Overview

The primary focus of this study is to investigate the generality of current detection techniques in the field of speech deepfake detection without focusing on specific forgery attack methods

Evaluation Metrics

Equal error rate, tandem detection cost function.

EER is a performance metric calculated when the false acceptance rate (FAR) and false rejection rate (FRR) are equal, reflecting the robustness and balance of the detection model at equal EER values. For speech deepfake detection tasks, FAR represents the probability of fake speech being classified as real speech, and FRR represents the probability of real speech being classified as incorrect speech. When the detection score s and threshold θ of the forgery detection model are known, let Pfa θ and Pfr θ correspond to FAR and FRR at the threshold θ, respectively.

$$P_{\mathrm{fa}}(\theta) = \frac{N_{\mathrm{f}}(s > \theta)}{N_{\mathrm{f}}}$$

$$P_{\mathrm{fr}}(\theta) = \frac{N_{\mathrm{r}}(s \leq \theta)}{N_{\mathrm{r}}}$$

In the formula, Nf and Nr are the total number of samples of fake speech and real speech, respectively. Nf (s>θ) is the number of fake speech samples with a detection score s greater than the threshold θ. Similarly, Nr (s ≤ θ) is the number of real speech samples with detection scores s less than or equal to the threshold θ. Pfa θ is a monotonically decreasing function with respect to the threshold θ, while Pfr θ is a monotonically increasing function with respect to the threshold θ. So when the threshold is θ, the EER of the model is;

$$f_{\mathrm{EER}} = P_{\mathrm{fa}}(\theta) = P_{\mathrm{fr}}(\theta)$$

T-DCF evaluates the reliability of forgery detection models using minimum risk Bayesian decision-making. Based on this, t-DCF comprehensively considers the cost of misjudgment in different situations. The calculation process of t-DCF is as follows:

$$f_{\text{t-DCF}} = C_{\mathrm{fa}} \times (1 - P_{\mathrm{ta}}) \times P_{\mathrm{fa}}(\theta) + C_{\mathrm{fr}} \times P_{\mathrm{ta}} \times P_{\mathrm{fr}}(\theta)$$

Data Augmentation Techniques

These include speech noise addition，mask enhancement，channel enhancement，and compression enhancement，each aiming to simulate complex real-world acoustic environments more effectively.

1. Speech noise addition，which aims to interfere with the speech signal by adding noise to simulate the complex acoustic environment of a real scenario as much as possible.

2. Mask enhancement is the masking operation on the time or frequency domain of speech to achieve noise suppression and enhancement of the speech signal for improving the accuracy and robustness of speech detection techniques.
3. Transmission channel enhancement focuses on solving the problems of signal attenuation，data loss, and noise interference caused by changes in the codec and transmission channel of speech data.
4. Compression enhancement techniques address the problem of degradation of speech quality during data compression. In particular，the main data compression methods are MP3，M4A，and OGG.

Technique	**What It Does**	**Why It Helps**	**Example**
Speech Noise Addition	Adds background noise	Makes the model robust to real-world noise	Adding traffic noise to speech
Mask Enhancement	Hides parts of the speech in time or frequency	Helps with noise suppression and robustness	Muting some frequencies in a word
Transmission Channel Enhancement	Simulates bad audio transmission	Trains models to handle network distortion	Making speech sound like a lagging Zoom call
Compression Enhancement	Simulates compressed audio quality	Prepares models for low-quality recordings	Converting speech to low-quality MP3

Forgery Detection Feature Extraction Approaches

Speech deepfake detection aims to identify whether a voice recording is real or generated by AI. To achieve this, we analyze specific features of the speech signal. These features can be extracted using different methods, which can be grouped into four main categories:

Approach	How It Works	Strengths	Limitations
Handcrafted Feature-Based	Manually extracts speech features like MFCCs, pitch, and formants	Explainable, computationally efficient	Less effective against advanced deepfakes
Hybrid Feature-Based	Combines handcrafted and deep learning features	Balances interpretability and automation	Computationally demanding
Deep Feature-Based	Uses deep learning models to automatically extract speech patterns	Highly accurate against complex deepfakes	Requires large datasets and training power

End-to-End Based	Automatically extracts deep features directly from raw speech/audio.	Fully Automated & Highly Accurate	High Computational Cost
Feature Fusion-Based	Merges multiple feature types for better accuracy	Most robust approach	Complex and data-intensive

Best Approaches For Our Use Case

Data Augmentation Techniques

Since our Use case focuses on detecting the AI generated speech in real-time or near real-time, following data augmentation techniques are most suitable;

1. Speech Noise Addition

   AI-generated speech lacks background noise and adding environmental noise (e.g., café noise, traffic sounds, or microphone static) helps the model to differentiate between AI generated and human spoken speech.

2. Mask Enhancement (Time-Frequency Masking)

   AI-generated voices are smooth while the real voice of a human being contains frequency fluctuations. Masking certain frequency domains in the speech makes the model robust to differentiate between AI-generated and human spoken speech.

3. Compression Enhancement

   AI-generated speech is often higher quality than real-world phone calls or recordings that undergo compression. Training the model with compressed versions of speech files helps simulate real-world audio distortions from platforms like Zoom, WhatsApp, or phone calls.

Feature Extraction Techniques

1. End-to-End Forgery Detection (Best Overall)

- Key Technical Innovation:
  - Learns deep, task-specific features directly from raw audio without handcrafted processing.
  - Uses self-supervised learning (e.g., Wav2Vec 2.0, Conformer) for improved feature extraction.
- Reported Performance Metrics:
  - Accuracy: ~95-98% on benchmark datasets like ASVspoof and FakeAVCeleb.

    - Real-time Processing: Achieves low latency when optimized for streaming.
- Why This Approach is Promising:
    - Highly adaptable to different AI-generated voices.
    - Real-time potential with optimized architectures.
    - Avoids manual feature selection, improving generalization across datasets.
- Potential Limitations:
    - High computational cost, requiring powerful GPUs and large datasets.
    - Less interpretable compared to feature-engineered methods.

2. Feature Fusion-Based Forgery Detection

- Key Technical Innovation:
    - Combines handcrafted (MFCC, CQCC) and deep learning-based features (CNN, ResNet, Transformers).
    - Captures both low-level audio artifacts and high-level phonetic structures.
- Reported Performance Metrics:
    - Accuracy: 92-96% on benchmark datasets.
    - Robustness: Performs well across multiple forgery types (text-to-speech, voice cloning).
- Why This Approach is Promising:
    - Balances traditional and deep learning techniques, making it effective for AI speech detection.
    - More interpretable compared to purely deep learning-based methods.
    - Lower computational cost than end-to-end models.
- Potential Limitations:
    - Requires manual feature engineering, making it less flexible for novel AI-generated speech.
    - Feature selection process can introduce bias, limiting generalization.

3. Hybrid Feature-Based Forgery Detection

- Key Technical Innovation:
    - Uses selective handcrafted and deep features, reducing computational cost while maintaining accuracy.
    - Often integrates LSTMs or GRUs for temporal modeling of speech patterns.
- Reported Performance Metrics:
    - Accuracy: 90-94% on datasets like ASVspoof and LJSpeech.
    - Efficiency: More computationally efficient than end-to-end methods.
- Why This Approach is Promising:
    - Optimized for real-time and low-resource environments.
    - More explainable than end-to-end deep learning models.
    - Can be fine-tuned for specific forgery detection tasks.

- Potential Limitations:
    - May sacrifice some accuracy compared to feature fusion or end-to-end approaches.
    - Hard to balance handcrafted and deep features optimally for different datasets.

Best Three Forgery Detection Implemented Papers

1. How to Boost Anti-Spoofing with X-Vectors
    - It (Ma et al) uses Hybrid feature technique for feature extraction, exctracting features both manually and using a deep learning model.
    - Utilizes LFCC (linear frequency cepstral coefficients) and MFCC (Mel-Frequency Cepstral Coefficient) features, which capture both spectral and cepstral properties, making it robust against various forgery techniques.
    - TDNN (Time-Delay Neural Network) and SENet34 techniques are being employed, which are effective for sequence modeling and feature enhancement.
    - Uses LCML (Loss Component-based Metric Learning), which improves discrimination between genuine and AI-generated speech.
    - EER and t-DCF are being used as evaluation metrics, LA (Logical access) has been observed as 0.83% suggesting its better performance in detecting AI-generated speech with minimal errors.
    - The combination of LFCC, MFCC feature extraction techniques and TDNN, SENet34 models ensures robustness and making this implementation suitable for real-time or near real-time detection.
    - Although the paper highlights better performance of the proposed method in detecting AI generated speech, it requires high computational power, which could be a challenge for real-time deployment on edge devices.
    - For better generalization, the method requires fine-tuning on large dataset.

2. Spoofing Attacker Also Benefits From Self-Supervised Pretrained Models
    - It is an end-to-end forgery detection technique, extracting features using AI algorithms and techniques, eliminates the need for manual feature extraction.
    - This study (Ito et al) suggests that simply adopting pre-trained models like Wav2Vec2, HuBERT, or WavLM for detection might not be enough—additional adversarial training, fine-tuning, and countermeasures (e.g., spectral or phase-based analysis) may be required.
    - Low EER rate (LA : 0.44) suggests a strong trade-off between false positive and false negatives.
    - The use of differen pre-tranied models for feature extraction makes the method robust and has real-time detection potential by incorporating Conv-TesNet model.
    - Conv-TasNet and self-supervised models may require significant GPU resources for training and real-time inference.

3. Betray oneself: A novel audio deepfake detection model via mono-to-stereo conversion
    - The paper (Liu et al) proposed a novel fusion-based anti-spoofing approach using the AASIST model and M2S Transformer.
    - Optimized using the Cross-Entropy (CE) loss function.
    - Achieved state-of-the-art performance for LA (Logical Access, 1.34) attacks.
    - It employs Sinc filters, which are learnable convolutional filters designed to mimic traditional spectral analysis techniques.

References

1. Liu, Rui, et al. "Betray oneself: A novel audio deepfake detection model via mono-to-stereo conversion." arXiv preprint arXiv:2305.16353 (2023).
2. Ito, Aoi, and Shota Horiguchi. "Spoofing attacker also benefits from self-supervised pretrained model." arXiv preprint arXiv:2305.15518 (2023).
3. Ma, Xinyue, et al. "How to boost anti-spoofing with X-vectors." 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023.