Speech Deepfakes Detection with Fast Fourier Transform using Complex Linear Algebra

Kevin Wirya Valerian - 13524019
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
kevin.wirya.valerian@gmail.com 13524019@std.stei.itb.ac.id

***Abstract*—Nowadays, the growing realism of speech deepfakes demands detection methods grounded in fundamental signal theory rather than purely data-driven models. A well-known speech deepfake detection approach is based on the Fast Fourier Transform (FFT), rooted in complex linear algebra and geometric signal representations. Speech signals are modeled as vectors in complex vector spaces, where the FFT acts as a structured linear transformation projecting time-domain data onto orthonormal bases of complex exponentials. From a geometric perspective, genuine and synthetic speech can be represented differently in high-dimensional spectral spaces. Simple inconsistencies in phase behavior and energy distribution can be analyzed through algebraic measures derived from the complex FFT representation. This study demonstrates the applicability of FFT-based complex spectral analysis for distinguishing AI-generated and genuine speech.**

***Index Terms*—Fast fourier transform, deepfakes detection, complex linear algebra, spectral geometry**

I. Introduction

Recent advances in speech synthesis and voice conversion have enabled machines to generate highly realistic human speech. Modern text-to-speech and voice cloning systems are capable of producing audio that is increasingly difficult to distinguish from genuine human recordings. A study reports that human participants were only able to accurately distinguish real from AI-generated voices with an accuracy of 70.4%. This leads to a serious risks that can be posed by speech deepfakes in areas such as biometric authentication and digital forensics. Reports indicate that voice-based fraud has increased significantly in recent years, with financial losses reaching billions of dollars globally. Deepfake-related losses have already reached $1.56 billion, with over $1 billion occurring in 2025 alone.

While deep learning models have achieved impressive performance in generating natural-sounding speech, detecting such synthetic audio remains a challenging task. Many existing detection methods rely heavily on large neural networks trained on specific datasets. Although effective, these approaches often lack interpretability and tend to degrade when exposed to unseen deepfake generation methods. Moreover, purely data-driven models can be computationally expensive and difficult to analyze, making them less suitable and reliable at present time.

Speech signals, however, are fundamentally mathematical objects. A digital audio signal can be represented as a finite sequence of samples, which naturally forms a vector in a high-dimensional space. Transforming this signal into the frequency domain using the Fast Fourier Transform (FFT) reveals its spectral structure, including phase behavior. These properties are governed by well-established principles of linear algebra and geometry, such as vector spaces, orthonormal bases, inner or dot products, and unitary transformations. From a linear algebra perspective, the FFT can be interpreted as a linear transformation that maps a time-domain signal into a complex frequency-domain vector. While synthetic speech models often reproduce spectral magnitudes effectively, differences in phase behavior may still arise.

Motivated by this observation, this paper proposes a speech deepfake detection approach grounded in the linear algebraic and geometric interpretation of the FFT. Rather than relying on large data-driven models, the proposed method focuses on analyzing basic algebraic properties of complex spectral representations. The approach aims to provide an interpretable demonstration of how concepts from complex linear algebra and geometry can be applied to distinguish genuine and synthetic speech signals.

II. Theoretical Framework

A. Complex Numbers and Geometric Representation

A complex number is defined as

$$z = a + jb, \quad a, b \in \mathbb{R} \tag{1}$$

where a is the real part and b is the imaginary part, with $j = \sqrt{-1}$. Both i and j represent the same imaginary unit. We use j since it is commonly used for signal processing.

Complex numbers admit a geometric interpretation in the complex plane, where the real part corresponds to the horizontal axis and the imaginary part to the vertical axis. This representation allows us to visualize complex numbers as vectors emanating from the origin.

Every non-zero complex number can be expressed in polar form

$$z = re^{j\theta} \tag{2}$$

where

- $r = |z| = \sqrt{a^2 + b^2}$ is the **magnitude**, representing the distance from the origin

- $\theta = \arg(z) = \arctan(b/a)$ is the **phase**, representing the angle from the positive real (horizontal) axis

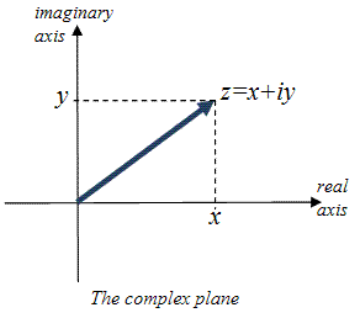

Fig. 1. Argand Plane (Complex Plane)
(Source: https://helpingwithmath.com/complex-plane/)

The exponential form relates to trigonometry via Euler's formula.

$$e^{j\theta} = \cos\theta + j\sin\theta \tag{3}$$

The FFT spectrum of audio signals consists of complex-valued coefficients. Each coefficient X_k can be decomposed into magnitude and phase components. The magnitude $|X_k|$ represents the energy at frequency bin k, while the phase $\angle X_k$ encodes temporal structure. A frequency bin is a specific range on the frequency axis used to group and analyze data. Our detection method exploits the observation that human speech exhibits **regular** phase structure, whereas AI-generated speech tends to have **irregular** phase structure.

B. Complex Vector Spaces $\mathbb{C}^N$

The set $\mathbb{C}^N$ of all N-tuples of complex numbers forms a vector space over the field $\mathbb{C}$. A vector $\mathbf{x} \in \mathbb{C}^N$ is written as

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}, \quad x_n \in \mathbb{C} \tag{4}$$

Digital audio signals are naturally elements of $\mathbb{R}^N$ in the time domain and $\mathbb{C}^N$ in the frequency domain after FFT transformation. The standard inner product on $\mathbb{C}^N$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=0}^{N-1} x_n \overline{y_n} \tag{5}$$

where $\overline{y_n}$ denotes the complex conjugate of y_n. This inner product induces the Euclidean norm.

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{n=0}^{N-1} |x_n|^2} \tag{6}$$

The squared norm $\|\mathbf{x}\|^2$ represents the total energy of the signal.

C. Discrete and Fast Fourier Transform

The Discrete Fourier Transform (DFT) converts a time-domain signal $\mathbf{x} \in \mathbb{C}^N$ to its frequency-domain representation $\mathbf{X} \in \mathbb{C}^N$ via:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1 \tag{7}$$

This transformation can be interpreted as computing the inner product of the signal with complex exponential basis vectors at each frequency k. The inverse DFT reconstructs the time-domain signal:

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}, \quad n = 0, 1, \ldots, N-1 \tag{8}$$

The naive DFT requires $O(N^2)$ complex multiplications. The Fast Fourier Transform (FFT) reduces this to $O(N \log N)$ using a divide-and-conquer strategy. The key insight is that DFT can be decomposed into smaller DFTs of even and odd indexed elements.

Let $\omega_N = e^{-j2\pi/N}$ be the primitive N-th root of unity. For $N = 2^m$, we split the DFT:

$$X_k = \sum_{n=0}^{N-1} x_n \omega_N^{kn} \tag{9}$$

$$= \sum_{n=0}^{N/2-1} x_{2n} \omega_N^{k(2n)} + \sum_{n=0}^{N/2-1} x_{2n+1} \omega_N^{k(2n+1)} \tag{10}$$

$$= \sum_{n=0}^{N/2-1} x_{2n} (\omega_N^2)^{kn} + \omega_N^k \sum_{n=0}^{N/2-1} x_{2n+1} (\omega_N^2)^{kn} \tag{11}$$

Since $\omega_N^2 = \omega_{N/2}$, this becomes

$$X_k = E_k + \omega_N^k O_k \tag{12}$$

where E_k is the DFT of even-indexed elements and O_k is the DFT of odd-indexed elements, both of size $N/2$.

Using the symmetry property $X_{k+N/2} = E_k - \omega_N^k O_k$, we compute both halves with a single recursion:

$$X_k = E_k + \omega_N^k O_k, \quad k = 0, \ldots, N/2-1 \tag{13}$$

$$X_{k+N/2} = E_k - \omega_N^k O_k, \quad k = 0, \ldots, N/2-1 \tag{14}$$

The recursion bottoms out at $N = 1$, where $X_0 = x_0$. The total complexity satisfies:

$$T(N) = 2T(N/2) + O(N) = O(N \log N) \tag{15}$$

The FFT preserves several important properties crucial for our detection method:

1. Energy Preservation (Parseval's Theorem)

$$\sum_{n=0}^{N-1} |x_n|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X_k|^2 \tag{16}$$

This ensures that total signal energy is conserved between time and frequency domains.

2. Linearity

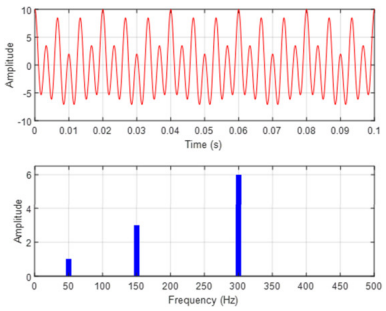

Fig. 2. Time to Frequency Mapping By FFT (Complex Plane) (Source: https://www.sciencedirect.com/topics/engineering/fast-fourier-transform)

$\text{FFT}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\text{FFT}(\mathbf{x}) + \beta\text{FFT}(\mathbf{y})$, allowing us to analyze signal components independently.

3. Symmetry for Real Signals:

When $x_n \in \mathbb{R}$, we have $X_{N-k} = \overline{X_k}$, so only the first $N/2$ coefficients need to be stored.

For our implementation, audio signals are real-valued in the time domain, so we extract magnitude and phase from the positive frequency components $k = 0, \ldots, N/2 - 1$.

D. Magnitude and Phase as Complex Algebraic Objects

Each FFT coefficient admits the polar decomposition:

$$X_k = |X_k| e^{j\theta_k} \tag{17}$$

where:

- $|X_k| = \sqrt{\text{Re}(X_k)^2 + \text{Im}(X_k)^2}$ is the spectral magnitude
- $\theta_k = \arctan\left(\frac{\text{Im}(X_k)}{\text{Re}(X_k)}\right)$ is the spectral phase

Phase coherence measures the consistency of phase relationships across frequency bins. For a phase spectrum $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{N-1})$, we define

$$C(\boldsymbol{\theta}) = \left| \frac{1}{N} \sum_{k=0}^{N-1} e^{j\theta_k} \right| \tag{18}$$

This metric ranges from 0 to 1:

- $C = 1$: Perfectly coherent
- $C = 0$: Completely random
- $0 < C < 1$: Partial coherence

The phase velocity (or phase derivative) measures the smoothness of phase progression

$$v_k = \theta_{k+1} - \theta_k \pmod{2\pi} \tag{19}$$

The variance of phase velocity indicates regularity:

$$\sigma_v^2 = \text{Var}(v_k) \tag{20}$$

Low variance suggests smooth and natural phase evolution, while high variance indicates abrupt phase changes typical of synthesis artifacts.

In our implementation, we quantify phase smoothness using the mean absolute phase gradient:

$$\bar{v} = \frac{1}{N-1} \sum_{k=0}^{N-2} |v_k| \tag{21}$$

where smaller values of $\bar{v}$ indicate smoother phase transitions characteristic of natural speech, while larger values suggest the discontinuous phase patterns often present in synthesized audio.

Modern AI voice synthesis (neural vocoders, GANs) primarily optimizes for perceptually accurate magnitude spectra because human hearing is more sensitive to magnitude than phase. However, these models often fail to produce coherent phase structure because phase reconstruction is mathematically ill-conditioned.

E. Spectral Entropy and Energy Distribution

Spectral entropy quantifies the concentration of energy across the frequency spectrum. For a magnitude spectrum $\mathbf{M} = (|X_0|, |X_1|, \ldots, |X_{N-1}|)$, we first normalize to obtain a probability distribution

$$p_k = \frac{|X_k|}{\sum_{i=0}^{N-1} |X_i|} \tag{22}$$

The spectral entropy is then defined as

$$H(\mathbf{M}) = -\sum_{k=0}^{N-1} p_k \log p_k \tag{23}$$

This metric characterizes the distribution of spectral energy

- **Low entropy**: Energy concentrated in few frequency bins (typical of human speech)
- **High entropy**: Energy spread uniformly across frequencies (may indicate synthesis artifacts)

The spectral L2 norm

$$\|\mathbf{M}\|_2 = \sqrt{\sum_{k=0}^{N-1} |X_k|^2} \tag{24}$$

represents the total energy in the signal and serves as a normalization factor for geometric distance computations.

F. Window-Based Phase Coherence Analysis

While global phase coherence provides an overall measure, local phase patterns can reveal subtle artifacts. We employ a sliding window approach to compute localized coherence. For a window size w, the local phase coherence at position i is

$$C_i^{(w)} = \frac{1}{w} \left| \sum_{k=i}^{i+w-1} e^{j\theta_k} \right| \tag{25}$$

The overall phase coherence is then the mean of all window coherences:

$$C_{\text{window}} = \frac{1}{N-w} \sum_{i=0}^{N-w-1} C_i^{(w)} \tag{26}$$

This windowed approach offers several advantages:

- Captures local phase consistency patterns
- More robust to isolated phase discontinuities
- Reveals frequency-dependent phase artifacts common in AI synthesis

The window size w is typically chosen based on the expected correlation length of natural speech phase patterns.

G. Data Clustering

In our framework, audio samples form clusters in the complex feature space. Let $\mathcal{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_M\}$ be the set of feature vectors from human speech samples, with centroid

$$\boldsymbol{\mu}_H = \frac{1}{M} \sum_{i=1}^{M} \mathbf{h}_i \tag{27}$$

Similarly, AI-generated samples form a cluster $\mathcal{A}$ with centroid $\boldsymbol{\mu}_A$.

H. Mahalanobis-Like Distance and Confidence Estimation

To quantify classification confidence, we compute standardized distances from the test sample to each cluster centroid. Given the phase coherence C of a test sample, and the statistics (μ_H, σ_H) and (μ_A, σ_A) from the human and AI training sets respectively, we define:

$$d_H(C) = \frac{|C - \mu_H|}{\sigma_H + \epsilon} \tag{28}$$

$$d_A(C) = \frac{|C - \mu_A|}{\sigma_A + \epsilon} \tag{29}$$

where ϵ is a small constant to prevent division by zero. These distances are analogous to Mahalanobis distances in one dimension, accounting for the variance of each class.

The confidence score is computed from the relative distances:

$$\text{confidence} = 1 - \frac{\min(d_H, d_A)}{\max(d_H, d_A) + \epsilon} \tag{30}$$

This confidence metric has the following properties:

- confidence $\in [0, 1]$
- High confidence (≈ 1): Test sample is much closer to one cluster than the other
- Low confidence (≈ 0): Test sample is equidistant from both clusters (ambiguous case)

The predicted class is the one with smaller distance:

$$\text{prediction} = \begin{cases} \text{HUMAN} & \text{if } d_H < d_A \\ \text{AI-GENERATED} & \text{otherwise} \end{cases} \tag{31}$$

This geometric approach provides both a classification decision and a measure of certainty, essential for practical deployment where uncertain predictions may require human review.

III. Methodology

The detection system operates through a four-stage pipeline. Those are audio preprocessing, frequency domain transformation, feature extraction, and threshold-based classification. The core hypothesis is that human speech exhibits distinct phase relationships in the frequency domain that differ systematically from AI-generated audio. These differences arise from fundamental mechanisms that human speech originates from physical vocal cord vibrations and acoustic resonance, while AI-generated speech is synthesized digitally, potentially introducing artifacts in phase structure.

A. Signal Processing

The audio signal is first loaded and normalized to a consistent sampling rate and amplitude range. Stereo recordings are converted to mono by averaging channels. The preprocessed time-domain signal then undergoes Fast Fourier Transform (FFT), converting it from amplitude-over-time representation to amplitude-and-phase-over-frequency representation.

The FFT produces a complex-valued vector where each element corresponds to a frequency component. The magnitude represents energy at that frequency, while the phase angle captures timing relationships. The system retains only positive frequency components, exploiting the symmetry property of real-valued signals. This frequency domain representation becomes the foundation for all subsequent analysis.

B. Feature Extraction

From the FFT output, the system extracts three key features characterizing the signal's geometric properties. First, phase coherence serves as the primary discriminative feature. They are used for quantify how consistently phase transitions across adjacent frequency bins. The computation analyzes phase values in sliding windows across the spectrum. High coherence indicates regular phase transitions typical of human speech. On the other hand, low coherence suggests random or artificial patterns common in nonhuman speech.

Second, phase velocity measures the rate of phase change across frequency. Human speech typically exhibits smoother phase curves due to continuous physical sound production, while AI-generated speech may show abrupt transitions.

Third, spectral entropy quantifies energy distribution across frequencies. Low entropy indicates concentration in specific bands, while high entropy suggests uniform distribution.

C. Classification Strategy

The classification employs a geometric threshold-based approach rather than machine learning. Before classification, the system computes reference statistics from known human and AI-generated speech samples. For each dataset, phase coherence values are extracted from multiple files, and statistical measures (mean and standard deviation) are computed. The decision threshold is established at the midpoint between the human and AI mean phase coherence values.

For a test audio sample, the system extracts its phase coherence and applies a simple decision rule. If phase coherence exceeds the threshold, it is classified as human. Otherwise, it is classified as nonhuman. To quantify confidence, the system computes normalized geometric distances from the test sample to each class reference, similar to Mahalanobis distance. The distance to each class measures how many standard deviations the test sample deviates from that class's mean. Confidence reflects how much closer the sample is to its predicted class compared to the alternative.

This threshold-based strategy offers immediate interpretability with a clear decision boundary, minimal calibration requirements, and avoidance of overfitting risks. The approach prioritizes mathematical transparency and geometric reasoning

over complex learned parameters, making the classification process fully explainable through linear algebra principles.

IV. Implementation

A. Signal Processing

```python
def load_wav(self, filepath):
        try:
            filepath = Path(filepath)
            file_ext = filepath.suffix.lower()
            # Load MP3 files with librosa
            if file_ext == '.mp3':
                if not HAS_LIBROSA:
                    raise ValueError("librosa not installed")
                signal, sr = librosa.load(str(filepath), sr=None, mono=True)
                return signal, sr
            # Load WAV files
            elif file_ext == '.wav':
                sr, signal = wavfile.read(filepath)
                # Convert stereo to mono if needed
                if len(signal.shape) > 1:
                    signal = np.mean(signal, axis=1)
                # Normalize to [-1, 1]
                if signal.dtype in [np.int16, np.int32]:
                    signal = signal.astype(np.float32) / np.max(np.abs(signal))
                else:
                    signal = signal.astype(np.float32)
                return signal, sr
            else:
                raise ValueError(f"Unsupported file format: {file_ext}. Use WAV or MP3.")
        except Exception as e:
            raise ValueError(f"Error loading audio file: {str(e)}")
```

Fig. 3. Loading Audio Files
(Source: Author)

First, audio file input was handled by first detecting the file extension. For WAV files, it uses scipy.io.wavfile.read to load the audio data and metadata, while MP3 files are loaded using the librosa library. The method then applies preprocessing steps. The signal is normalized to a floating-point range of [-1, 1].

```python
def compute_spectral_features(self, signal, sr=None):
        # Apply FFT
        X = fft(signal)
        # Keep only positive frequencies
        N = len(X)
        X = X[:N//2]
        # Compute magnitude and phase in polar form
        magnitude = np.abs(X)
        phase = np.angle(X)
        # Frequency bins (in Hz)
        if sr is None:
            sr = self.target_sr
        freq = np.fft.fftfreq(N, d=1/sr)[:N//2]
        return {
            'X': X,                          # Complex spectral vector
            'magnitude': magnitude,          # Energy spectrum
            'phase': phase,                  # Phase spectrum
            'freq': freq,                    # Frequency axis
            'N': N                           # Original signal length
        }
```

Fig. 4. Computing Mathematical Features of Audio
(Source: Author)

There is a part of the code to transform the time-domain signal into the frequency domain using Fast Fourier Transform (FFT) via scipy.fftpack.fft. All results—complex coefficients, magnitude, phase, and frequency—are organized into a dictionary for downstream feature calculations.

Spectral coherence is computed using a sliding window approach. The method returns both the overall coherence as the mean of all window coherences and an array of per-window values for diagnostic purposes. Next, the the smoothness of phase is quantified by computing the firt-order difference between adjacent phase. Last, there exists a part for extracting L2 norm, then normalize the magnitude vector using it. By

```python
def compute_phase_coherence(self, phase, window_size=5):
        try:
            # Remove NaN and infinite values
            phase = np.nan_to_num(phase, nan=0.0, posinf=0.0, neginf=0.0)
            # Ensure phase is valid
            if len(phase) < window_size:
                return 0.5, np.array([0.5])

            # Compute coherence in sliding windows
            coherence_bins = []
            for i in range(len(phase) - window_size):
                window_phase = phase[i:i+window_size]
                # Sum phase vectors: Σ exp(j*∠X(k))
                phase_sum = np.abs(np.sum(np.exp(1j * window_phase)))
                # Normalize by window size
                coherence = phase_sum / window_size
                coherence_bins.append(coherence)

            # Overall coherence: mean of window coherences
            overall_coherence = np.mean(coherence_bins) if coherence_bins else 0.5
            overall_coherence = np.clip(overall_coherence, 0.0, 1.0)

            return overall_coherence, np.array(coherence_bins)
        except Exception as e:
            return 0.5, np.array([0.5])

def compute_phase_velocity(self, phase):
        # Phase difference between adjacent frequency bins
        phase_velocity = np.diff(phase)
        phase_velocity = np.angle(np.exp(1j * phase_velocity))
        # Mean absolute velocity / smoothness metric
        velocity_smoothness = np.mean(np.abs(phase_velocity))

        return velocity_smoothness

def compute_spectral_inner_products(self, magnitude):
    # Euclidean norm of magnitude vector
    l2_norm = np.linalg.norm(magnitude, ord=2)
    # Normalized spectral shape (unit vector for cosine similarity)
    if l2_norm > 0:
        spectral_shape = magnitude / l2_norm
    else:
        spectral_shape = magnitude

    # Spectral entropy
    prob = (magnitude + 1e-10) / (np.sum(magnitude) + 1e-10)
    entropy = -np.sum(prob * np.log(prob))
    return {
        'l2_norm': l2_norm,
        'spectral_shape': spectral_shape,
        'entropy': entropy
    }
```

Fig. 5. Computing Phase Coherence, Phase Velocity, and Spectral Entropy
(Source: Author)

using entropy formula, in the end, three metrics are returned in a dictionary. Extraction of all features from a file need to be executed for a complete information gathering.

```python
def extract_all_features(self, filepath):
        # Extracting all features

def extract_features_from_file(filepath):
        # Extracting from file
```

Fig. 6. Extraction
(Source: Author)

B. Reference Statistics

```python
def get_wav_files(self, directory):
        audio_files = []
        if os.path.exists(directory):
            audio_files = list(Path(directory).glob('**/*.wav'))
            audio_files += list(Path(directory).glob('**/*.mp3'))
        return sorted(audio_files)
```

Fig. 7. Obtaining WAV Files
(Source: Author)

All audio files inside a certain directory are recursively searched by using glob patterns to find files matching *.wav and *.mp3. The method returns a sorted list of Path objects representing all discovered audio files.

```python
def compute_statistics(self, verbose=True):
        stats = {
            'human': {'coherences': [], 'velocities': [], 'entropies': []},
            'nonhuman': {'coherences': [], 'velocities': [], 'entropies': []}
        }
        # Process human speech
        for filepath in human_files:
            try:
                features = self.processor.extract_all_features(str(filepath))
                stats['human']['coherences'].append(features['phase_coherence'])
                stats['human']['velocities'].append(features['phase_velocity'])
                stats['human']['entropies'].append(features['spectral_entropy'])
            except Exception as e:

        # Process AI-generated speech
        if verbose:
        nonhuman_files = self.get_wav_files(self.nonhuman_dir)
        for filepath in nonhuman_files:
            try:
                features = self.processor.extract_all_features(str(filepath))
                stats['nonhuman']['coherences'].append(features['phase_coherence'])
                stats['nonhuman']['velocities'].append(features['phase_velocity'])
                stats['nonhuman']['entropies'].append(features['spectral_entropy'])
            except Exception as e:

        # Compute aggregate statistics
        result = {}

        for label in ['human', 'nonhuman']:
            coherences = np.array(stats[label]['coherences'])
            velocities = np.array(stats[label]['velocities'])
            entropies = np.array(stats[label]['entropies'])
            # Return all results

        # Print results

        return result
```

Fig. 8. Computing Statistics
(Source: Author)

Baseline statistics must be computed to process human and nonhuman datasets. After processing all files, the method computes aggregate statistics for features such as mean, standard deviation, minimum, and maximum values.

```python
def compute_and_save_reference_stats(human_dir='../data/human',
                                     nonhuman_dir='../data/nonhuman',
                                     output_file='reference_stats.json'):
    computer = ReferenceStatisticsComputer(human_dir, nonhuman_dir)
    stats = computer.compute_statistics(verbose=True)
    computer.save_statistics(stats, output_file)
    return stats
```

Fig. 9. Computing and Saving Reference Statistics
(Source: Author)

The rest of the code is used to save and load the statistics. When the application is first started, the data will get computed by the functions there.

C. Detector

```python
def _compute_geometric_distance(self, features):
        # Weights based on discriminative power
        distances_human = []
        distances_ai = []
        metrics = ['phase_coherence', 'phase_velocity', 'spectral_entropy']
        for metric in metrics:
            # Standardized distances per metric
            # Apply weights

        # Weighted Euclidean distance
        d_human = np.sqrt(np.sum(np.array(distances_human)**2))
        d_ai = np.sqrt(np.sum(np.array(distances_ai)**2))

        # Confidence based on relative distances
        min_dist = min(d_human, d_ai)
        max_dist = max(d_human, d_ai)
        confidence = 1.0 - (min_dist / (max_dist + 1e-6))

        return d_human, d_ai, confidence
```

Fig. 10. Computing Geometric Distance
(Source: Author)

Geometric distance calculation is used for confidence classification. It is derived from the formula mentioned in the theoretical framework. It handles errors and also division by zero. In addition, we use weights approach to compute the distance in order to maximize the usage of all three metrices.

```python
    def predict(self, audio_filepath, verbose=False):
        features = self.processor.extract_all_features(audio_filepath)
        # Compute geometric distances
        d_h, d_ai, confidence = self._compute_geometric_distance(features)
        # Decision using geometric distance
        if d_ai < d_h:
            primary_prediction = 'ai'
        else:
            primary_prediction = 'human'
        result = {
            # Result details
        }
        return result

    def predict_batch(self, audio_files_list, verbose=False):
        predictions = []
        for filepath in audio_files_list:
            try:
                result = self.predict(filepath, verbose=verbose)
                result['filepath'] = str(filepath)
                predictions.append(result)
            except Exception as e:
                predictions.append({
                    'filepath': str(filepath),
                    'error': str(e),
                    'prediction': None,
                    'confidence': None
                })
        return predictions
```

Fig. 11. Classification Based on Geometric Distance
(Source: Author)

Prediction performs the main classification task. The method applies computed geometric distance to human and to AI to classify whether the audio is human or nonhuman. It calculates the confidence score via this data. We used the help of an additional function to predict multiple audio files.

D. Application

```python
from flask import Flask, request, jsonify # building rest APIs and handling HTTP
from flask_cors import CORS # communicate with backend API
import os # OS interface
import tempfile # creates temporary directories for storing uploaded files
from pathlib import Path # OO file path manipulation
import json # reading config files
from detector import DeepfakeDetector
from reference_stats import compute_and_save_reference_stats

# Initialize Flask app
app = Flask(__name__)
CORS(app)

# Configuration
UPLOAD_FOLDER = tempfile.gettempdir()
ALLOWED_EXTENSIONS = {'wav', 'mp3'}
REFERENCE_STATS_FILE = 'reference_stats.json'

# Global detector instance
detector = None

def allowed_file(filename):
    return '.' in filename and filename.rsplit('.', 1)[1].lower() in ALLOWED_EXTENSIONS

def initialize_detector():
# GET and POST methods handling
# Error handling
# Main functions
```

Fig. 12. Application System
(Source: Author)

The Flask web application serves as the user interface for the deepfake detection system. When initializing detector, system automatically checks for the existence of reference statistics and computes them from the training datasets if unavailable, ensuring the system operates well on first launch. Upon receiving a file, the endpoint validates the file format,

saves it temporarily to disk, invokes the deepfake detector to perform classification, and returns a JSON response containing the prediction label (human or AI-generated), confidence score, and detailed spectral metrics including phase coherence, geometric distances to both classes, phase velocity, spectral entropy, and L2 norm. Additional endpoints provide system status checks and access to reference statistics for diagnostic purposes.

V. Case Analysis

This case analysis evaluates the robustness of the detection system through testing of both human and nonhuman audio samples. The human samples are sourced from real people, while the nonhuman samples sourced from **Elevenlabs**. The author picked Elevenlabs since it let user to vary the components of the generated audio. The nonhuman sample set comprises three distinct categories speed, stability, and similarity. By introducing controlled variations, we isolate the individual factors that influence detection performance. This approach allows us to understand not only the system's ability to distinguish human from AI-generated audio under ideal conditions, but also its resilience to common variations that AI systems might employ to evade detection. Consequently, this comprehensive evaluation reveals whether the detection metrics are robust indicators of audio authenticity or if they can be manipulated through algorithmic parameter adjustments.

There are several variables inside the tables to represent each computed value. Variables, C, V, and E, consecutively, represent phase coherence, phase velocity, and spectral entropy. While d_H and d_{AI} represent the geometric distance to human and AI. As primary indicator, we use the abbreviation $Pred.$ for prediciton and Cf for confidence rate in percentage.

A. Case 1: Human Samples

TABLE I
Detection Performance for Human Samples

Human Samples							
Sample	**C**	**V**	**E**	$\mathbf{d_H}$	$\mathbf{d_{AI}}$	**Pred.**	**Cf (%)**
H-1	0.496	1.390	10.008	0.804	1.324	Human	39.3
H-2	0.468	1.309	10.015	0.796	1.317	Human	39.5
H-3	0.631	0.973	9.988	0.895	1.380	Human	35.1
H-4	0.704	0.842	9.911	1.006	1.480	Human	32.0
H-5	0.694	0.799	10.348	0.673	1.057	Human	36.3
H-6	0.365	1.603	10.257	0.550	1.062	Human	48.2
H-7	0.246	1.838	9.460	1.324	1.894	Human	30.1
H-8	0.249	1.984	9.909	0.910	1.439	Human	36.8

The detection system correctly identified all eight human audio samples with 100% accuracy. The phase coherence values range from 0.246 (H-7) to 0.704 (H-4), phase velocity spans from 0.799 (H-5) to 1.984 (H-8), and spectral entropy varies between 9.460 (H-7) and 10.348 (H-5). Across all samples, the geometric distance to human class (d_H) ranges from 0.550 to 1.324, while the distance to AI class (d_{AI}) ranges from 1.057 to 1.894. All samples consistently maintain $d_H < d_{AI}$, which results in correct human predictions. The confidence rates vary from 30.1% (H-7) to 48.2% (H-6). The data demonstrates that despite significant variations in individual feature values, the geometric distance metric successfully distinguishes human speech from AI-generated audio in all test cases.

B. Case 2: Nonhuman Samples with Speed Variation

TABLE II
Detection Performance for AI Samples with Speed Variation

Nonhuman Samples								
Sample	Speed	**C**	**V**	**E**	$\mathbf{d_H}$	$\mathbf{d_{AI}}$	**Pred.**	**Cf (%)**
AI-1	0.70x	0.379	1.590	11.924	1.043	0.663	AI	36.5
AI-2	0.80x	0.430	1.506	11.899	1.023	0.639	AI	37.5
AI-3	0.90x	0.381	1.607	11.752	0.880	0.486	AI	44.8
AI-4	1.00x	0.421	1.612	11.673	0.806	0.405	AI	49.8
AI-5	1.10x	0.401	1.632	11.533	0.671	0.260	AI	61.2
AI-6	1.20x	0.379	1.697	11.380	0.526	0.115	AI	78.1

Using stability of 30 and similarity of 20, all six AI-generated samples with varying speed parameters were correctly classified as AI with 100% accuracy. As speed increases from 0.70x to 1.20x, spectral entropy decreases from 11.924 to 11.380, and phase velocity increases from 1.590 to 1.697. Phase coherence fluctuates between 0.379 and 0.430 without a consistent trend. The geometric distance to human class (d_H) decreases from 1.043 to 0.526, and distance to AI class (d_{AI}) decreases from 0.663 to 0.115 as speed increases. The confidence rate shows a strong positive correlation with speed, rising from 36.5% at 0.70x to 78.1% at 1.20x. The data indicates that faster playback speeds produce samples that are more distinguishable as AI-generated, with the system achieving its highest confidence at 1.20x speed.

C. Case 3: Nonhuman Samples with Stability Variation

TABLE III
Detection Performance for AI Samples with Stability Variation

Human Samples								
Sample	Stab.	**C**	**V**	**E**	$\mathbf{d_H}$	$\mathbf{d_{AI}}$	**Pred.**	**Cf (%)**
AI-1	0	0.424	1.464	11.968	1.089	0.710	AI	34.8
AI-2	20	0.390	1.575	11.871	0.993	0.608	AI	38.8
AI-3	40	0.392	1.557	11.979	1.096	0.719	AI	34.4
AI-4	60	0.438	1.478	11.811	0.939	0.548	AI	41.6
AI-5	80	0.401	1.558	11.871	0.994	0.609	AI	38.8
AI-6	100	0.399	1.594	11.917	1.037	0.656	AI	36.8

Using normal speed and similarity of 20, all six AI-generated samples with stability parameters ranging from 0 to 100 were correctly classified as AI with 100% accuracy. The phase coherence values range from 0.390 to 0.438, phase velocity varies between 1.464 and 1.594, and spectral entropy spans from 11.811 to 11.979. The geometric distance to human class (d_H) ranges from 0.939 to 1.096, while distance to AI class (d_{AI}) ranges from 0.548 to 0.719. Confidence rates fluctuate between 34.4% (stability 40) and 41.6% (stability 60), showing no clear monotonic relationship with the stability parameter. Sample AI-4 with stability 60 achieves the highest confidence at 41.6%, while samples AI-1 and AI-3 show the lowest confidence at 34.8% and 34.4% respectively. The data reveals that stability variations do not significantly impact the detection system's ability to classify AI audio, with all samples maintaining $d_{AI} < d_H$ and relatively consistent confidence levels.

D. Case 4: Nonhuman Samples with Similarity Variation

TABLE IV
DETECTION PERFORMANCE FOR AI SAMPLES WITH SIMILARITY VARIATION

Human Samples								
Sample	Sim.	**C**	**V**	**E**	$\mathbf{d_H}$	$\mathbf{d_{AI}}$	**Pred.**	**Cf (%)**
AI-1	0	0.395	1.609	11.912	1.033	0.651	AI	37.0
AI-2	20	0.392	1.657	11.909	1.030	0.648	AI	37.1
AI-3	40	0.398	1.610	11.892	1.014	0.630	AI	37.8
AI-4	60	0.399	1.575	11.892	1.014	0.630	AI	37.9
AI-5	80	0.390	1.624	11.866	0.988	0.603	AI	39.0
AI-6	100	0.393	1.638	11.926	1.046	0.665	AI	36.4

Using normal speed and stability of 30, all six AI-generated samples with similarity parameters ranging from 0 to 100 were correctly classified as AI with 100% accuracy. Phase coherence values remain narrow between 0.390 and 0.399, phase velocity ranges from 1.575 to 1.657, and spectral entropy varies minimally from 11.866 to 11.926. The geometric distance to human class (d_H) ranges from 0.988 to 1.046, and distance to AI class (d_{AI}) ranges from 0.603 to 0.665. Confidence rates show minimal variation, spanning from 36.4% (similarity 100) to 39.0% (similarity 80), with most samples clustered around 37%. The maximum confidence difference across all similarity levels is only 2.6 percentage points. The data demonstrates that similarity parameter adjustments have negligible impact on the detection system's performance, with all samples exhibiting nearly identical feature characteristics and consistently maintaining $d_{AI} < d_H$.

VI. CONCLUSION

In conclusion, this paper presents a novel approach to AI-generated audio detection by leveraging phase-based analysis through the fast fourier transform (FFT). The system created by the author employs three key metrics—phase coherence, phase velocity, and spectral entropy to distinguish between human and nonhuman speech. The mathematical foundation utilizes windowing techniques to segment audio signals, applies FFT for frequency domain transformation, and extracts phase information to identify subtle artifacts in synthetic speech. These metrics are integrated through a weighted classification approach that evaluates the temporal consistency, spectral distribution, and phase relationships inherent in audio signals, providing a computationally efficient framework for deepfake detection.

However, the system exhibits certain limitations that affect its detection accuracy regarding the stability and similarity parameter since it shows no major findings. Needless to say, this system will be obsolete. The system might fail to distinguish authentic audio from synthetic speech when environmental factors such as background noise, low-quality recordings, or post-processing effects are present, as these conditions can introduce phase distortions similar.

Several improvements are recommended to increase system's performance. First, incorporating additional spectral features such as mel-frequency cepstral coefficients (MFCCs) or formant analysis could provide complementary information to phase-based metrics. Second, implementing machine learning classifiers trained on diverse datasets would enable adaptive threshold determination and better generalization across different AI synthesis models. Third, developing noise-robust preprocessing techniques would improve performance in real-world scenarios with varying audio quality. Last but not least, there are many approaches for detecting speech deepfakes, using FFT is not always the most optimal approach. Nevertheless, as long as technology advancement is concerned, FFT is one of the most efficient algorithms to detect AI-generated speech in an optimized time.

ATTACHMENT

Github: Source Code of "Speech Deepfakes Detection with Fast Fourier Transform using Complex Linear Algebra"
Here

Youtube Video: Demonstration on The Source Code "Speech Deepfakes Detection with Fast Fourier Transform using Complex Linear Algebra" using Python
Here

ACKNOWLEDGMENT

The author gratefully acknowledges the blessings and strength granted by God Almighty, which enabled the successful completion of this paper. Sincere appreciation is also extended to Dr. Ir. Rinaldi, M.T., lecturer of the IF2123 Linear and Geometric Algebra course, for his continuous support, guidance, and encouragement throughout the semester and in the writing of this paper. Next, the author would thank both of the author's parents who always give the author supports during ups and downs during the process of this work. Last but not least, the author would also like to thank James Cooley and John Tukey, which significantly inspired this work by popularizing FFT algorithm. It impacts meaningfully to the author's competitive programming experience.

REFERENCES

[1] Burrus, C. S. (1989). Algorithms for Discrete Fourier Transform and Convolution. Springer Science+Business Media New York. (Accessed on December 20, 2025). https://link.springer.com/book/10.1007/978-1-4757-3854-4

[2] Christensen, M. G. (2019). Introduction to Audio Processing. Springer Nature Switzerland AG. (Accessed on December 19, 2025). https://doi.org/10.1007/978-3-030-11781-8

[3] Mai, K. T., Bray, S., Davies, T., and Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. PLOS ONE, 18(8):1–20. (Accessed on December 19, 2025). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285333

[4] Munir, R. (2025). IF2123 Aljabar Linier dan Geometri - Semester I Tahun 2025/2026: Aljabar Kompleks (Update 2024). (Accessed on December 19, 2025). https://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2025-2026/Algeo-25-Aljabar-Kompleks-2025.pdf

[5] Ni, Y., et al. (2024). A Deepfake Detection Algorithm Based on Fourier Transform of Biological Signal. Tech Science Press. (Accessed on December 16, 2025). https://www.techscience.com/cmc/v79n3/57116

[6] Nordholm, S., Togneri, R., Toh, A. M. (2005). Spectral Entropy as Speech Features for Speech Recognition. PEECS. (Accessed on December 20, 2025). https://www.academia.edu/download/95656320/Spectral_entropy_as_speech_features_for_speech_rec.pdf

[7] Uppada, S. K. (2014). Centroid Based Clustering Algorithms- A Clarion Study. International Journal of Computer Science and Information Technologies, Vol. 5 (6), 7309–7313. (Accessed on December 20, 2025). https://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit2014050688.pdf

[8] Xiang, S., et al. (2008). Learning a Mahalanobis Distance Metric for Data Clustering and Classification. Tsinghua National Laboratory for Information Science and Technology (TNList). (Accessed on December 20, 2025). https://doi.org/10.1016/j.jesp.2017.09.011

STATEMENT

Hereby I declare that this paper that I have written is my own work, not a reproduction or translation of someone else's work and not plagiarized.

Bandung, 24 December 2025

Kevin Wirya Valerian
13524019