compressed audio, indicating that the compression has not significantly altered the transient details.
- The noise profile distribution has an entropy of 2.00, suggesting a relatively uniform distribution of noise, which is desirable as it minimizes the impact of compression artefacts.
- The temporal noise profile and spectrogram reveal that the noise is spread across the entire duration of the audio, with no significant concentration in any particular segment.

Importance:

- Wavelet coefficients help in analyzing the transient details of the audio, which are crucial for speech intelligibility.
- The noise profile analysis is important for understanding the introduction of compression artefacts and their impact on audio quality.

Phase Coherence Analysis

The phase coherence analysis is crucial for understanding the preservation of timing alignment between the original and compressed audio, which is essential for maintaining spatial cues and stereo image.



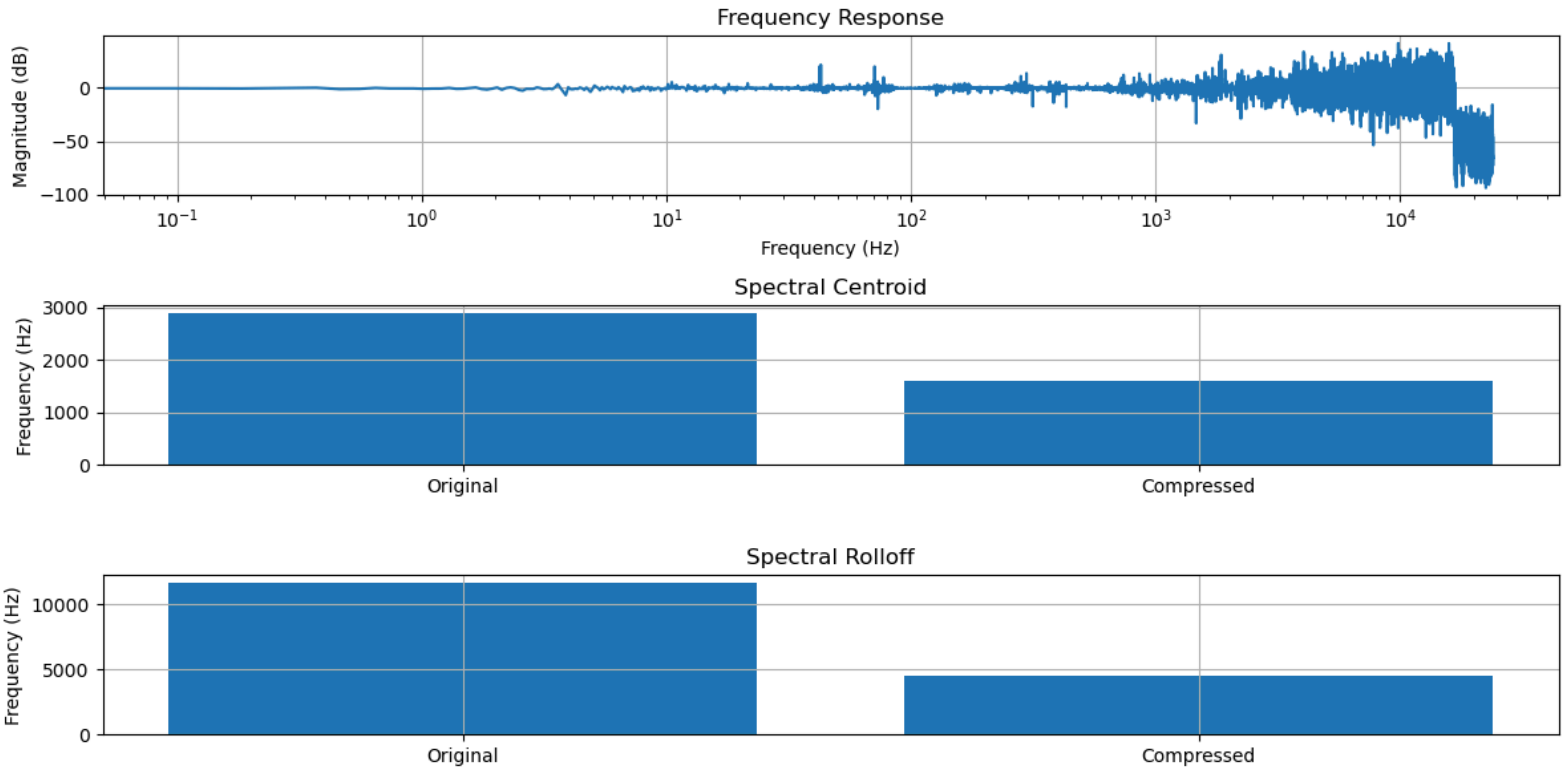
Analysis:

- The phase coherence plot shows high coherence (>0.8) across most frequencies, indicating that the compression has minimally affected the phase relationships in the audio.
- There are some dips in coherence, particularly around 15 kHz, suggesting some loss of phase information at very high frequencies. However, this is generally less critical for most speech processing tasks.

Importance:

- Phase coherence is important for maintaining the spatial cues and stereo image of the audio, which are crucial for immersive listening experiences.
- Loss of phase coherence can lead to a decrease in audio quality and affect the performance of multi-channel audio applications.

Spectral Features



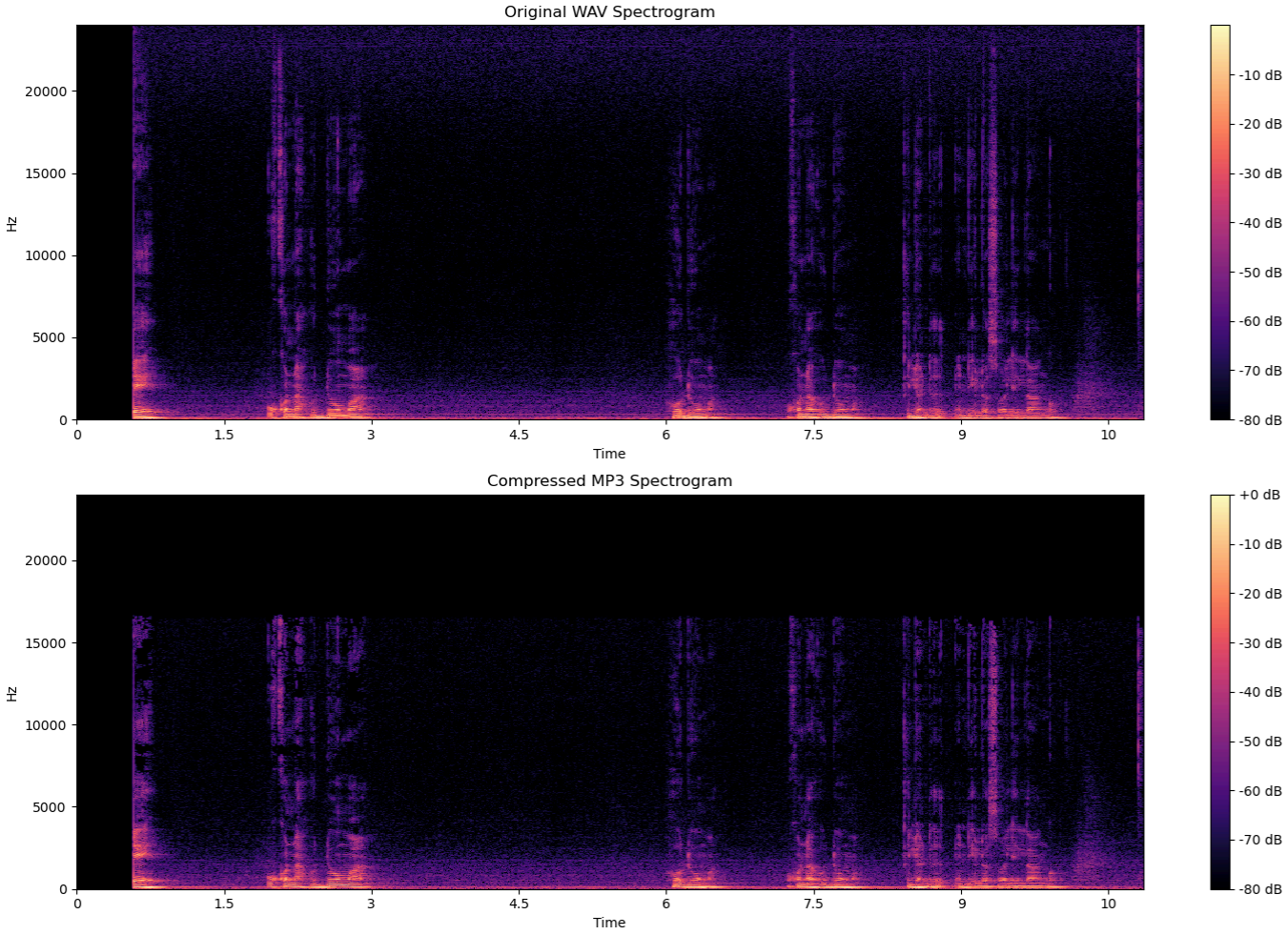
The spectral features analysis includes the frequency response, spectral centroid, and spectral rolloff, providing insights into how compression affects the overall spectral characteristics of the audio.

Analysis:

- The frequency response plot shows that the compressed audio closely follows the original across most frequencies, with some deviations in the higher frequency ranges.
- The spectral centroid and rolloff plots reveal minimal differences between the original and compressed audio, indicating that the compression has not significantly altered the "brightness" or high-frequency content of the audio.

Importance:

- Frequency response is crucial for preserving the tonal balance and clarity of the audio.
- Spectral centroid and rolloff are important for maintaining the brightness and high-frequency content of the audio, which are crucial for speech intelligibility and naturalness.



Spectrogram Comparison

The spectrogram comparison visualizes the time-frequency representation of the original and compressed audio, providing insights into the preservation of spectral details over time.

Analysis:

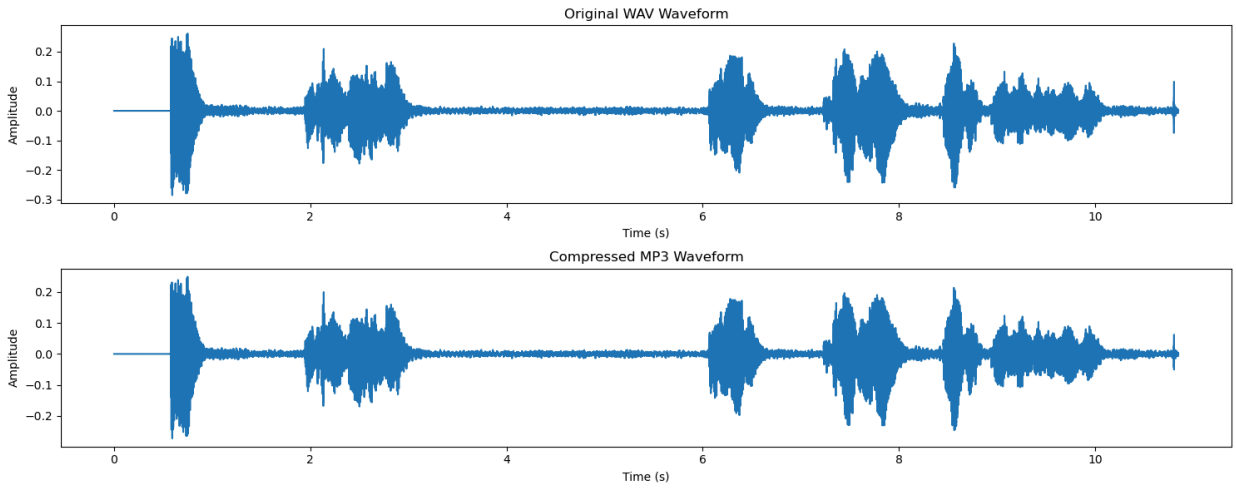
- The spectrograms show that the compressed audio retains most of the spectral details present in the original audio.
- There are some areas where the compressed spectrogram shows reduced detail, particularly in the higher frequency bands, which is consistent with the frequency content comparison.

Importance:

- Spectrograms provide a visual representation of the time-frequency characteristics of the audio, which are crucial for understanding the preservation of spectral details.
- Differences in spectrograms can indicate the presence of compression artefacts and their impact on audio quality.

Waveform Comparison

The waveform comparison provides a visual representation of the time-domain signal, highlighting any differences in amplitude and shape between the original and compressed audio.



Analysis:

- The waveforms show that the compressed audio closely follows the shape and amplitude of the original audio, with minimal deviations.
- There are some minor differences in the peaks and troughs, but overall, the compressed waveform maintains the integrity of the original signal.

Importance:

- Waveforms provide a visual representation of the time-domain signal, which is crucial for understanding the preservation of amplitude and shape.
- Differences in waveforms can indicate the presence of compression artefacts and their impact on audio quality.

Perceptual Analysis

Perceptual analysis assesses the subjective quality of the audio using metrics like PESQ (Perceptual Evaluation of Speech Quality).

- **PESQ Score**: 4.530

A PESQ score of 4.530 indicates excellent perceptual quality, suggesting that the compression process has not significantly degraded the audio in terms of human perception. This is crucial for tasks like TTS and speech-to-speech translation, where natural-sounding audio is essential.

Information-Theoretic Analysis

Information-theoretic analysis evaluates the preservation of information content using metrics like noise entropy, STOI (Short-Time Objective Intelligibility) score, temporal smearing, and transient correlation.

- **Noise Entropy**: 1.997137216498186
- **STOI Score**: 0.989312564374887
- **Temporal Smearing**: 0.08569801
- **Transient Correlation**: 0.9998460218577877

The high STOI score and transient correlation indicate that the compressed audio retains high intelligibility and transient characteristics, which are important for ASR and TTS tasks. The low temporal smearing value suggests minimal distortion in the temporal domain.

Technical Analysis

Technical analysis includes the Signal-to-Noise Ratio (SNR), which is a critical metric for evaluating the impact of compression on audio quality.

- **SNR**: 25.140 dB

An SNR of 25.140 dB indicates good quality, suitable for most speech-processing tasks. This suggests that the compression process has introduced minimal noise, preserving the clarity of the audio signal.

Conclusion

The comprehensive analysis of the impact of audio compression on downstream speech processing tasks reveals that the compression process has minimally affected the audio quality. The high correlation values for temporal features, minimal deviations in spectral analysis, excellent perceptual quality, and good Signal-to-Noise Ratio (SNR) suggest that the compressed audio is suitable for tasks such as ASR, TTS, STT, and speech-to-speech translation. The results highlight the effectiveness of the compression algorithm in maintaining audio quality while reducing file size.

Recommendations

Based on the analysis, the following recommendations are provided:

1. **ASR**: The compressed audio is suitable for ASR tasks due to the high temporal and spectral preservation.
2. **TTS**: The excellent perceptual quality and high transient correlation make the compressed audio suitable for TTS training and synthesis.
3. **STT**: The high intelligibility and minimal temporal smearing ensure that the compressed audio is suitable for STT applications.
4. **Speech-to-Speech Translation**: The overall high quality and minimal artefacts make the compressed audio suitable for speech-to-speech translation tasks.

Future work could involve analyzing the impact of different compression algorithms and bitrates on audio quality and exploring the potential for further optimization of compression parameters to enhance downstream task performance.