SonicShield

AI-Powered Guardian Against DeepFake Speech

Team AudioSentinels

Chandranath Bhattacharya
MDS202318

Salokya Deb
MDS202341

Soumyajoy Kundu
MDS202349

MSc. Data Science

Applied Machine Learning

May 07, 2025

CHENNAI MATHEMATICAL INSTITUTE

Outline

Problem Statement

AI-Powered Guardian Against DeepFake Speech

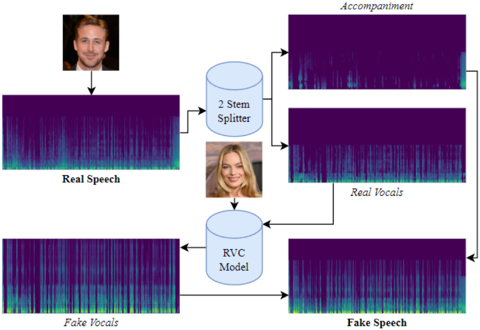
To address the growing misuse of generative AI for real-time voice cloning and DeepFake attacks by developing a machine learning model capable of detecting AI-generated speech in real time.

Goal!!!

- Ensure reliable, low-latency identification of synthetic audio for enhanced privacy and security.

Data – Kaggle

This dataset contains examples of real human speech, and DeepFake versions of those speeches by using Retrieval-based Voice Conversion.



REAL

1. Biden
2. Margot
3. Ryan
4. Musk
5. Obama
6. Taylor
7. Trump
8. Linus

FAKE

- $8 * 7 = 56$ fake voice conversions

All audio files are of duration max 10 mins with a size ~ 100 MB

Workflow

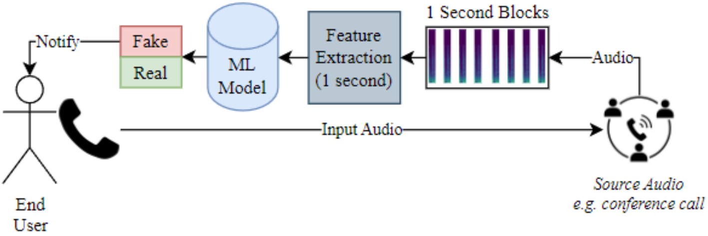
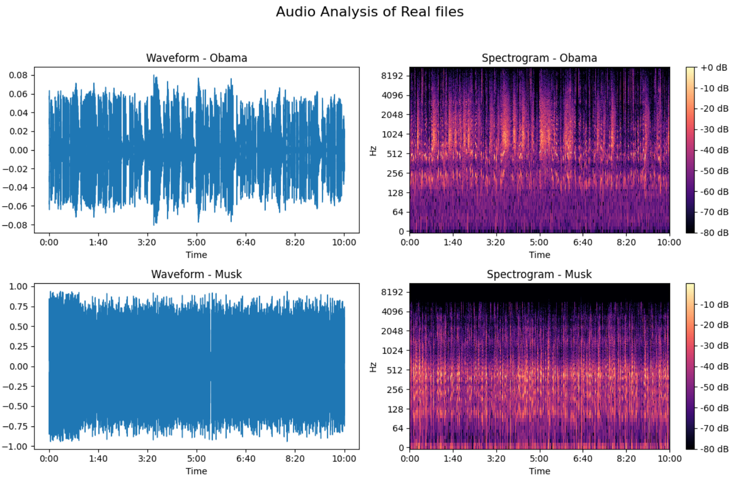


Figure: Potential use of a successful system

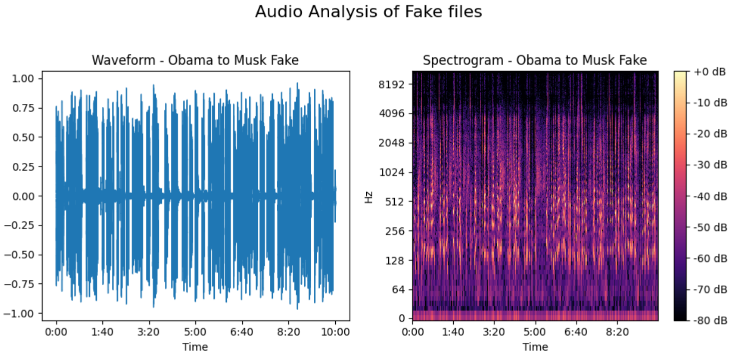
Data Visualisation

Waveform & Spectrogram



Data Visualisation

Waveform & Spectrogram



Data Preprocessing

Feature Extraction

1. **Audio Loading**
   - Load each ‘.wav‘ file using `librosa` at sampling rate of 22050 Hz.
2. **Windowing**
   - Divide each audio file into 1-second non-overlapping segments.
3. **Feature Extraction:** For each 1-second segment:
   - Chromagram
   - Root Mean Square Energy
   - Spectral Centroid
   - Spectral Bandwith Rolloff
   - Zero Crossing Rate (with silence check)
   - Mel-Frequency Cepstral Coefficients (20 coefficients)

Data Preprocessing

Feature Extraction

4. **Feature Aggregation**
   - Compute the mean of each feature over the window.
5. **Label Assignment**
   - Tag each window with its corresponding label (e.g., Real or Fake).
6. **Batch Processing**
   - Use `joblib.Parallel` for parallel processing across all audio files.
7. **Output**
   - A dataframe of randomly sampled files of size 11778 using stratification

Feature Analysis

Correlation & Significance (Unpaired t-test) Analysis

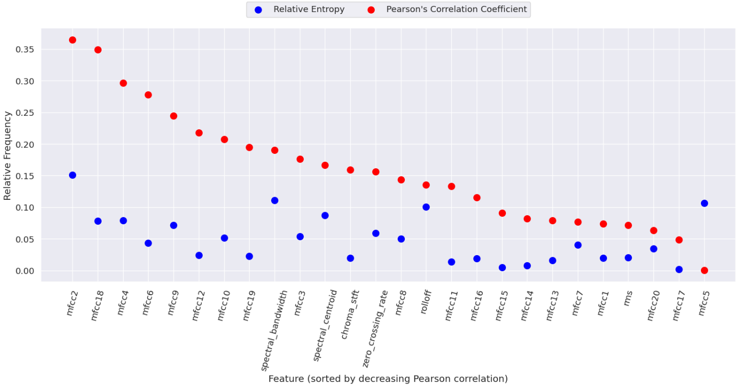


Figure: Pearson's Correlation Coefficient and Relative entropy for all of the extracted features when used for binary classification of real or AI-generated vocals

Feature Analysis

Correlation & Significance (Unpaired t-test) Analysis

Attribute	Signif?	Attribute	Signif?
Chromagram	✓	*MFCC 8*	✓
Root Mean Square	✓	*MFCC 9*	✓
Spectral Centroid	✓	*MFCC 10*	✓
Spectral Bandwidth	✓	*MFCC 11*	✓
Rolloff	✓	*MFCC 12*	✓
Zero Crossing Rate	✓	*MFCC 13*	✓
MFCC 1	✓	*MFCC 14*	✓
MFCC 2	✓	*MFCC 15*	✓
MFCC 3	✓	*MFCC 16*	✓
MFCC 4	✓	*MFCC 17*	✓
MFCC 5	✗	*MFCC 18*	✓
MFCC 6	✓	*MFCC 19*	✓
MFCC 7	✓	*MFCC 20*	✓

Model Building

Benchmark Models

1. **Data Preparation**
   - Feature matrix `X` from all columns except `LABEL`.
   - Encode `LABEL`: `REAL = 1`, `FAKE = 0`.
2. **Data Splitting**
   - Split dataset into Train, Validation, and Test sets using stratified sampling.
3. **Hyperparameter Tuning**
   - Perform grid search using `ParameterGrid` on training set.
   - Evaluate models on the validation set using accuracy or AUC.
4. **Learning Curve Analysis**
   - Plot training vs. validation accuracy.
   - Understand model behavior with increasing training size.
5. **Evaluation on Test Set**
   - *Metrics*: Accuracy, Classification Report, Confusion Matrix.
   - *Visualization*: ROC Curve and AUC score.

Model Building

Benchmark Models

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
XGBoost (200)	**0.993**	0.998	0.991	0.993	0.993
Random Forest (200)	**0.990**	0.995	0.983	0.989	0.989
Naïve Bayes (Gaussian)	0.830	0.864	0.784	0.822	0.830
Logistic Regression	0.820	0.884	0.882	0.883	0.883
Support Vector Machine (SVM)	0.723	0.815	0.576	0.675	0.723

Table: Comparison of Models – Mean over 5-fold CV

Model Building

Transfer Learning using DistilBERT

1. **Feature Extraction**
2. **Preprocessing :**
   - Normalize features with `StandardScaler`.
   - Encode labels (FAKE=0, REAL=1) using `LabelEncoder`.
3. **Model Architecture :**
   - Project 26-dim audio features to match DistilBERT hidden size.
   - Use frozen `distilbert-base-uncased` as a feature extractor.
   - Add classification head with fully connected layers.

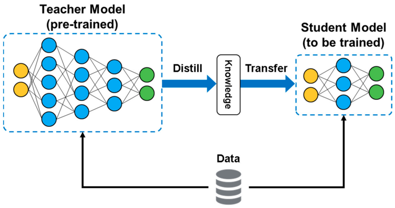


Figure: DistilBERT

Model Building

Transfer Learning using DistilBERT

4. **Training :**
   - Use HuggingFace `Trainer` with 5 epochs and stratified split.
   - Log performance via Weights Biases.
5. **Inference :** Classify new audio files by:
   - Extracting and scaling features,
   - Running inference through the trained model,
   - Decoding predicted label (FAKE/REAL).
6. **Persistence :**
   - Save model, scaler, and label encoder for future use.

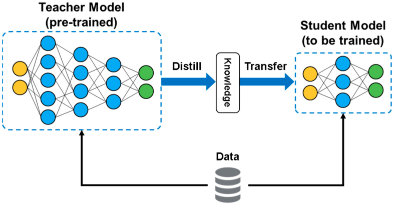


Figure: DistilBERT

Model Building

Transfer Learning using DistilBERT

Epoch	Training Loss	Validation Loss	Accuracy
1	0.472	0.391	0.846
2	0.359	0.215	0.924
3	0.295	0.163	0.941
4	0.271	0.144	0.949
5	**0.267**	**0.139**	**0.953**

Table: Performance over Epochs for DistilBERT

App

Flask & Streamlit

Flask
web development,
one drop at a time

Streamlit

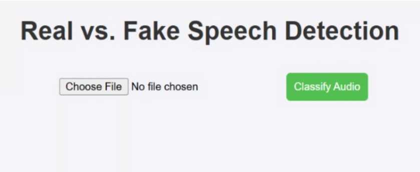


Figure: XgBoost App

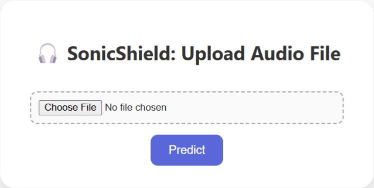


Figure: DistilBERT App

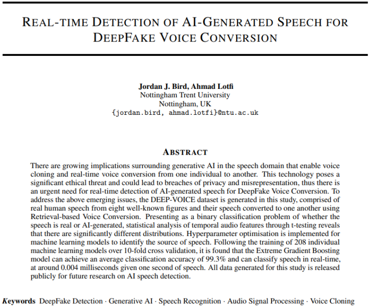
Challenges

1. Extracting relevant features from raw audio.
2. Augmenting data by splitting audio into 1-second chunks.
3. Incorporating ensemble methods for robustness.
4. Generalizing using transfer learning with DistilBERT.
5. Managing limited GPU/compute resources during training.

Future Scope

1. **Model Enhancement**
   - Explore audio-specific transformers like Wav2Vec2
2. **Dataset Expansion**
   - Collect more varied audio samples with real-world noise for better generalization.
3. **Deployment Improvements**
   - Integrate models into a Flask app; deploy on cloud platforms like AWS or Heroku.
4. **Performance Comparison**
   - Benchmark DistilBERT vs. audio-native models for accuracy and efficiency.

References



Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion

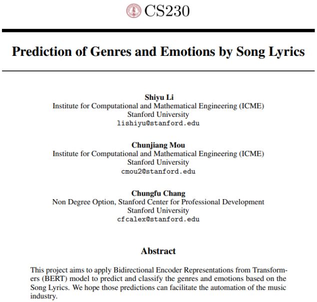
Jordan J. Bird, Ahmad Lotfi
Nottingham Trent University
Nottingham, UK
{jordan.bird, ahmad.lotfi}@ntu.ac.uk

Abstract

There are growing implications surrounding generative AI in the speech domain that enable voice cloning and real-time voice conversion from one individual to another. This technology poses a significant ethical threat and could lead to breaches of privacy and misrepresentation, thus there is an urgent need for real-time detection of AI-generated speech for DeepFake Voice Conversion. To address the above emerging issues, the DEEP-VOICE dataset is generated in this study, comprised of real human speech from eight well-known figures and their speech converted to one another using Retrieval-based Voice Conversion. Presenting as a binary classification problem of whether the speech is real or AI-generated, statistical analysis of temporal audio features through t-testing reveals that there are significantly different distributions. Hyperparameter optimisation is implemented for machine learning models to identify the source of speech. Following the training of 208 individual machine learning models over 10-fold cross validation, it is found that the Extreme Gradient Boosting model can achieve an average classification accuracy of 99.3% and can classify speech in real-time, at around 0.004 milliseconds given one second of speech. All data generated for this study is released publicly for future research on AI speech detection.

Keywords DeepFake Detection · Generative AI · Speech Recognition · Audio Signal Processing · Voice Cloning

Figure: Reference 1



CS230

Prediction of Genres and Emotions by Song Lyrics

Shiyu Li
Institute for Computational and Mathematical Engineering (ICME)
Stanford University
lishiyu@stanford.edu

Chunjiang Mou
Institute for Computational and Mathematical Engineering (ICME)
Stanford University
cmou2@stanford.edu

Chungfu Chang
Non Degree Option, Stanford Center for Professional Development
Stanford University
cfcalex@stanford.edu

Abstract

This project aims to apply Bidirectional Encoder Representations from Transformers (BERT) model to predict and classify the genres and emotions based on the Song Lyrics. We hope those predictions can facilitate the automation of the music industry.

Figure: Reference 2

Thank You

soumyajoy.mds2023@cmi.ac.in
chandranath.mds2023@cmi.ac.in
salokya.mds2023@cmi.ac.in