SIGNAL AND IMAGE PROCESSING PROJECT REPORT

REAL OR FAKE? - DETECTION OF SYNTHETIC VOICES GENERATED BY CHATTERBOX

Project Team:

- **Eda TEKEŞ** (eda.t.23@ogr.iu.edu.tr)
- **Selen GÜNEL** (seleng@ogr.iu.edu.tr)
- **Zehra ÖZTÜRK** (zehraozturk2023@ogr.iu.edu.tr)

1. INTRODUCTION AND OBJECTIVE

The rapid advancement in Text-to-Speech (TTS) systems has enabled the production of synthetic voices that are nearly indistinguishable from human speech. This evolution brings significant security risks, including identity theft and social engineering attacks.

The objective of this project is to develop a robust system capable of distinguishing between real human voices and deepfake voices generated by the **Chatterbox TTS** model. The system utilizes digital signal processing techniques and classical machine learning algorithms to achieve high-accuracy classification.

2. DATASET CREATION

2.1 Real Voice Recordings

Real voice samples were collected from volunteers under ethical guidelines. The recordings follow these technical specifications:

- **Format:** .wav (16-bit PCM)
- **Channel:** Mono
- **Sampling Rate:** 16 kHz
- **Duration:** 5–15 seconds per clip

2.2 Deepfake Voice Generation

Synthetic voices were generated using the **Chatterbox TTS** model. Real recordings were provided as audio prompts to the model to generate similar synthetic counterparts. This ensured a balanced dataset where each real voice has a corresponding deepfake version.

3. DATA PRE-PROCESSING

All audio files were standardized to 16 kHz and converted to mono. Silent segments at the beginning and end of recordings were removed using `librosa.effects.trim` to eliminate noise and improve feature extraction efficiency.

4. FEATURE EXTRACTION

The project focuses on both spectral and time-frequency domain features:

4.1 Spectral and Time-Frequency Features

- **MFCC (13 coefficients):** Captures the power spectrum of the sound.
- **Delta & Delta-Delta MFCC:** Represents the trajectories of the MFCC coefficients.
- **Zero Crossing Rate (ZCR):** Detects the rate of sign-changes in the signal.
- **Spectral Centroid / Flatness / Rolloff:** Describes the shape and "brightness" of the spectrum.
- **RMS Energy:** Represents the volume/energy levels.

4.2 Statistical Summarization

For each feature matrix, the following statistics were calculated to create a fixed-length feature vector:

- Mean, Standard Deviation, Maximum, and Minimum.

5. MODEL DEVELOPMENT

- **Data Splitting:** The dataset was divided into 80% training and 20% testing using **stratified splitting** to maintain class balance.
- **Scaling:** `StandardScaler` was applied to normalize features for optimal model performance.
- **Classifier:** A **Support Vector Machine (SVM)** with an **RBF kernel** was selected.
    - *Parameters:* C=15.0, Gamma='scale'.

6. PERFORMANCE AND EVALUATION

The model demonstrated high precision in identifying real voices without producing false alarms.

- **Overall Accuracy:** 93.75%
- **Confusion Matrix Analysis:** 8 real and 7 fake voices were correctly classified, with only one fake voice misclassified as real.

7. ETHICAL CONSIDERATIONS

This project adheres to ethical principles:

- All participants provided informed consent.
- Voice data was anonymized to protect personal information.
- Synthetic voices were generated and used strictly for research purposes.

8. LITERATURE COMPARISON

The approach of combining MFCCs with an SVM classifier is a well-established method in audio forensics. This project confirms that classical signal processing provides a stable and interpretable solution for specialized deepfake detection tasks compared to data-heavy deep learning models.

9. RESULTS

The system achieved an **Overall Accuracy of 93.75%**. For synthetic voices, the precision was 1.00, and the F1-score was 0.93, indicating high reliability in detecting voices generated by the Chatterbox architecture.

10. FUTURE WORK

Future enhancements may include:

- Expanding the dataset with diverse languages and speakers.
- Integrating prosodic features like pitch and formant frequencies.
- Developing a real-time detection application.

11. REFERENCES

- Kinnunen, T., & Li, H. (2010). *An overview of text-independent speaker recognition: From features to supervectors.* Speech Communication.
- Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition.* Academic Press.
- McFee, B., et al. (2015). *librosa: Audio and Music Signal Analysis in Python.* SciPy Conference.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.
- Wu, Z., et al. (2015). *Spoofing and countermeasures for speaker verification: A survey.* Speech Communication.
- Chatterbox TTS GitHub Repository: https://github.com/resemble-ai/chatterbox