Final Project

Topic Voice Synthesis: Generating Human Like Speech from Text

Okoronkwo, Alfred

CSCI E-89 Deep Learning, Fall 2025

Harvard University Extension School

Prof. Zoran B. Djordjević & prof. Rahul Joglekar

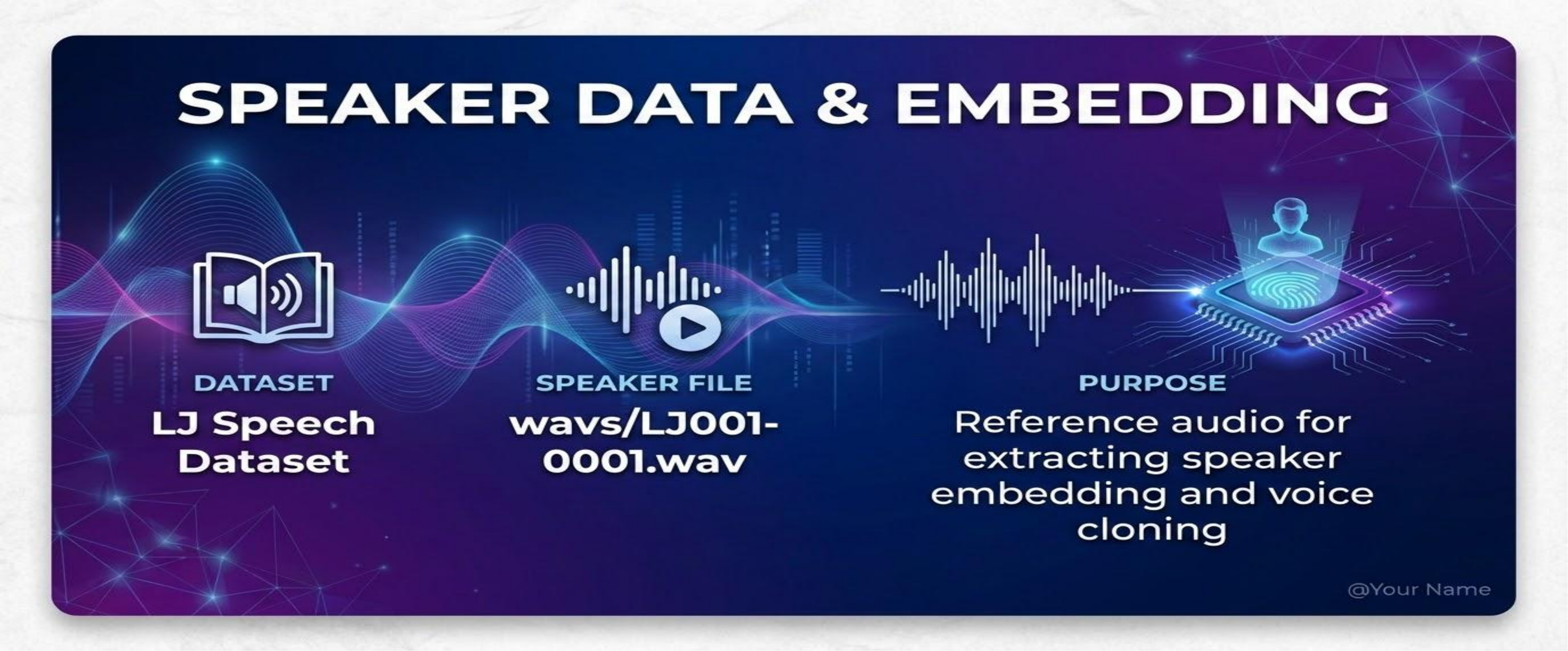
Introduction

Goal:

develop a system that can clone a target speaker's voice and generate speech from arbitrary text in the speaker's voice.

Model:

the project uses the tts_models/multilingual/multi-dataset/your_tts pretrained model



SPEAKER DATA & EMBEDDING

DATASET

LJ Speech Dataset

SPEAKER FILE

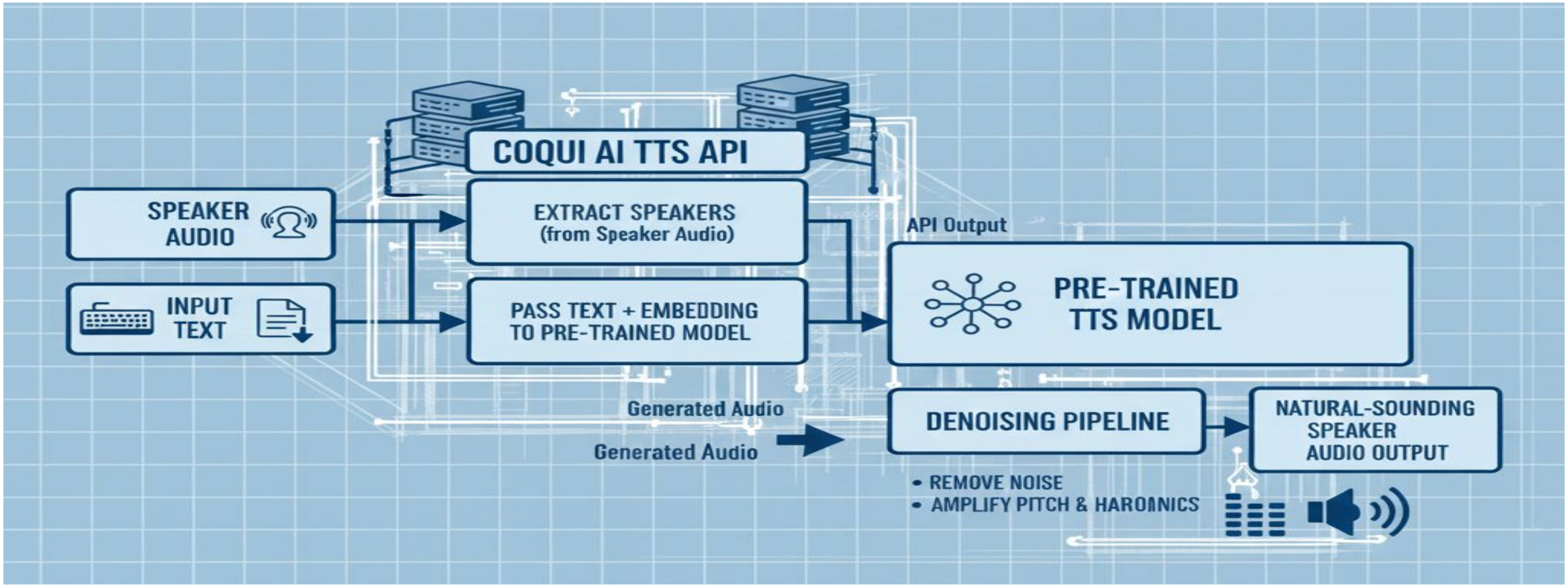
wavs/LJ001-0001.wav

PURPOSE

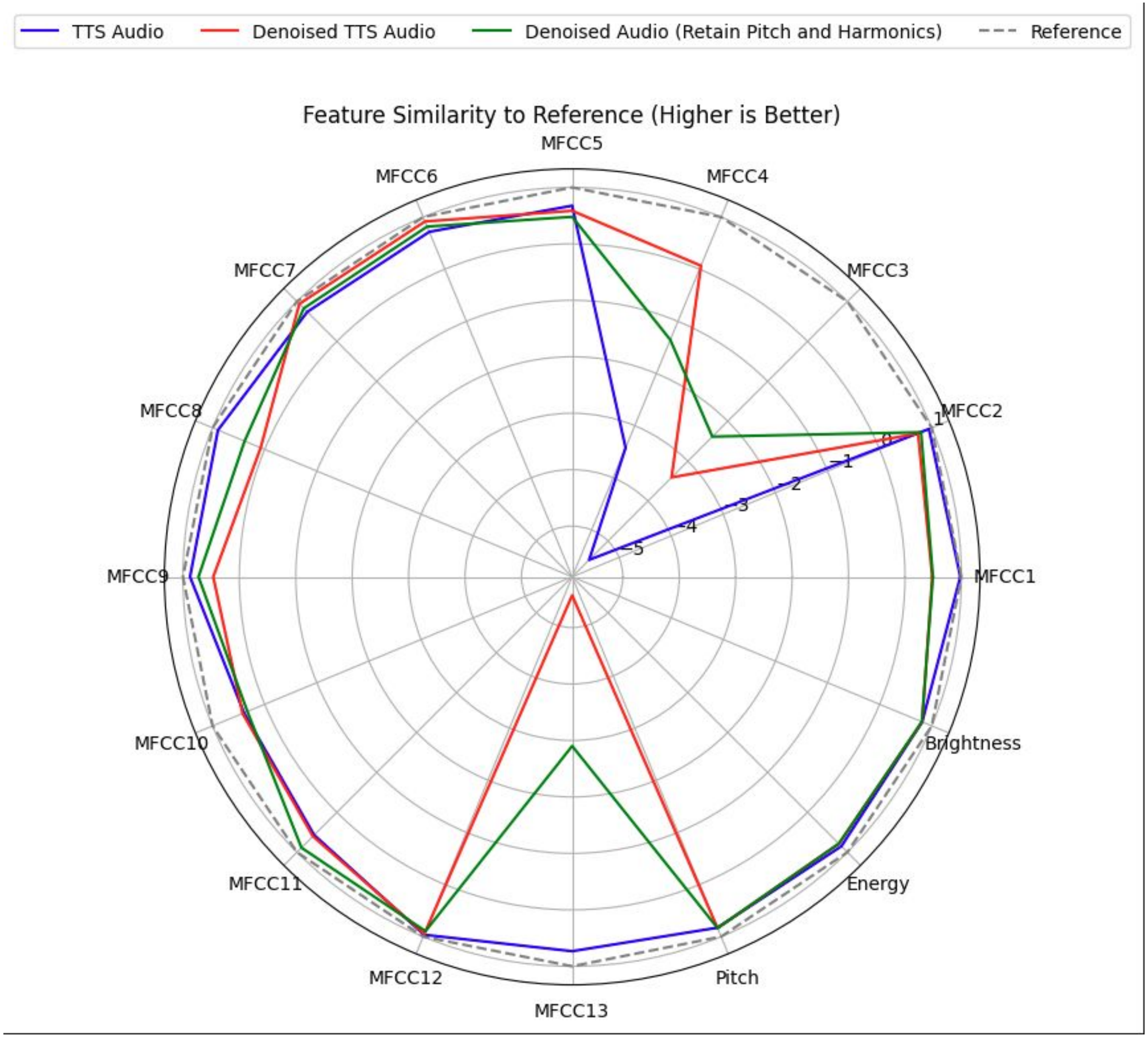
Reference audio for extracting speaker embedding and voice cloning

@Your Name

Zero-Shot Voice Cloning


COQUI AI TTS API
SPEAKER AUDIO
INPUT TEXT
EXTRACT SPEAKERS (from Speaker Audio)
PASS TEXT + EMBEDDING TO PRE-TRAINED MODEL
API Output
PRE-TRAINED TTS MODEL
Generated Audio
Generated Audio
DENOISING PIPELINE
NATURAL-SOUNDING SPEAKER AUDIO OUTPUT
• REMOVE NOISE
• AMPLIFY PITCH & HAROINNICS

Audio Similarity Comparison



Conclusion

By preserving the speaker's harmonics and pitch, we were able to improve the TTS-generate audio. Listening to the audio confirms that the resulting speech is much closer to the target speaker, sounding significantly better than both the original TTS-generated audio and the initially denoised TTS output.

YouTube URLs, Last Page

- 2 minute (short): https://youtu.be/7qUkhFdp_Js
- 15 minutes (long):