Improving Speech Naturalness and Nuance using HiFiGAN-Hubert-Soft Vocoder: A Case Study of the Voicebox TTS model

Kulboboev Shukhrat , Mohammed Salah Al-Radhi

shukhrat.kulboboev@edu.bme.hu

SmartLab
Intelligent Interactions

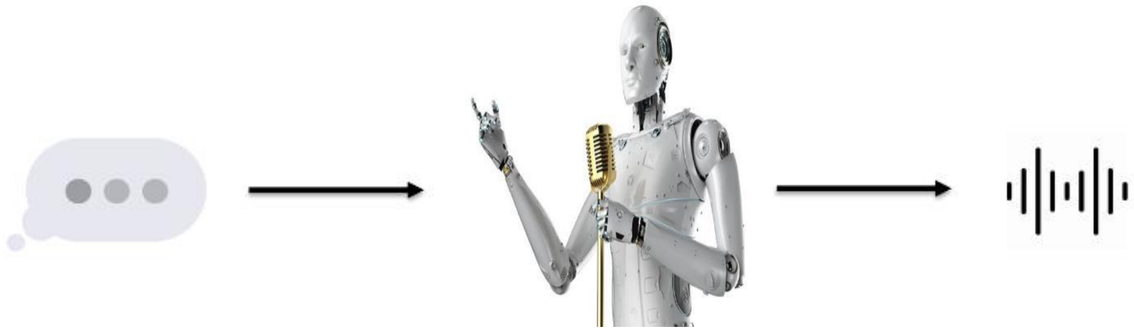
Contents

- Introduction
- Methodology
- Experimental Evaluation and Results
- Conclusion and Future Work

Introduction

Introduction

- Virtual assistant with human-like voice closer to reality
- Text-to-Speech (TTS) technology advancing, but some challenges persist
- Voicebox, a popular TTS model, offers flexibility but has limitations
- Voice may sound mechanical and lack emotional nuance
- Solution: HiFiGAN Hubert Soft Vocoder improves naturalness and expressiveness of generated speech



Voicebox TTS model

- Voicebox is a multilingual, non-autoregressive TTS model that generates natural-sounding speech. It will be used to easily edit audio tracks and allow visually impaired people to hear written messages from friends in their voices and enable people to speak any foreign language in their own voice.

Model	ZS TTS	Denoise	Partial Edit	Sampling
VALL-E	✓	✗	✗	✓
YourTTS	✓	✗	✗	✓
A3T	✓	✓(short)	✓	✗
Demucs	✗	✓	✗	✗
Voicebox	✓	✓(short)	✓	✓

Problems and Targets

Problems

- Traditional TTS systems face inherent limitations in achieving genuine human-like speech.
- Existing TTS systems often produce voices that sound robotic and lack natural expressiveness.

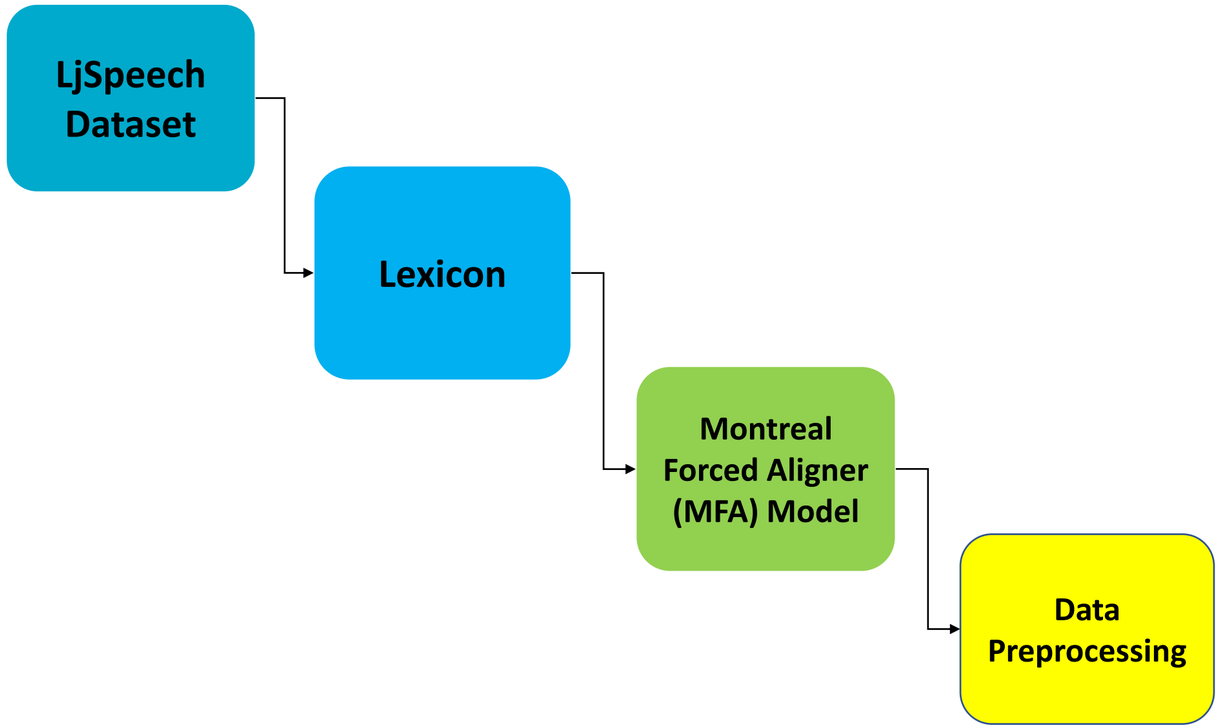
Targets

- Train Voicebox from scratch and implement it in practices
- Implement the HiFiGAN-Hubert-Soft vocoder and HiFiGAN-Hubert-Discrete vocoder to address robotic and inexpressive qualities.
- Choose which vocoder well done in real cases

PROPOSED METHODOLOGY

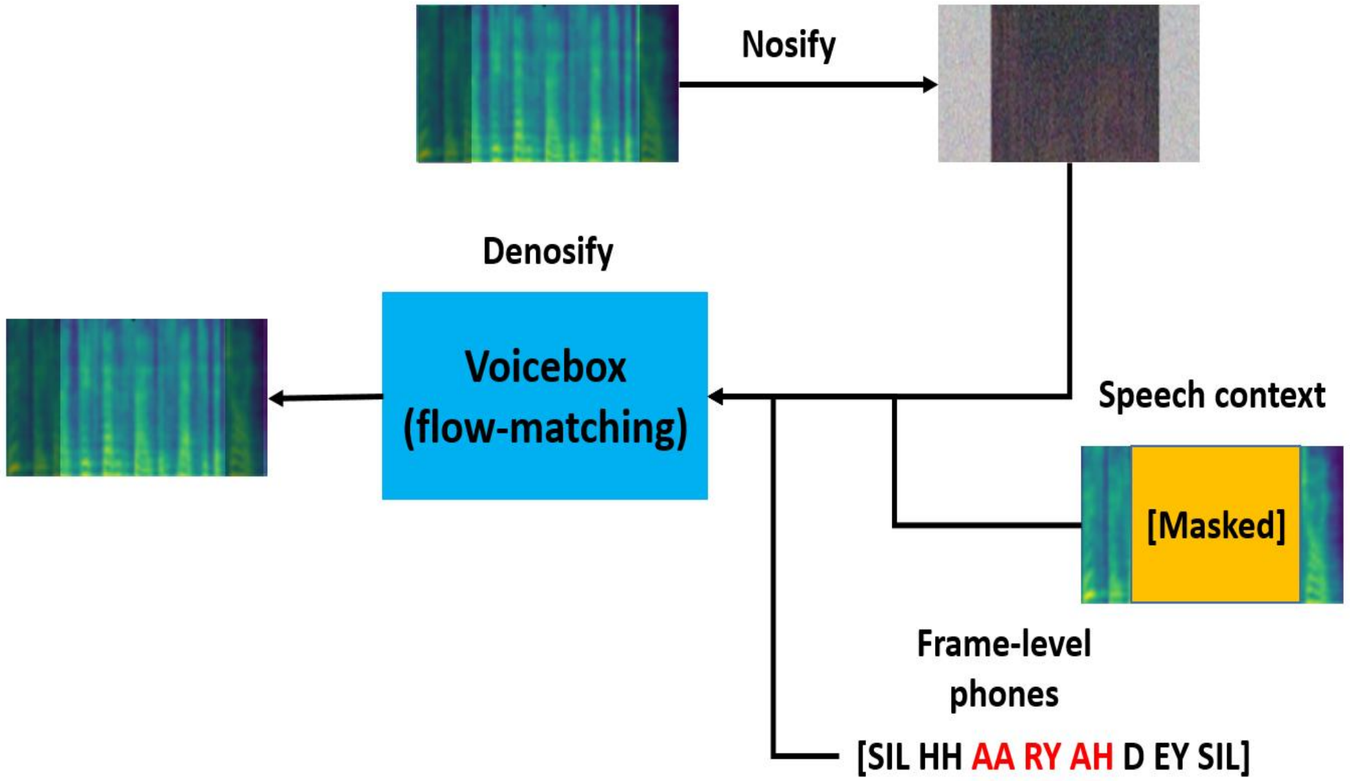
Data Preparation

The data preparation phase stands as the cornerstone of our research, comprising the deliberate and systematic handling of essential datasets, namely the LJSpeech dataset, lexicon, and Montreal Forced Aligner (MFA) model. LjSpeech dataset consists of 13,100 short audio clips of a single female speaker reading passages from 7 non-fiction books.



Workflow of Model

From a noisy distribution (top right in the image), it progressively builds the masked part of the sequence or the whole sequence by denoising it using the provided text and the unmasked parts of the audio sequence.



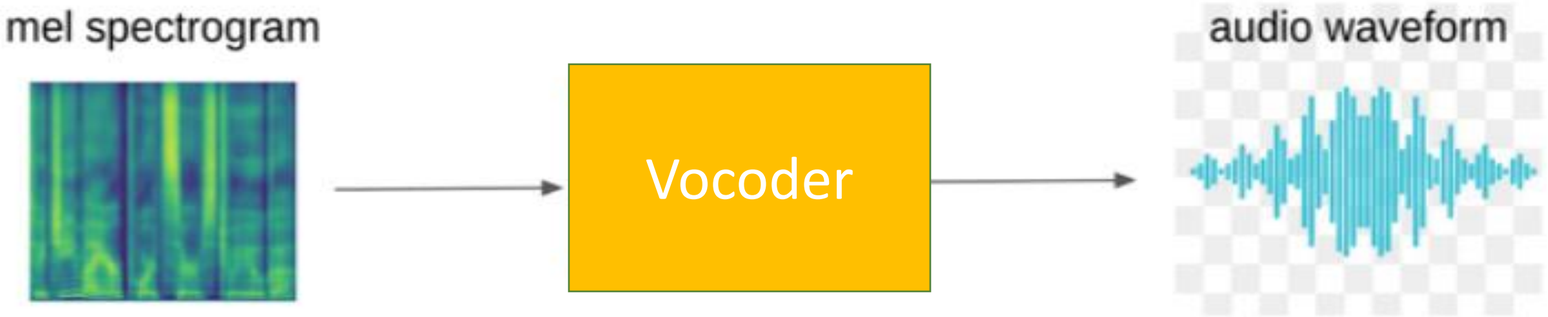
Proposed Methodology

Hifigan HuBERT Discrete

- Integer IDs
- Less expressive
- Lower quality
- Faster and easier

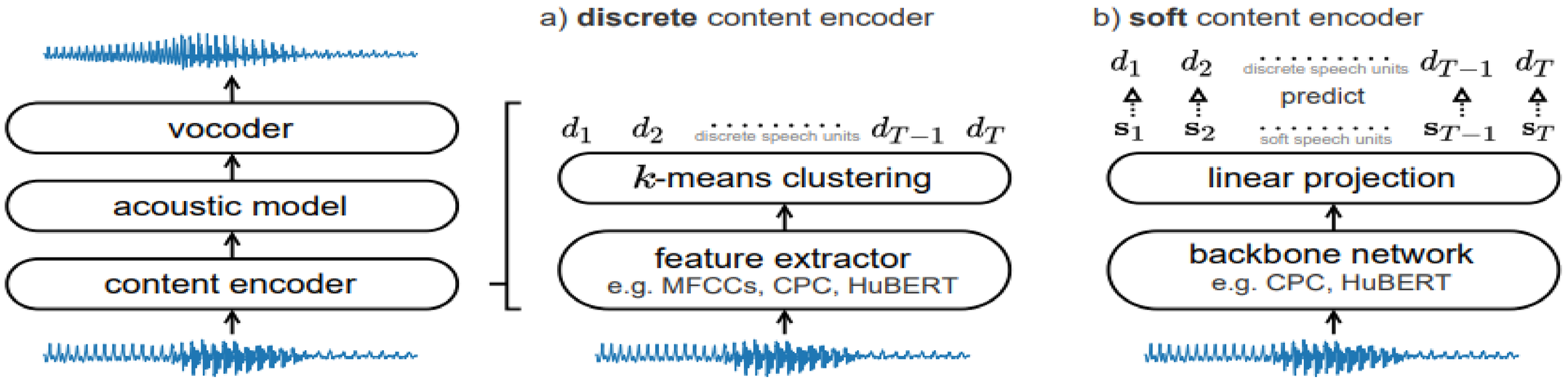
Hifigan HuBERT Soft

- Probability distributions over discrete speech units
- More expressive
- Higher quality
- Slower and more difficult



Proposed Methodology

a) The discrete content encoder clusters audio features to produce a sequence of discrete speech units.
b) The soft content encoder is trained to predict the discrete units. The acoustic model transforms the discrete/soft speech units into a target spectrogram. The vocoder converts the spectrogram into an audio waveform.



EXPERIMENTAL SETTING AND EVALUATION

Setup

Training Schedule

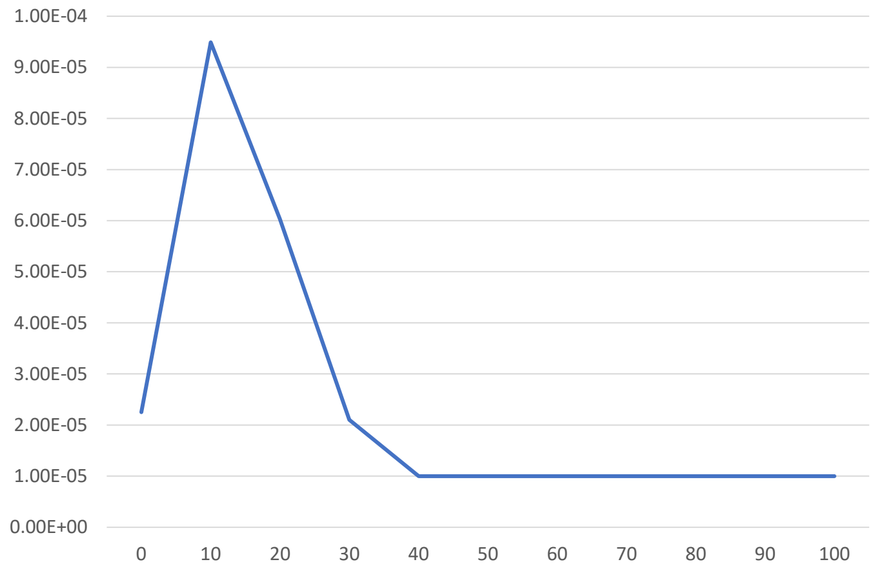
- Model was trained during about 48 hours without early stopping over 100 epochs
- NVIDIA TITAN Xp GPU and CUDA Version: 11.4 is used for training

Hyperparameters and Optimization

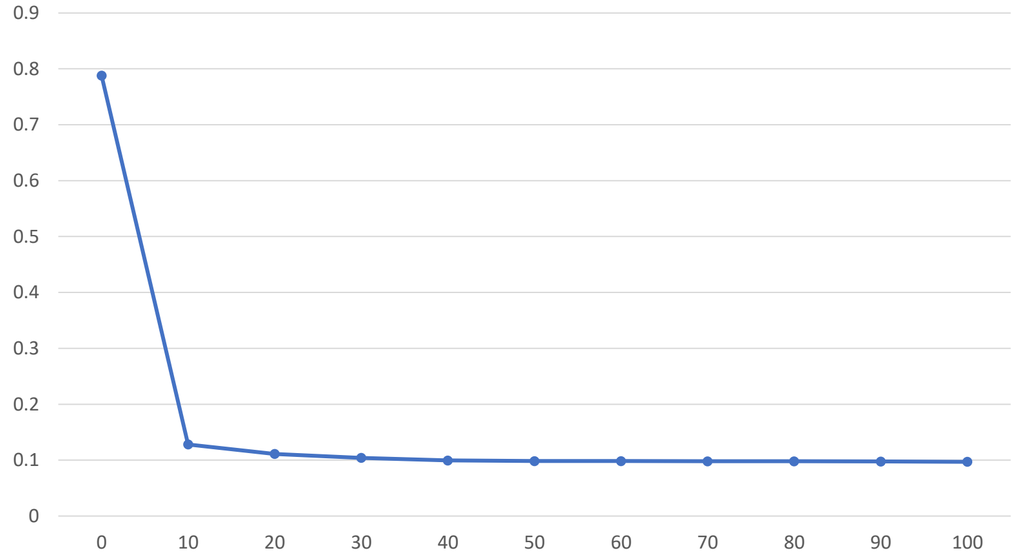
- started the learning rate with 1.18e-5 and epoch loss 0.788 after the first epoch training
- Adam optimizer was used with a peak learning rate of le-4
- Trained with mini batch size of 12 because the server could manage train with 16 batch size
- Audio content was represented as an 80-dimensional log Mel spectrogram

Change of Parameters

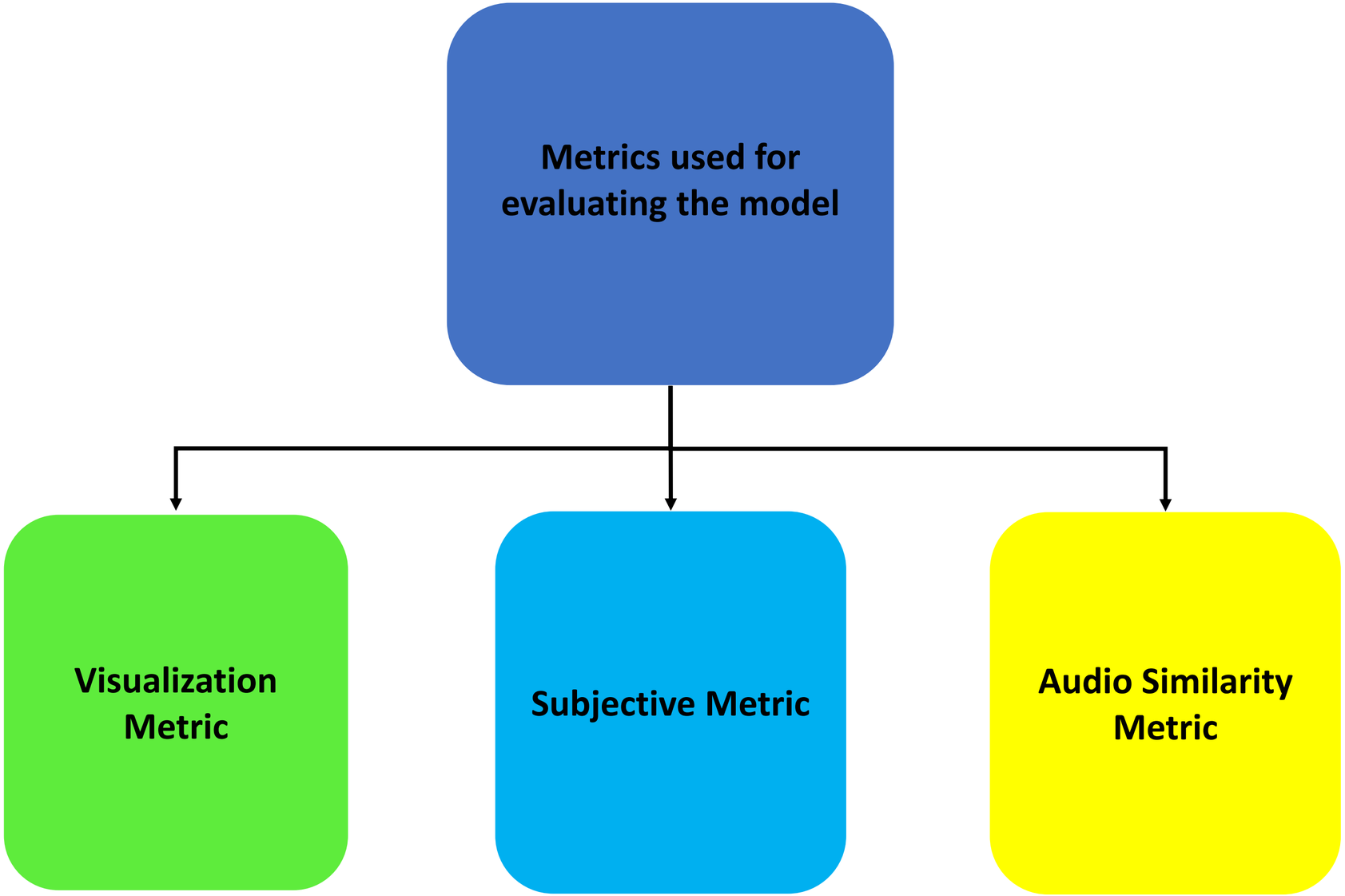
Learning rate changes



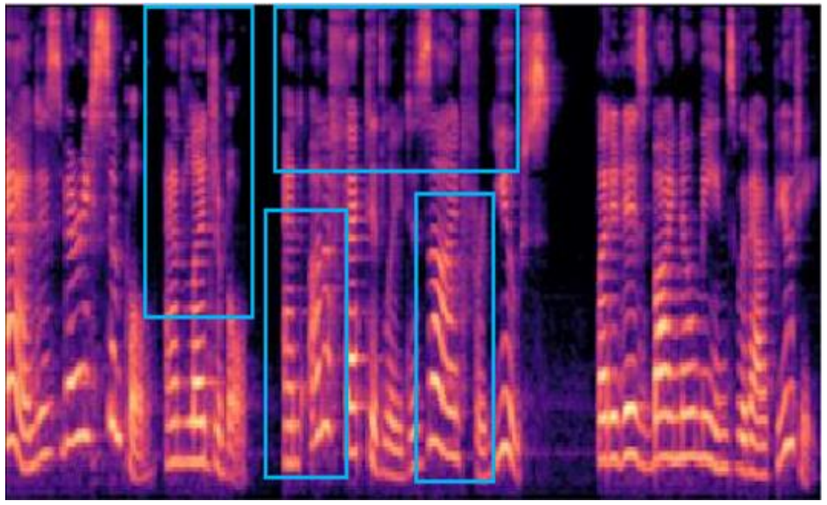
Loss in epochs



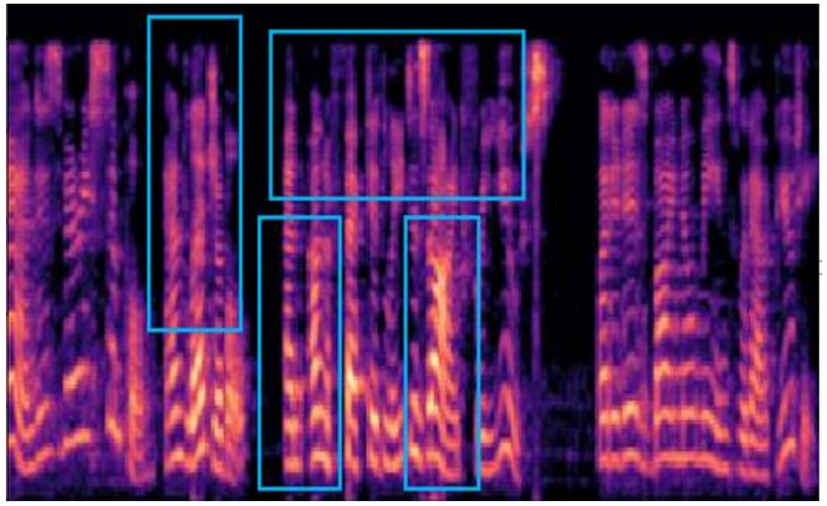
Evaluation Metric



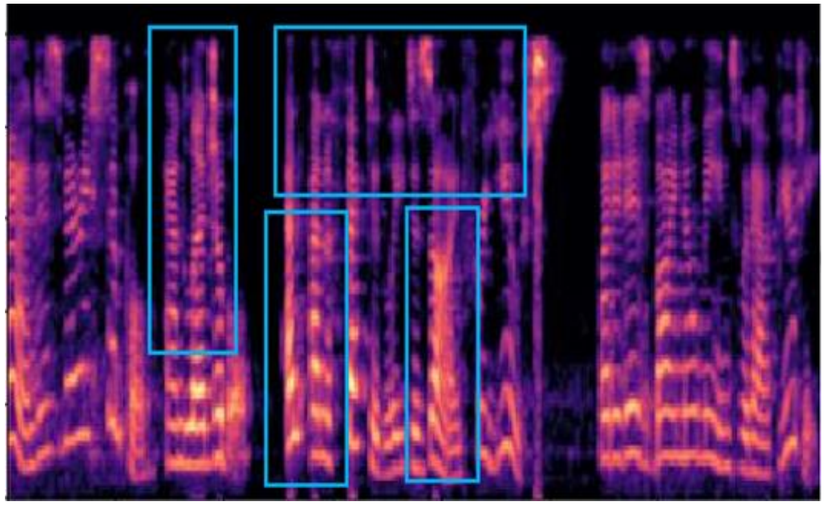
Visualization Metric



Original Voice

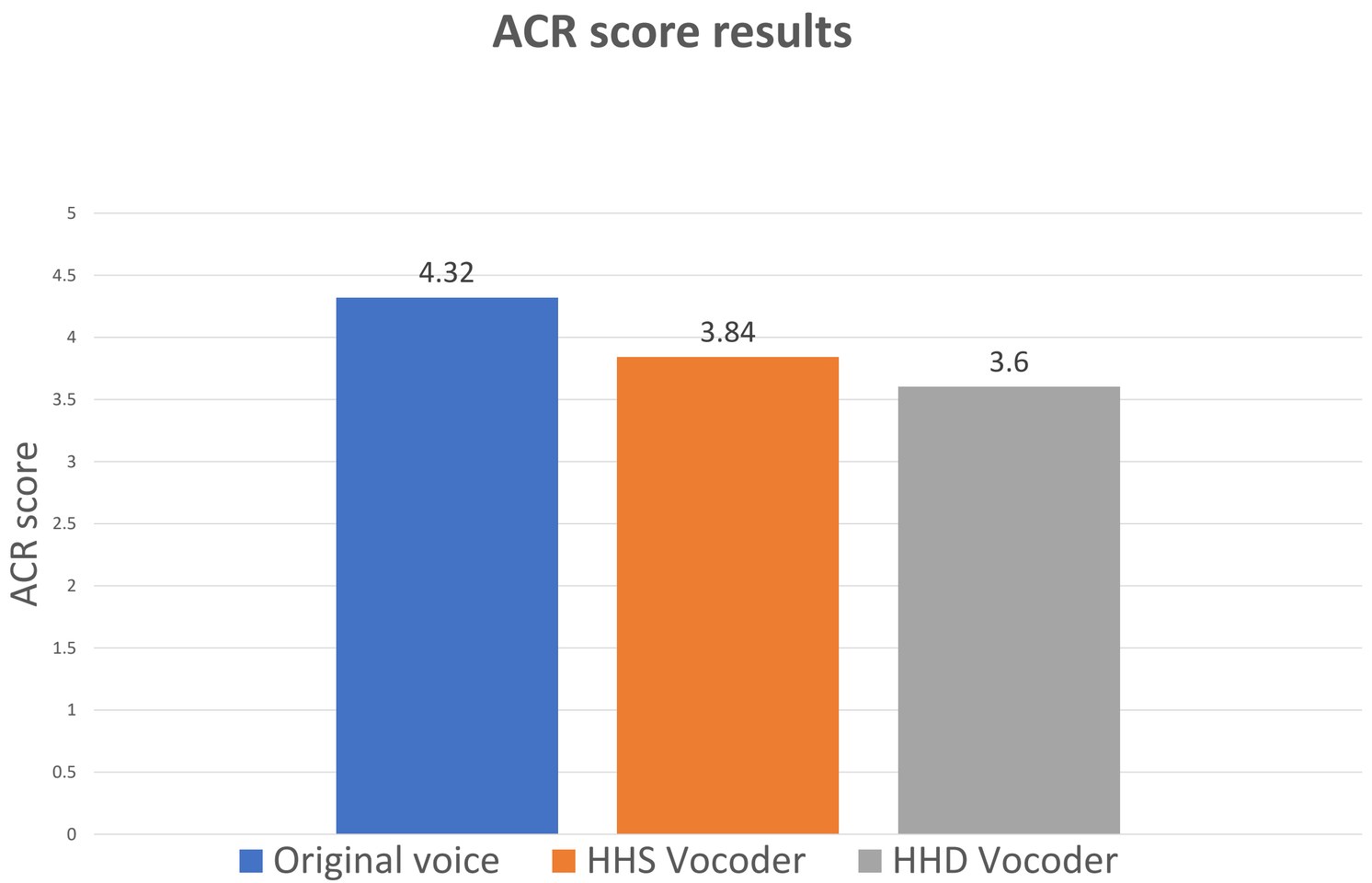


HHS Vocoder

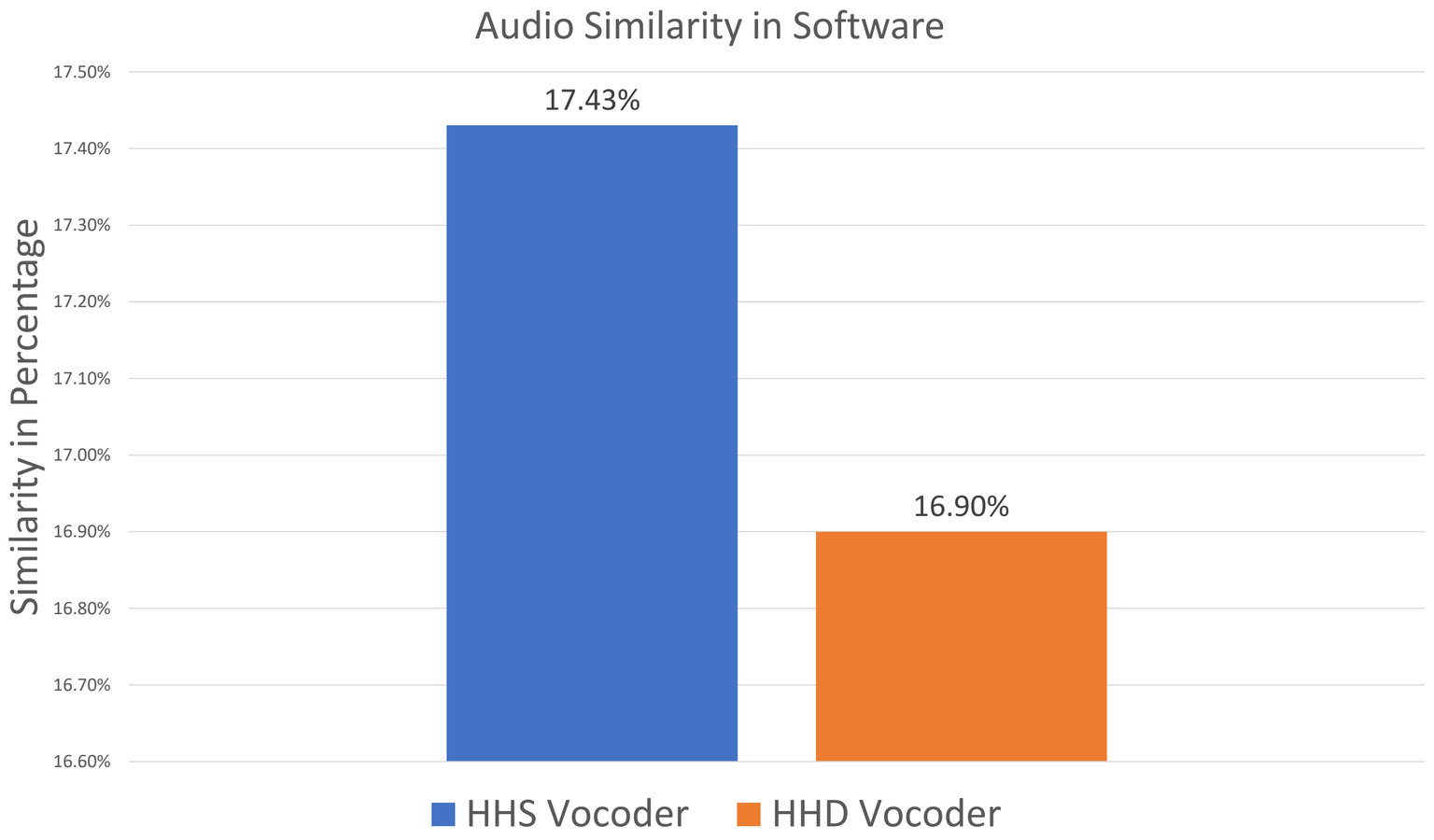


HHD Vocoder

Subjective Metric



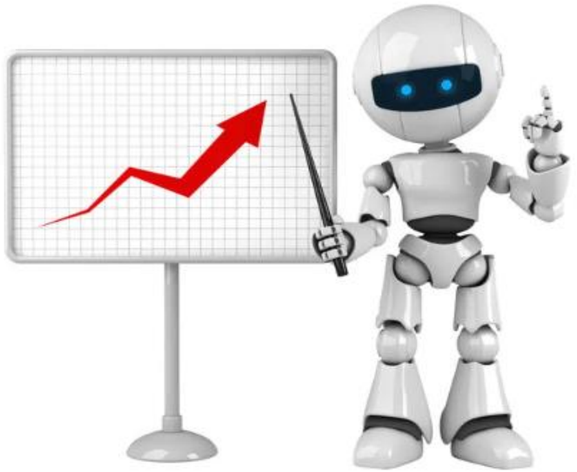
Audio Similarity Metric



Conclusion and Future Work

Conclusion

- Successfully enhanced the Voicebox TTS model by replacing the HHD vocoder with HHS, resulting in superior speech precision, particularly in nuanced scenarios.
- Hold substantial implications for advancing high-quality TTS systems and promises more realistic and engaging human-machine speech interactions.



Future Work

- Evaluating the impact of Universal Vocoder on speech naturalness and nuance
- Training and adding duration model on Voicebox
- Training Voicebox model on another datasets

Thank you for your attention

shukhrat.kulboboev@edu.bme.hu