https://doi.org/10.2352/EI.2025.37.3.MOBMU-310
© 2025, Society for Imaging Science and Technology

Detecting Voice Cloning and Text to Speech Audio in Real Time on Mobile Devices

Waldemar Berchtold, Julian Heeger, Simon Bugert, Martin Steinebach,; Fraunhofer Institute for Secure Information Technology SIT / ATHENE; Darmstadt, Hesse/Germany

Abstract

In this paper, we present a method that analyzes an audio stream in real time and provides an indication of whether the voice is synthetic generated by a voice clone or a text to speech model. Unlike state-of-the-art techniques that rely on self-supervised (SSL) or non-self-supervised learning, this method is deterministic and focuses on the analysis of tonal and non-tonal components within an audio stream. By leveraging principles from the MPEG-1 global masking threshold, the algorithm systematically evaluates tonal and noise components within a defined frequency range. The underlying hypothesis is that synthesized audio exhibits distinct tonal and non-tonal characteristics compared to original human speech, which can be quantified for classification.

This interpretable, deterministic framework addresses key limitations of existing SSL-based approaches, including high computational costs and limited transparency. Beyond detecting synthesized speech, the method provides insights into the likely model used for generation. Experimental evaluations demonstrate the algorithm's effectiveness, revealing distinct and consistent patterns across various TTS and voice conversion (VC) models, thereby offering a reliable and computationally efficient solution for audio authenticity verification. The proposed algorithm is developed and tested on a small dataset and show an excellent separation between different solution providers and genuine voices.

Introduction

Recent incidents involving identity attacks through audio manipulation have become increasingly prevalent. For instance, the New York Post reported on an incident where a 15-year-old girl's voice was cloned in an extortion attempt, with the perpetrators demanding a million-dollar ransom from her mother[1].

Recent advancements in artificial intelligence (AI) have enabled cybercriminals to employ voice cloning techniques, leading to a novel class of scams. According to a McAfee report, with as little as three seconds of audio, malicious actors can replicate an individual's voice to craft deceptive messages delivered via voicemail or voice messaging services.

A global survey conducted by McAfee, encompassing 7,000 participants, revealed that approximately 25% had either encountered an AI voice cloning scam personally or knew someone who had. Of the people who reported losing money, 36% said they lost between $\$500$ and $\$3,000$, while 7% got taken for sums anywhere between $\$5,000$ and $\$15,000$[2]. This underscores the growing prevalence of such fraudulent activities.

The implications of AI-driven voice cloning are profound, as they facilitate more convincing impersonation attacks, thereby increasing the likelihood of successful deception. This development necessitates heightened awareness and the implementation of robust security measures to mitigate the risks associated with these sophisticated scams.

State of the Art

Various methods exist for detecting manipulated voice recordings. These detection methods can be categorized into handcrafted feature-based approaches, deep learning techniques, end-to-end models, self-supervised learning (SSL) approaches, and further innovative strategies.

Handcrafted Feature-Based Detection Traditional detection techniques rely on handcrafted acoustic features such as linear frequency cepstral coefficients (LFCC), mel-frequency cepstral coefficients (MFCC), and constant Q cepstral coefficients (CQCC), often processed using machine learning classifiers like Gaussian Mixture Models (GMM) and Support Vector Machines (SVM)[8], [15]. Studies indicate that CQCC-based features offer superior performance among handcrafted features. Feature-based approaches such as those evaluated by [11] using SVM, XGBoost, and Random Forest classifiers, show limited generalizability to unknown attacks. These models achieve moderate accuracy (e.g., SVM with validation accuracy of 0.85 and test accuracy of 0.67), while deep learning models such as Temporal Convolutional Networks (TCN) outperform them significantly.

Deep Learning Approaches Deep learning models using spectrogram-based representations such as mel-spectrograms and short-time Fourier transform (STFT) have demonstrated superior detection capabilities. ResNet-34[3], EfficientCNN[14], and RES-EfficientCNN[9] models trained on the ASVspoof dataset achieved low Equal Error Rates (EERs), indicating effective spoofing detection. DeepSonar[17] introduces an alternative approach by monitoring deep neural network (DNN) neuron behavior, achieving high robustness against various distortions. However, vulnerability to environmental noises such as wind and rain remains a challenge.

[1] https://nypost.com/2023/04/12/ai-clones-teen-girls-voice-in-1m-kidnapping-scam/

[2] https://www.mcafee.com/blogs/privacy-identity-protection/artificial-imposters-cybercriminals-turn-to-ai-voice-cloning-for-a-new-breed-of-scam/

End-to-End Models End-to-end detection models process raw audio waveforms directly without requiring feature extraction. Examples include Res-TSSDNet[8], RawNet2[15], RawGAT-ST[16], and AASIST[10]. RawNet2, an extension of the RawNet architecture, employs a convolutional structure and achieves a competitive EER of 3.15% on ASVspoof 2019. Spectro-temporal graph attention networks such as RawGAT-ST significantly enhance spoofing detection, outperforming earlier models by combining spectral and temporal information. AASIST further improves detection rates by employing heterogeneous graph structures, reducing EERs to 0.83%.

Self-Supervised Learning (SSL) SSL-based approaches leverage pre-trained speech models such as Wav2Vec 2.0[13][5], HuBERT, and WavLM[7] to enhance robustness. Fine-tuning these models with spoofed and bona fide speech data significantly improves detection accuracy. Studies by [18] confirm that SSL-based models outperform conventional feature-based classifiers in generalization to unseen attacks, demonstrating lower error rates across various datasets.

While significant advancements have been made in detecting manipulated voice recordings, a major challenge remains: generalizability. Most existing methods rely on ASVspoof datasets, limiting their applicability to real-world scenarios. Future research should focus on developing robust countermeasures that generalize across datasets, attack types, and languages.

Requirements

In order to establish a real-time capable recognition system for a smartphone that can provide an assessment on short sequences of a few minutes, there are a number of requirements.

Real-Time Capability The algorithm must be capable of detecting synthetic speech with minimal delay to support real-time applications, such as fraud detection in phone calls, live-stream moderation, and security in voice authentication systems. Achieving this requires efficient signal processing techniques and optimized neural architectures that can process audio streams instantaneously. The detection system should integrate seamlessly into existing communication infrastructures without introducing significant latency. A delay in decision-making could compromise security-sensitive applications, making immediate response times essential for practical deployment.

Low Computational Complexity To ensure broad applicability, the detection algorithm must be computationally efficient, enabling deployment on a wide range of devices, including embedded systems, mobile platforms, and cloud-based architectures. Deep learning-based approaches should prioritize lightweight architectures, using quantization, pruning, or knowledge distillation techniques to minimize computational overhead. High complexity can hinder real-world adoption, particularly in resource-constrained environments where real-time detection is necessary. Scalability is also a key consideration, as the algorithm should be able to analyze large volumes of audio data without excessive power consumption or infrastructure costs.

Explainability A major challenge in synthetic speech detection is ensuring that the model's decisions are interpretable. Explainability is crucial for building trust among users, regulators, and stakeholders, particularly in forensic and legal applications. The algorithm should provide insights into its decision-making process, such as identifying key acoustic or prosodic features indicative of synthetic speech. Methods such as attention heatmaps, feature importance visualization, and interpretable model architectures should be integrated to allow users to verify why a particular voice sample was classified as synthetic. Without explainability, even highly accurate systems may face skepticism and resistance to adoption.

Reliable Decision on Short Audio Sequences Synthetic speech detection must be robust even when analyzing brief audio clips. In many practical scenarios, such as scam call prevention and real-time authentication, only a few seconds of audio are available for analysis. The algorithm must extract key distinguishing features from short utterances, distinguishing human speech from AI-generated audio with high confidence. Techniques such as frame-level analysis, phoneme-level anomaly detection, and spectral consistency checks can improve performance under short-duration constraints. Ensuring reliable detection in minimal time is critical to maintaining security and usability in real-world applications.

Likelihood Estimation of Synthetic Speech Instead of providing a binary classification (synthetic vs. real), the algorithm should output a likelihood score indicating the probability that a given voice sample is artificially generated. This probabilistic approach allows for more flexible decision-making, enabling adaptive security thresholds and risk-based assessments. For instance, high-confidence cases can trigger automated responses, while lower-confidence cases may require additional verification. Providing likelihood estimations also enhances interpretability, allowing users to understand the degree of certainty in the model's predictions and adjust detection sensitivity based on operational requirements.

Method

In speech production, the balance between tonal and noise components significantly influences the perceived naturalness and expressiveness of a voice. A higher proportion of noise components results in a more monotonous voice, characterized by reduced variation in articulation. This monotonicity stems from the diminished harmonic structure, leading to less dynamic spectral changes across phonemes.

Furthermore, phoneme realization is inherently variable. Each phoneme is pronounced differently depending on linguistic and contextual factors, such as the specific word in which it appears and its position within a sentence. This variability is crucial for natural-sounding speech, as it reflects the continuous adaptation of articulation based on coarticulatory effects, prosodic cues, and speaker intent. The suppression of this variability, either due to increased noise components or artificial processing constraints, may lead to robotic or unnatural speech synthesis, which lacks the fluidity and expressiveness of human speech.

The detection algorithm processes incoming audio streams in real time on a smartphone, applying a series of transformations and statistical analyses to differentiate between synthetic and natural speech. The procedure follows a structured sequence of operations to ensure accurate classification.

First, the incoming audio stream is captured continuously, with each segment of data processed in discrete frames. To facilitate frequency-domain analysis, each frame is segmented into 2048-sample windows, which are then transformed into the spectral domain using a Fast Fourier Transform (FFT). This step enables the extraction of key acoustic characteristics necessary for distinguishing synthetic from real speech.

Following the transformation, tonal and noise components are extracted from each frame. The tonal components refer to well-defined harmonic structures present in natural speech, whereas noise components correspond to broadband or non-harmonic spectral energy. By isolating these two signal characteristics, the algorithm can analyze the underlying structure of the speech sample.

Once extracted, the algorithm counts the number of tonal and noise components per frame, providing a frame-wise distribution of signal characteristics over time. This count allows for statistical assessment of how structured (tonal) or unstructured (noisy) the signal appears within a short-term analysis window.

To achieve a more robust classification, the algorithm then compares the relation between the mean and standard deviation of noise to tonal components over a 10-second segment of audio. By computing these statistical measures over a longer period, the system accounts for variations in speech dynamics and ensures that transient noise artifacts do not unduly influence the classification. Synthetic speech, which typically exhibits more uniform spectral characteristics due to its generation process, often shows distinct statistical patterns compared to natural speech, where phonemes exhibit dynamic variation.

Finally, a decision is made based on predefined thresholds. These thresholds define acceptable ranges for the statistical relationship between tonal and noise components, allowing the system to determine whether the analyzed speech sample is likely synthetic or real. If the computed measures fall within the predefined range for natural speech, the sample is classified as real. Conversely, if deviations indicative of synthetic generation are detected, the sample is flagged accordingly.

This structured approach ensures that the system remains efficient and interpretable while maintaining high reliability in distinguishing synthetic speech from human-generated audio.

Implementation

The method is related to the global masking of the MPEG-1 model. The global masking is determined by accumulating the impact of tonal maskers, noise maskers and the hearing threshold in silence. The proposed algorithm simply counts the tonal and noise components with a specific width in a specific frequency range for each frame. The fundamental hypothesis is that the tonal and non-tonal components of synthesized audio and original audio differ in count and appearance.

To utilize this, we use the mean of the count and the deviation of both the tonal and noise components only in the range of 0-4000Hz. The determination of the global threshold is based on the following equation:

$$\mathrm{v}_f(z_{i,j}) := \begin{cases} b(z_{i,j}+1)-(aX_j+6)+c, & d <= z_{i,j} < -1 \\ (aX_j+6)z_{i,j}+c, & -1 <= z_{i,j} < 0 \\ -bz_{i,j}+c, & 0 <= z_{i,j} < 1 \\ -(z_{i,j}-1)(b-a/2X_j)-b-c, & 1 <= z_{i,j} < e \end{cases} \tag{1}$$

where a=0.4, b=17, c=0, d=3, e=8.These values were selected as proposed in the original method in ISO/IEC 11172 [2], and are applied to both tonal and non-tonal components. The determination of tonal and non-tonal components happen as described in the MPEG-1 standard.

To do so, the proposed algorithm transforms a frame of 2048 samples to the power spectrum. The power spectrum is then used to determine tonal and non-tonal components as proposed in ISO/IEC 11172. Afterward, we count the number of tonal and non-tonal components that fulfill the properties of equation 1 for each frame. The algorithms next calculate the mean and standard deviation of the counted tonal and non-tonal components.

In the last step the algorithm uses the six resulting values, namely the mean count of tonal m_{tonal} and non-tonal $m_{non-tonal}$ components, the relation of both to each other $p_m = m_{non-tonal}/m_{tonal}$ as well as the standard deviation of the tonal s_{tonal} and non-tonal $s_{non-tonal}$ components and the relation of both to each other $p_s = s_{non-tonal}/s_{tonal}$.

Evaluation

In the following, we will discuss the test setup and the evaluation results.

Test Setup

For the test setup we use 6 different TTS and VC models beside genuine data. The audio is trimmed to 10 seconds audio play length. We tested 10 different speakers and 10 different text passages of 10 different books resulting in 1000 files for each model as well as for the genuine data. The different tested models are the TTS from *ElevenLabs*[3] and *XTTS*[6] as well as the voice conversion model from *ElevenLabs*, *FreeVC*[12], *KnnVC*[4] and *QuickVC*[1].

Test Results

The initial results of the proposed algorithm are promising. The findings further indicate that it is possible to infer which model was most likely used. Figure 1 show one dot for each 10 second sample. The x-axis gives the relation of the non-tonal to tonal components $p_m = m_{non-tonal}/m_{tonal}$ and the y-axis the relation of the standard deviation from the non-tonal to tonal components $p_s = s_{non-tonal}/s_{tonal}$. The genuine data are around 36 on the y-axis and 96 on the x-axis. All models show different values for p_m and p_s except *QuickVC* and *KnnVC* have overlapping values.

The proposed approach enables a user to reliably determine whether the caller is a genuine individual or if an identity attack is occurring, based on 10 seconds of spoken audio. The proposed approach is faster than real time and run smoothly on a smartphone.

The authors acknowledge that the disclosure of such forensic methods may lead to targeted improvements in models, potentially rendering the proposed feature ineffective in reliably distin-

[3] elevenlabs.io

guishing genuine voices from voice conversion (VC) or text-to-speech (TTS) models in the future.

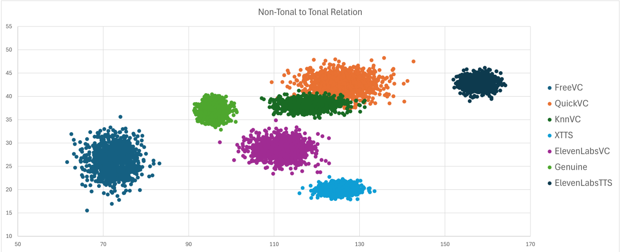

Figure 1: Each dot represents a 10-second sample. The x-axis gives the relation of the non-tonal to tonal components $p_m = m_{non-tonal}/m_{tonal}$ and the y-axis the relation of the standard deviation from the non-tonal to tonal components $p_s = s_{non-tonal}/s_{tonal}$.

Conclusion and Future work

The results of this study demonstrate the effectiveness of the proposed real-time algorithm for distinguishing synthetic from genuine speech based on tonal and non-tonal component analysis. By leveraging spectral-domain transformations and statistical measures, the approach provides a robust and interpretable method for identifying AI-generated speech, achieving real-time performance on a smartphone.

Experimental evaluation across multiple text-to-speech (TTS) and voice conversion (VC) models indicates clear separability between genuine and synthetic speech in most cases. The algorithm successfully captures differences in the spectral characteristics of synthetic and natural speech, enabling classification within a 10-second window. The distinct clustering of various synthetic models in the feature space suggests that not only can synthetic speech be detected, but also the specific model used for generation can be inferred. However, some overlap between certain models, such as QuickVC and KnnVC, suggests that further refinements may be necessary for more precise classification.

A key advantage of this method lies in its deterministic nature, offering interpretability over deep learning-based approaches, which often lack transparency. Furthermore, its computational efficiency enables deployment on mobile devices, making it a practical tool for real-time authentication and identity verification.

Despite these promising results, the authors acknowledge the potential for adversarial adaptation. The public disclosure of such forensic techniques may incentivize AI developers to modify their models to evade detection. This underscores the need for ongoing research into adaptive forensic techniques that can evolve alongside advancements in synthetic speech synthesis. Future work should explore more refined statistical features and hybrid approaches that integrate machine learning while maintaining explainability and efficiency.

The approach must be evaluated on a larger database and possibly further optimized.

Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Quickvc: Any-to-many voice conversion using inverse short-time fourier transform for faster conversion, 2023.

[2] ISO/IEC 11172. Information technology - coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s - part 3: Audio. In *Standard and International Organization for Standardization, Geneva, CH*.

[3] PR Aravind, Usamath Nechiyil, Nandakumar Paramparambath, et al. Audio spoofing verification using deep convolutional neural networks by transfer learning. *arXiv preprint arXiv:2008.03464*, 2020.

[4] Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors, 2023.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[6] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. Xtts: a massively multilingual zero-shot text-to-speech model, 2024.

[7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[8] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters*, 28:1265–1269, 2021.

[9] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

[10] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

[11] Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar, and Faruk Kazi. A deep learning framework for audio deepfake detection. *Arabian Journal for Science and Engineering*, pages 1–12, 2021.

[12] Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion, 2022.

[13] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[14] Nishant Subramani and Delip Rao. Learning efficient representations for fake speech detection. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 34, pages 5859–5866, 2020.

[15] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

[16] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 1–8, 2021.

[17] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1207–1216, 2020.

[18] Xin Wang and Junichi Yamagishi. Investigating self-supervised front ends for speech spoofing countermeasures. *arXiv preprint arXiv:2111.07725*, 2021.

Author Biography

Waldemar Berchtold has headed the Multimedia Security research group since 2022 at the Fraunhofer Institute for Secure Information Technology (SIT) and is a researcher at the National Research Center for Applied Cybersecurity (ATHENE) in Darmstadt, Germany. He received his diploma in mathematics in 2008 and his Ph.D. in 2022 at TU Darmstadt. The focus of his research is in various areas of multimedia security for authenticity and integrity proof and digital watermarking. He has led numerous projects in the field of media security with a focus on audio and video.

Julian Heeger is a research associate in the Media Security and IT Forensics department at the Fraunhofer Institute for Secure Information Technology (SIT) and a researcher at the National Research Center for Applied Cybersecurity (ATHENE) in Darmstadt, Germany. He holds a Master's degree in IT security from the Technical University of Darmstadt.

Simon Bugert received his Master's degree in computer science from the Technical University of Darmstadt, Germany in 2021. Since then, has been a research associate in the Media Security and IT Forensics department at the Fraunhofer Institute for Secure Information Technology (SIT) and at the ATHENE National Research Center for Applied Cybersecurity.

JOIN US AT THE NEXT EI!

Imaging across applications . . . Where industry and academia meet!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org