Detecting AI-Generated Speech Using Supervised Deep Learning

Steven Chen (SC311), Patrick Chiu (PC82), YungHsin Tsai (YT66), Jitao Huang (JH322)

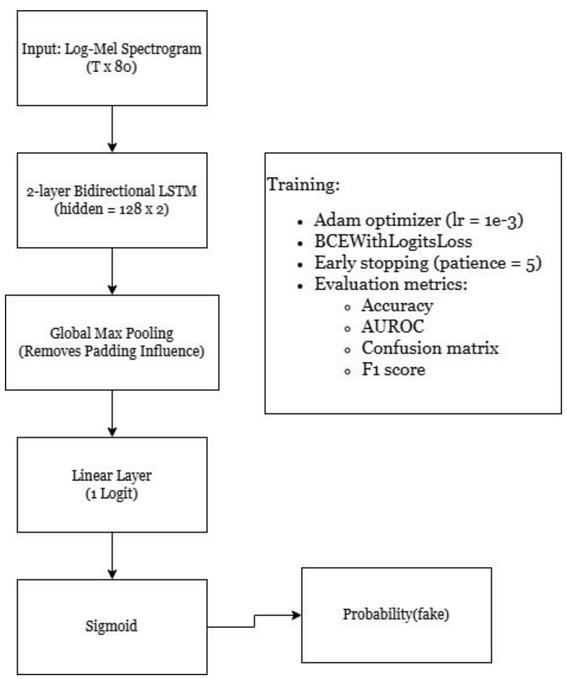
Abstract

We develop a lightweight deep learning model to distinguish real speech from AI-generated audio. Using a custom dataset of human recordings and FastSpeech + Flowavenet outputs, our bidirectional LSTM achieves 92% accuracy and strong class separation in ROC and t-SNE analyses. The model performs competitively with larger baselines while offering faster inference and smaller size.

Background

Modern text-to-speech systems can generate highly realistic audio, making synthetic speech increasingly difficult to distinguish from real human recordings. This poses risks for impersonation, fraud, and misinformation. To address this, we develop a supervised deep learning model that detects whether an audio clip is real or AI-generated.

Experimentation and Model



METHODS

Dataset Construction

- Built a custom two-class dataset of **real** and **FastSpeech + Flowavenet-generated** speech
- Ensures the model learns **acoustic** differences rather than textual cues

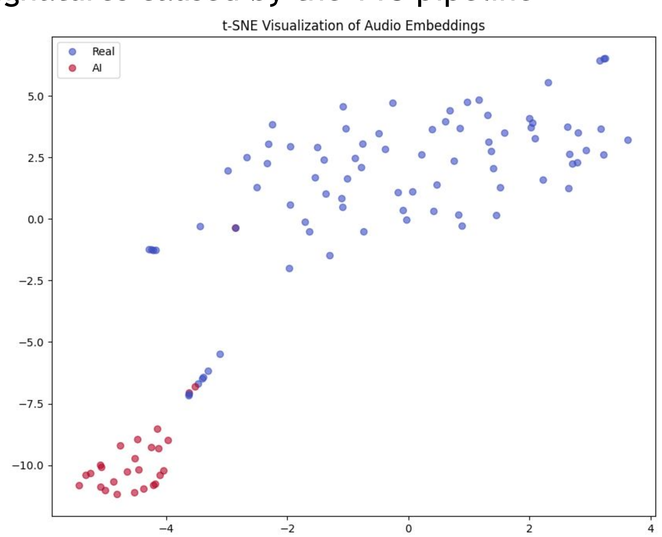
Preprocessing

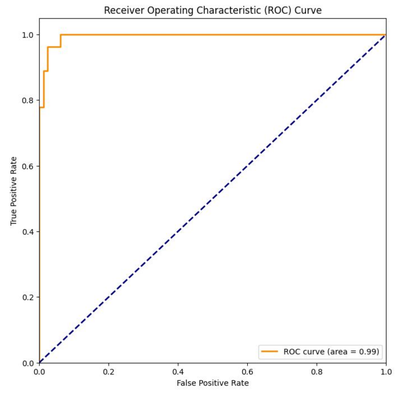
- Load audio at 16 kHz using librosa
- Convert waveform → **log-Mel spectrograms**
- Normalize each spectrogram
- Zero-pad sequences with PyTorch pad_sequence for batching
- Return (mel, label) pairs via a PyTorch Dataset

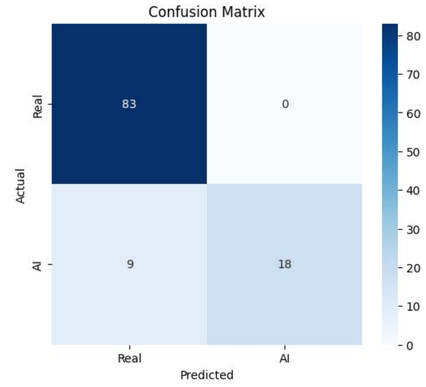
RESULTS

Model Performance

- The bidirectional LSTM effectively separates real vs. FastSpeech+Flowavenet speech
- The ROC curve shows strong class discriminability
- t-SNE reveals **two well-defined clusters**, indicating that the model learns acoustic signatures caused by the TTS pipeline







Accuracy	0.92
F1 Score	0.80
Equal Error Rate	0.037

Conclusion

Comparison With Other Models (Using 3rd Dataset)

Model	Our Model	AASIST
Accuracy	0.92	0.98
F1 Score	0.80	0.9474
Inference Time	685.78 ms	3351.30 ms
Model Size	2.33 MB	28.09 MB

ACKNOWLEDGEMENT/ BIBLIOGRAPHY

- Azis, Huzain; Rismayanti, Nurul; Binti Abdullah, Munaisyah; Binti Ismail, Suriana (2025), "Speech Dataset of Human and AI-Generated Voices", Mendeley Data, V2, doi: 10.17632/5czyx2vppv.2
- FastSpeech : Fast, Robust and Controllable Text to Speech https://doi.org/10.48550/arXiv.1905.09263
- FloWaveNet : A Generative Flow for Raw Audio https://doi.org/10.48550/arXiv.1811.02155
- Combination of FastSpeech and FloWaveNet https://github.com/cjlin8787/FastSpeech-FloWaveNet?tab=readme-ov-file
- dataset https://github.com/WaliMuhammadAhmad/AIVoice-Detection
- AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks https://doi.org/10.48550/arXiv.2110.01200