Speech Synthesis using deep learning

Why use deep learning for speech synthesis?

1.

Overcoming Limitations of Traditional Techniques

•Traditional speech synthesis methods often result in robotic-sounding speech lacking natural prosody and expressiveness.
•Deep learning-based approaches have revolutionized speech synthesis by addressing these limitations.
•Neural networks capture intricate patterns in speech, allowing for more natural and emotionally engaging voices.

2.

High-Quality Deep Learning Text-to-Speech (TTS) Models

•Deep learning empowers the development of high-quality TTS models with remarkable realism and clarity.
•Neural TTS models effectively capture the intricate nuances of language and acoustic features, resulting in improved intonation, rhythm, and pronunciation.
•These advancements contribute to more lifelike and intelligible synthesized speech, enhancing the overall user experience and engagement.

Speech Synthesis Process explained

Mel-Spectrogram Generation

First, you need to generate a mel-spectrogram from the input text using a text-to-speech (TTS) model such as Tacotron.

This involves encoding the text into linguistic and acoustic features and decoding them into a mel-spectrogram representation.

Inverse Mel-Spectrogram

Next, you perform an inverse operation on the mel-spectrogram to obtain a linear-scaled spectrogram.

This step is necessary because most vocoders operate on linear-scale spectrograms.

Vocoder Model

You then feed the linear-scaled spectrogram into a vocoder model. A popular choice is the WaveNet vocoder, which is a deep generative model that can synthesize high-quality audio.

The vocoder model takes the linear-scaled spectrogram as input and generates a time-domain waveform which is your output audio file.

Making our own transformer from scratch

We used the LJ Speech and LibriSpeech datasets to train our model, here’s a look into LibriSpeech

OVERVIEW

•Size: Approximately 1,000 hours of clean English speech data.
•Structure: Divided into subsets for training, development, and evaluation.
•Chapter-Based Organization: Audio files paired with corresponding text transcripts.

DATA SPLITS

•Training Sets: The LibriSpeech dataset provides various subsets for training ASR models, such as "train-clean-100" and additional sets with varying levels of noise and transcription errors.
•Test Set: The test subsets serve as the final evaluation benchmark to assess the accuracy and generalization of TTS models.

TRANSCRIPTION

Manual Transcription:
"clean" subsets have undergone careful manual transcription, ensuring high-quality and accurate transcriptions.

Automatic Transcription:
"other" subsets include automatic transcriptions, presenting a more challenging evaluation set with potential errors.

APPLICATIONS

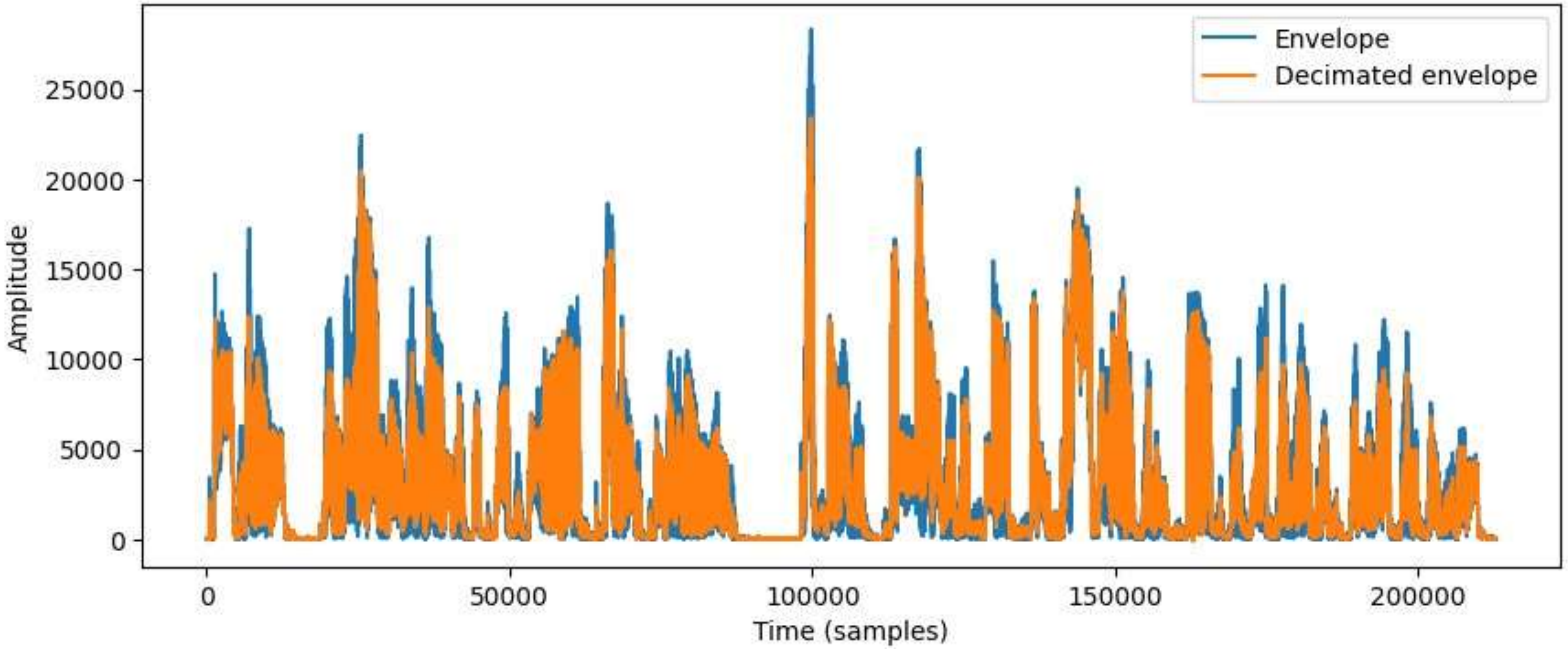
Automatic Speech Recognition Models
LibriSpeech serves as a standard benchmark dataset for training and evaluating ASR models.

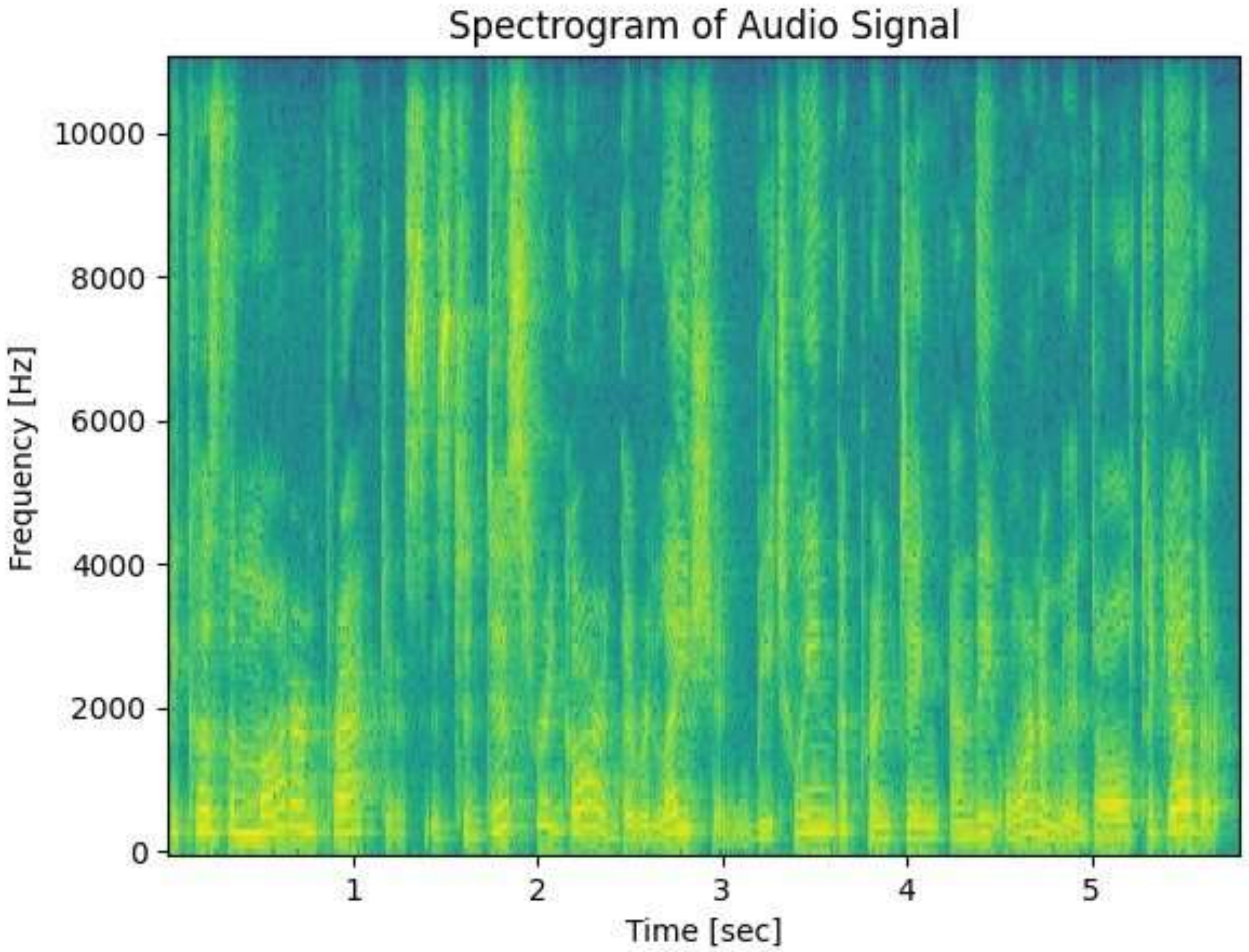
TTS Model Training:
The dataset is utilized for training text-to-speech (TTS) models to synthesize high-quality, natural-sounding speech.

Making our own transformer from scratch

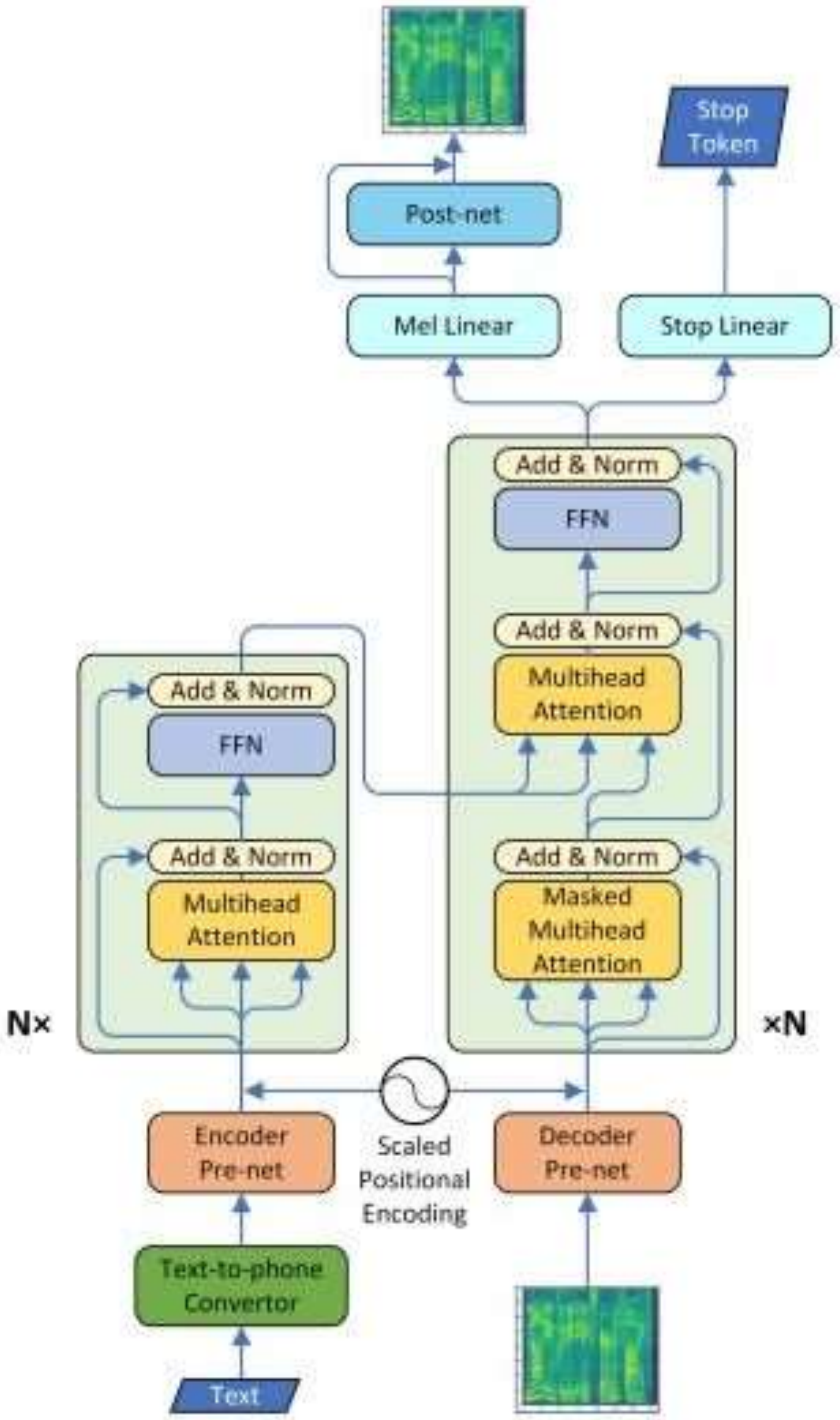
A look into LJ Speech dataset

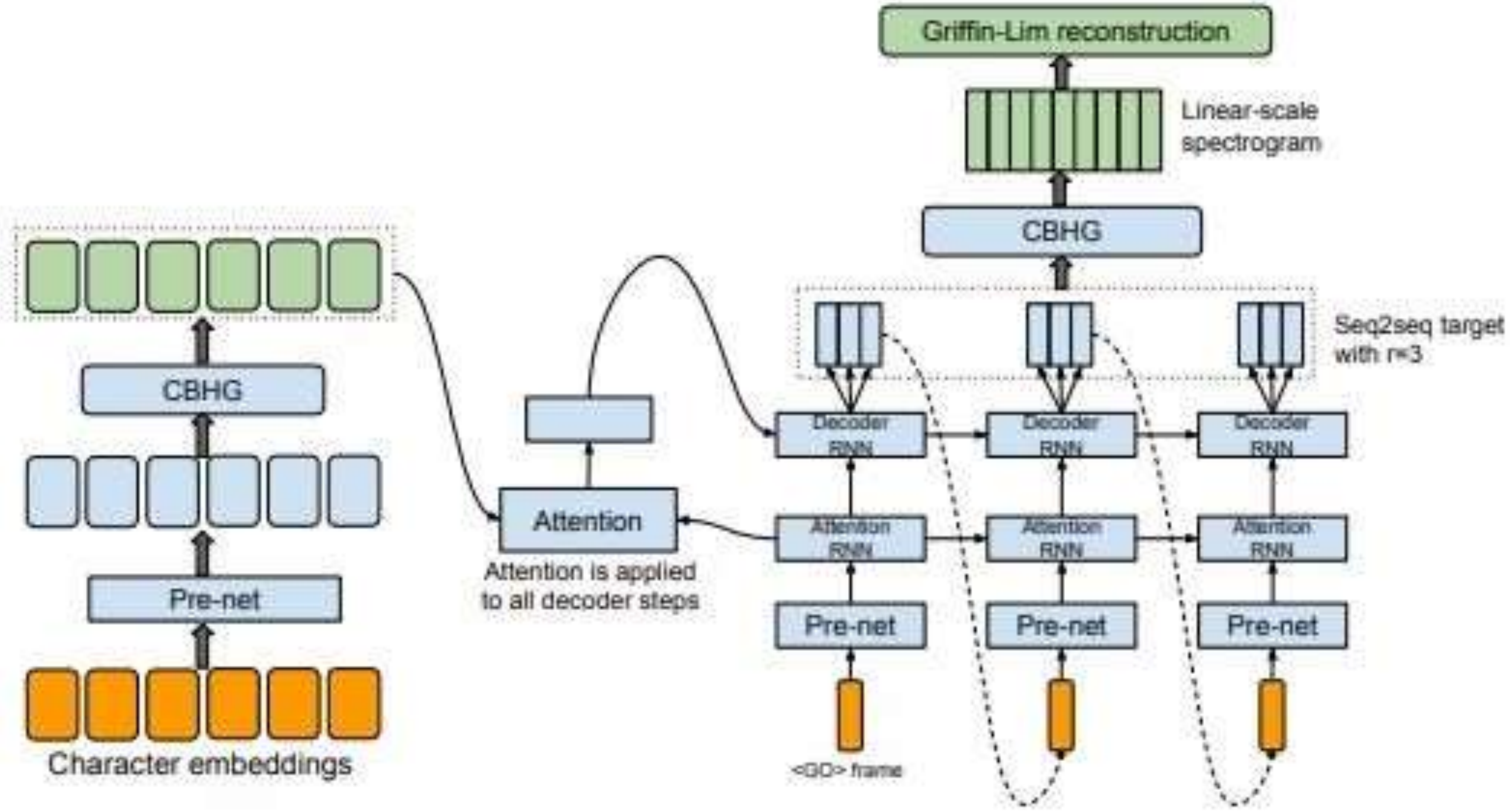
Magnitude vaíiations of the audio signal oveí time





Transformer Architecture





Transformer Architecture explained

Encoder-Decoder Framework

The custom transformer model follows an encoder-decoder architecture. The encoder processes the input text, generating a contextual representation. The decoder then generates mel-spectrograms, capturing the acoustic features of the speech.

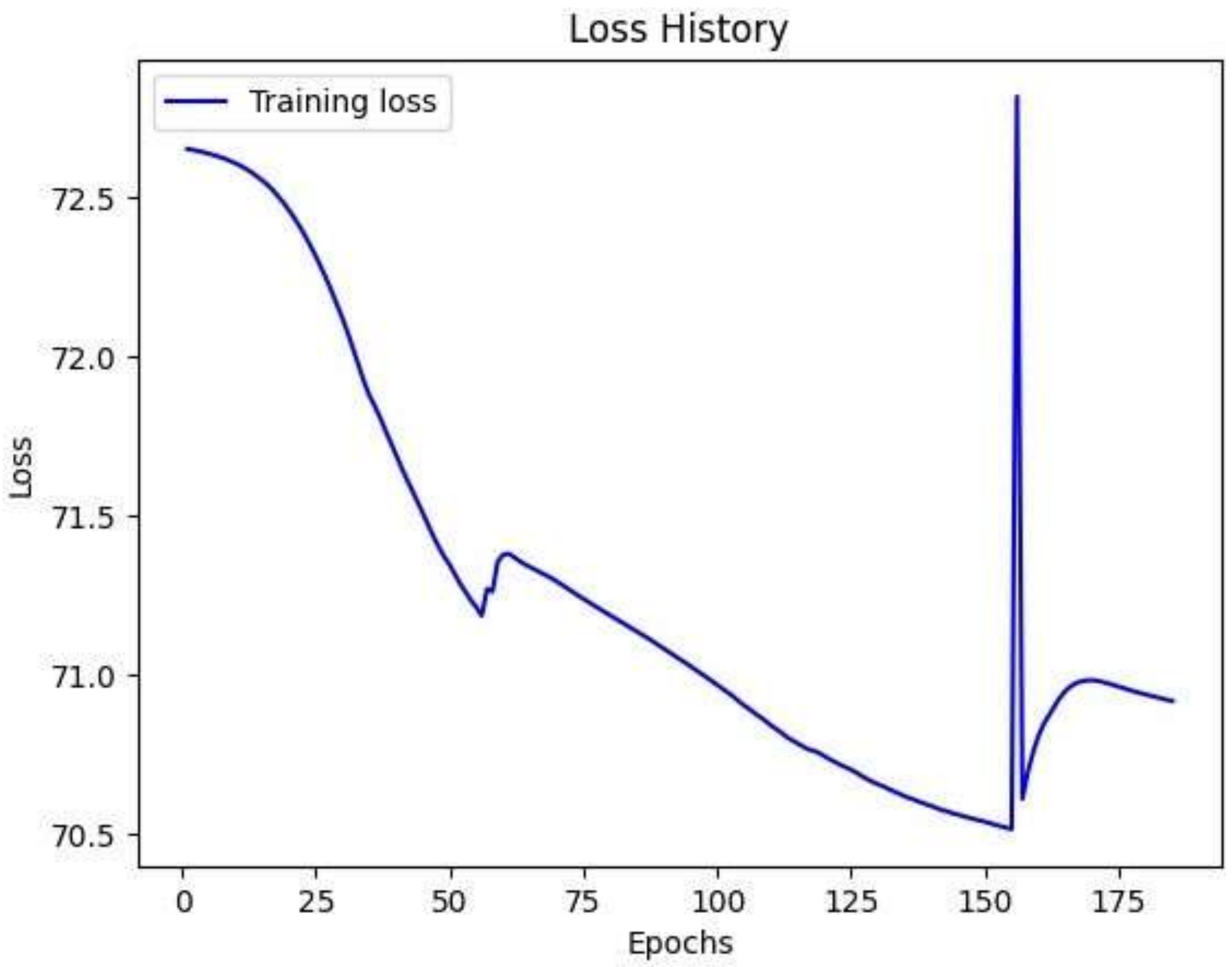
Attention Mechanism

The custom transformer model incorporates an attention mechanism that aligns the input text with the generated mel-spectrograms. This allows the model to attend to relevant parts of the input text during the synthesis process, ensuring accurate alignment between the text and the synthesized speech.

Training for Alignment

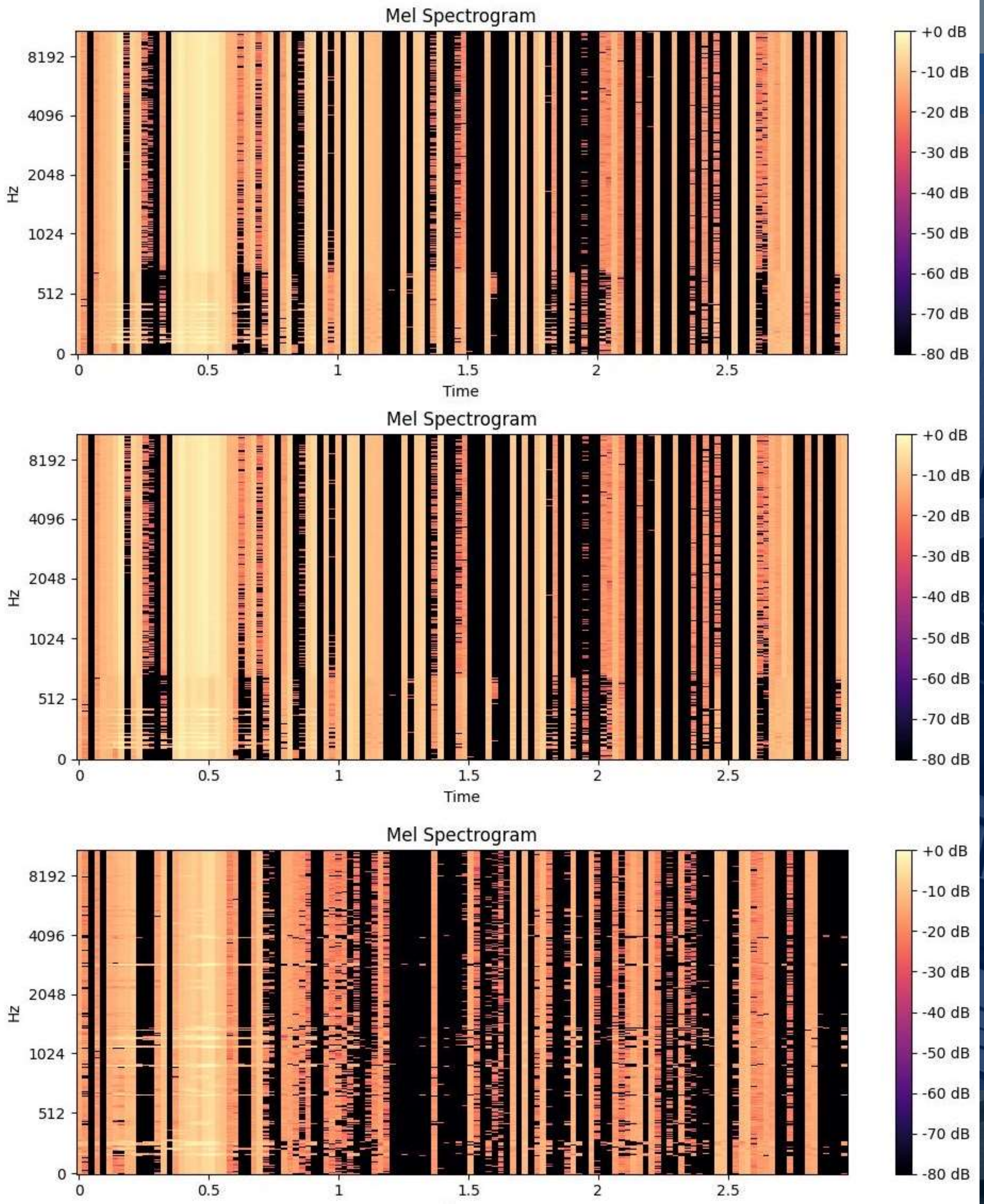
During training, the custom transformer model learns to align the input text with the target mel-spectrograms using supervised learning. The model's parameters are optimized, enabling the generation of high-quality speech that closely matches the desired acoustic features.

Model Training



Custom transformer Outputs

After training on the LibriSpeech dataset our transformer was giving us these spectrogram outputs



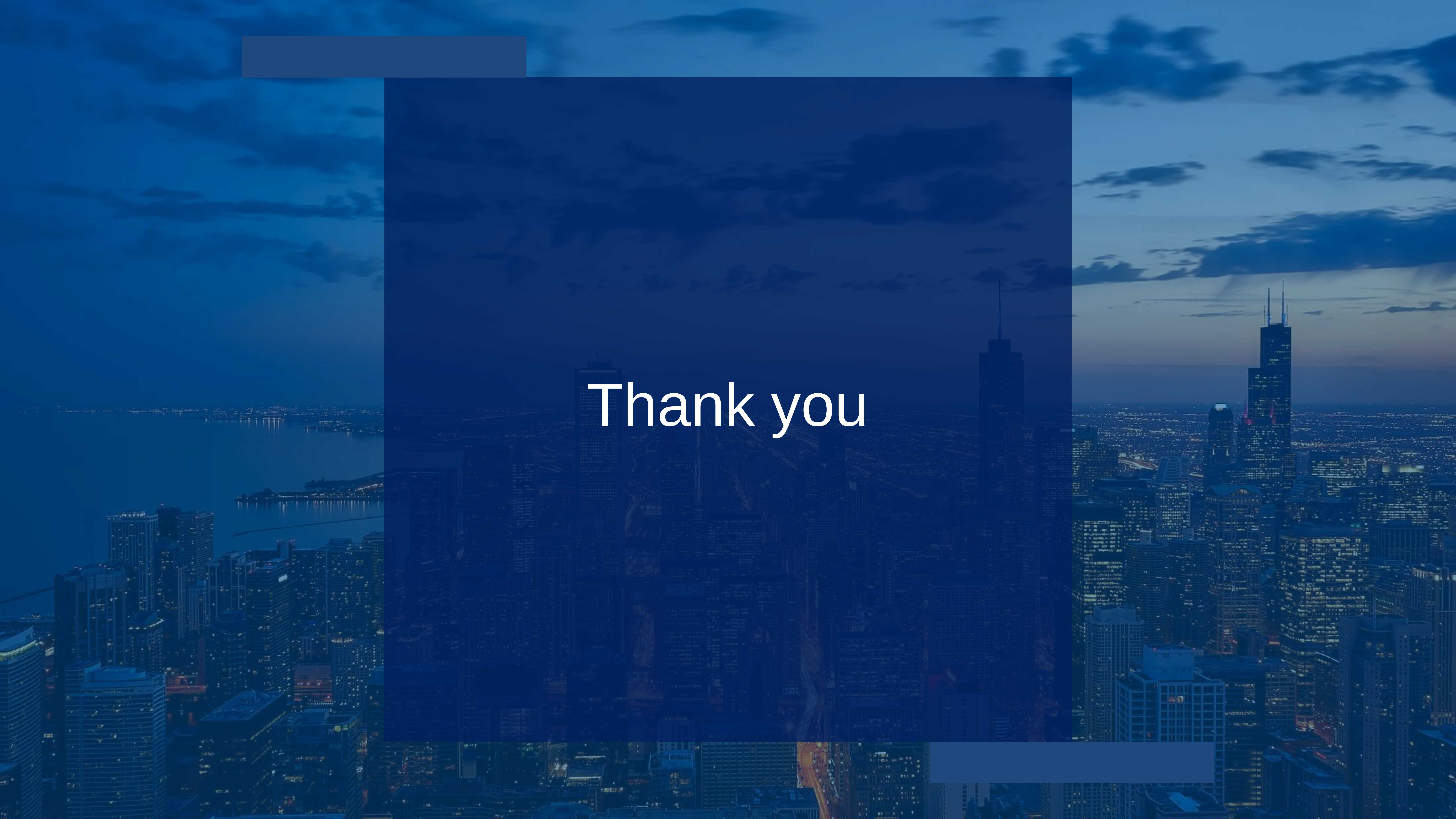
SOTA Models

Spectogram Generation Models

1. Tacotron 2: Tacotron 2 is an advanced text-to-speech model that can generate mel-spectrograms.
2. FastSpeech: FastSpeech is a fast and efficient text-to-speech model that can generate mel-spectrograms quickly. It utilizes a non-autoregressive framework and parallelizes the generation process.
3. Transformer TTS: Transformer TTS is a text-to-speech model based on the Transformer architecture. It offers flexibility and achieves excellent performance in generating mel-spectrograms.

Vocoder Models

1. WaveRNN: WaveRNN is a recurrent neural network-based vocoder model known for its ability to generate high-fidelity audio waveforms. It operates directly on the raw waveform and can be conditioned on mel-spectrograms.
2. Parallel WaveGAN: Parallel WaveGAN is a GAN-based vocoder model that can synthesize high-quality audio waveforms. It utilizes a multi-resolution structure and parallelization techniques for efficient and effective waveform generation.
3. MelGAN: MelGAN is a generative adversarial network-based vocoder model specifically designed for mel-spectrogram inversion. It can generate high-quality audio waveforms from mel-spectrograms.


Thank you