# Audio Deepfake Detection using Reflection Coefficient Vocal Tract Modeling

Jasper Wilkerson

## Goals

The goal of this project is to propose a system for audio deepfake detection, in addition to:

- Investigating a novel input feature
- Demonstrating effectiveness across datasets
- Performing an ablation study to determine specific discriminative features common across deepfakes

## Datasets

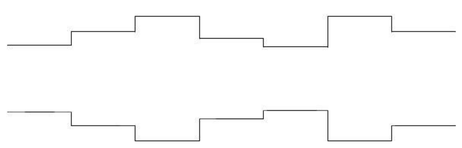
Two Datasets were used, one historical and one current, covering TTS, voice conversion, and codec generation techniques

- ASVSpoof2015
  - ~300,000 samples covering 10 methods of audio generation
- CodecFake (2025)
  - ~1,000,000 samples covering 6 different methods of transformer-based audio generation

## LPC Reflection Coefficients

Instead of using traditional MFCC's, this project used 16 order LPC Reflection Coefficients which can approximate the vocal tract.

At each 10ms segment of audio, we calculate a lossless 16 tube model



## Model Architecture

Three Bi-LSTM models were trained on the ASVSpoof dataset, CodecFake dataset, and a combined dataset

Weighted sampling was performed to account for the unequal distribution, but the data was otherwise unchanged

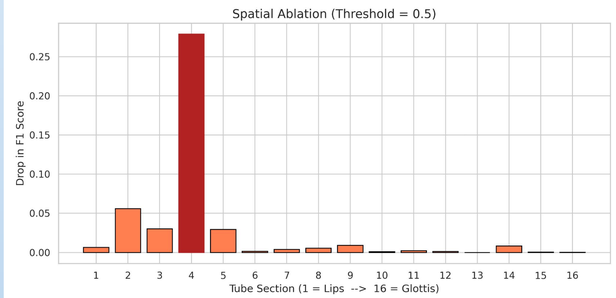



Figure 1: Ablation study on the ASVSpoof dataset

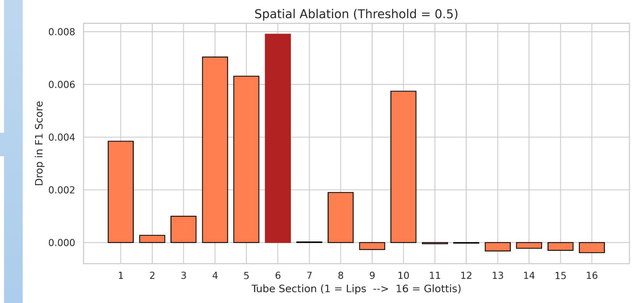



Figure 2: Ablation study on the CodecFake dataset

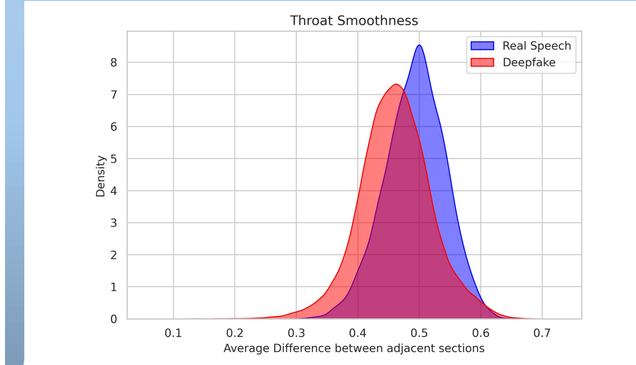



Figure 3: Density chart of average difference between sections

## Analysis

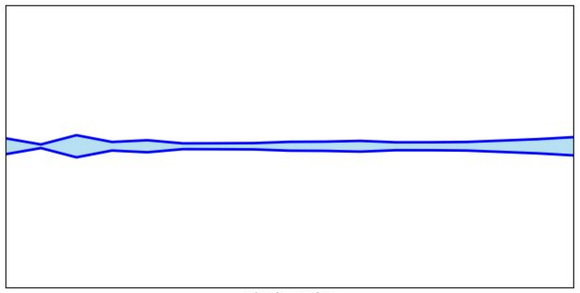



Figure 4: Vocal Tract reconstruction of deepfake audio

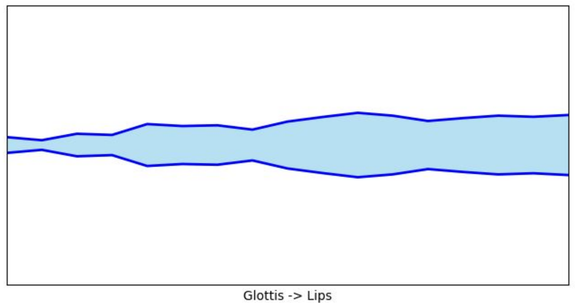



Figure 5: Vocal Tract reconstruction of real audio

To investigate which anatomical features are most discriminative, the evaluation was run with individual tube sections masked *(Figures 1 and 2)*. For both models, Section 4 (the hard palate/tongue blade) caused the largest drop in accuracy when masked. This reveals that synthetic algorithms fail primarily at modeling the complex, turbulent constrictions in the front of the mouth.

When calculating the spacial jaggedness (the average $\Delta k$ difference between tube sections), we found that deepfakes generate vocal tracts that are mathematically smoother than genuine audio *(Figure 3)*

## Results

| | ASVSpoof Model | CodecFake Model | Combined Model |
|---|---|---|---|
| F1 | 0.9938 | 0.9979 | 0.9973 |
| EER | 0.0107 | 0.0065 | 0.0142 |

All three LSTM models achieved near-perfect detection on both targeted datasets, proving the usefulness of biologically based features in this task.

- Both standalone models successfully detected almost all generated audio within their specific datasets, yielding Macro F1 scores >0.99.
- When trained on a combined dataset (Vocoders + Codecs), the model was able to generalize, despite major differences in recording and sampling. This demonstrates that compression artifacts do not overwrite the underlying physical anomalies left by deepfake synthesis

## Conclusion

LPC Reflection coefficients were shown to be:

- Novel, biologically based audio input features
- Effective in distinguishing deepfakes even with a simple model
- Useful for all tested types of audio generation and >0.96 Macro F1 for all 16 generation methods
- More interpretable than other end to end techniques and faster than other linguistically-based techniques