OPEN ACCESS

Volume: 11

Special Issue: 3

Month: July

Year: 2024

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Received: 16.05.2024

Accepted: 20.06.2024

Published: 08.07.2024

Citation:
Bhaskaran, N.A., et al. "Detecting Deep Fake Voice Using Machine Learning." *Shanlax International Journal of Arts Science and Humanities*, vol. 11, no. S3, 2024, pp. 53–60.

DOI:
https://doi.org/10.34293/sijash.v11iS3-July.7919

Detecting Deep Fake Voice using Machine Learning

N. A. Bhaskaran
Associate Professor, Department of AI & DS, Arjun College of Technology Coimbatore

M. Srinadh
Department of AI & DS, Arjun College of Technology, Coimbatore

K. Dhamodhar
Department of AI & DS, Arjun College of Technology, Coimbatore

Mani Maran R
Department of AI & DS, Arjun College of Technology, Coimbatore

Abstract

With the rapid advancement of deep learning techniques, the creation of synthetic media, particularly deep fake voices, has become increasingly sophisticated and accessible. This poses significant challenges in maintaining trust and authenticity in audio-based content. In response, this project proposes a machine learning-based approach to detect deep fake voices. The project begins by curating a diverse dataset consisting of genuine and deep fake voice samples, covering various demographics, accents, and emotional expressions. Pre-processing techniques are applied to clean and standardize the audio data, followed by feature extraction to capture relevant characteristics of the voice signals. For model development, a Convolutional Neural Network (CNN) architecture augmented with recurrent layers is employed, leveraging its ability to learn spatial and temporal features from the spectrogram representations of the audio. The model is trained on the prepared dataset using categorical cross-entropy loss and optimized through backpropagation.Evaluation of the trained model is conducted on a separate test set, measuring performance metrics such as accuracy, precision, recall, and F1-score. Post- processing methods, including thresholding and smoothing, are applied to refine the model's predictions and enhance robustness. The proposed approach offers a promising framework for detecting deep fake voices in audio content, contributing to the ongoing efforts to combat the spread of misinformation and preserve the integrity of digital media. However, ongoing research and collaboration across disciplines are essential to address emerging challenges and ensure the responsible deployment of deep fake detection technologies.

Introduction

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 Deep Fake Voice Conversion.

To address the above emerging issues, we are introducing the DEEP-VOICE dataset. DEEP-VOICE is comprised of real human speech from eight well-known figures and their speech converted to one another using Retrieval-based Voice Conversion.

For each speech, the accompaniment ("background noise") was removed before conversion using RVC. The original accompaniment is then added back to the Deep fake speech:
In recent years, the proliferation of deep fake technology has raised significant concerns regarding its potential misuse, particularly in the realm of voice manipulation. Deep fake voices, generated using advanced machine learning algorithms, can convincingly mimic real voices, presenting serious threats to various sectors such as cybersecurity, media integrity, and personal privacy. As deep fake technology becomes increasingly sophisticated and accessible, the need for robust detection methods becomes paramount.

Example 1: SecurityRisksinAudioAuthentication Systems with the rise of voice-based authentication systems in sectors like banking, telecommunications, and government agencies, the threat posed by deep fake voices cannot be overstated. Imagine a scenario where an individual's voice is seamlessly replicated by a malicious actor using deep fake technology, thereby gaining unauthorized access to sensitive information or financial resources. The potential ramifications of such security breaches underscore the urgency of developing effective mechanisms to detect and mitigate deep fake voices.

Example 2: Media Integrity and Misinformation in an era characterized by the rampant spread of misinformation and fake news, deep fake voices pose a grave threat to media integrity. Imagine a world where political leaders, celebrities, or public figures can be made to say anything through manipulated audio recordings. These fabricated audio clips, if not properly identified and debunked, could significantly erode public trust and exacerbate societal divisions.

Detecting deep fake voices in audio recordings is thus essential for preserving the authenticity and credibility of media content.

Example 3: Personal Privacy and Consent Deep fake technology raises serious concerns regarding personal privacy and consent, particularly in the context of voice recordings.

With the ability to generate highly realistic voice replicas, individuals may find themselves unwittingly implicated in false or compromising scenarios. For instance, deep fake voices could be used to create fake audio evidence of incriminating conversations or illicit activities, potentially tarnishing one's reputation or causing irreparable harm. Safeguarding personal privacy and ensuring informed consent necessitate the development of reliable methods for detecting and verifying the authenticity of voice recordings.

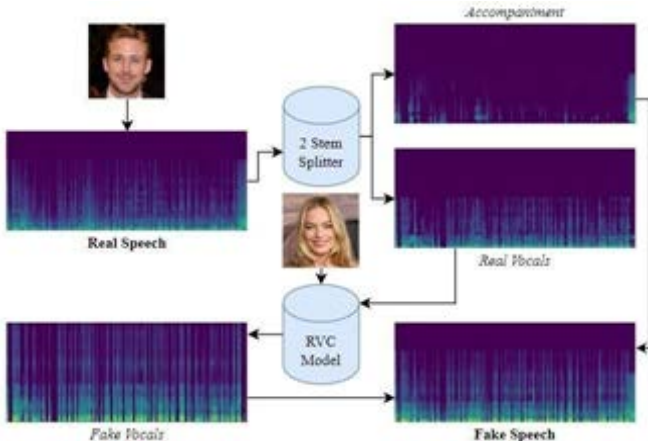
In light of these challenges, machine learning techniques offer promising avenues for detecting deep fake voices. By leveraging advanced algorithms and large datasets of authentic and manipulated audio samples, researchers can train models to discern subtle inconsistencies and artefacts characteristic of deep fake voice synthesis. Through interdisciplinary collaboration and ongoing innovation, we can strive to stay one step ahead of malicious actors and safeguard the integrity of audio content in an increasingly digitized world.

Literature Survey

A comprehensive literature survey reveals a growing body of research focused on detecting deep fake voices using machine learning techniques. Studies have explored various approaches, including feature-based analysis, deep neural networks, and spectrogram-based methods. Researchers have investigated the effectiveness of different algorithms in distinguishing between genuine and manipulated audio recordings. Additionally, efforts have been made to create benchmark datasets for training and evaluating detection models, facilitating advancements in this

critical field. By synthesizing findings from these studies, researchers gain insights into the state-of-the-art techniques and emerging challenges in deep fake voice detection.

Network and Connected to the Auto Power Backup



With technological products such as Google Audio LM it is possible to produce realistic sounds, well-structured, and consistent sound sequences. While this technology helps with many issues such as speech disorders, it is also possibleto use it maliciously such as information pollution, phishing, and telephony scams. In our study, we tried to detect audio sequences produced with Google Audio LM.

This study is about detecting the imitated sound sequences produced later by using feature extraction, machine learning, and deep learning methods and being able to tell the difference from the original sound sequences. Our proposal can be used in many areas. For example;

It can be used on social media platforms to prevent information pollution, fake news, and destruction, and to prevent the manipulation of speeches of important people, with warnings for video and audio resources uploaded to the platforms.

In the field of cyber security, it may be possible to prevent phone scams, detect phishing attacks, prevent identitytheft, and detect whether malware is installed in audio files. In mobile phones, it is possible to analyse incoming calls and protect privacy by developing models that can work on mobile phones without the need for a cloudenvironment.

There are audio deep fake detection solutions and they deter- mine whether the given audio is real or fake. For this, machine learning and deep learning models have been developed. Whiledeveloping these models, the features used come to the fore. The features used by the models are divided into low-leveland high-level features. In the low- level feature extraction ap- preach, artifacts and traces are searched during the production of the audio file. The high- level feature extraction approach focuses on features such as semantic content. In addition, values such as Mean Opinion Score (MOS) value, which indicates the quality of the sound produced, Sam-piling rate (Sampling Rate - can vary from 8 kilohertz (kHz)to 48 kilohertz (kHz), mono and stereo), are used as features. Features such as languages spoken, words used, intonations, and accents that show the way words are pronounced are also used.

The general framework used for fake audio detection is as shown in Figure 1. At the Audio Feature Extractor step, the above-mentioned features are obtained. In the Training Process step, many machine learning and deep learning models are used. In the last step, the prediction probability is calculated and it is decided whether the audio is fake or real.

This study consists of the following sections; the Next section gives related work. Section III presents the proposed detection model. In Section IV, we shared experimentation results for the proposed model. The study ends with Section V.

Methodological Diversity: The literature survey underscores the diverse range of methodologies employed in deep fake voice detection research. Techniques include traditional signal processing methods, machine learning algorithms such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), as well as novel approaches like adversarial training.

Diagram

Background Information

The Mel-Frequency Cepstral Coefficients (MFCC) constitute a logarithmic scale-adjusted and non-linear representation of the power spectrum inherent in acoustic signals, as expounded upon by Bodda parietal. [15]. The Mel frequency scale serves as a perceptual model of sound frequency variations, mirroring the human auditory system's response to sound. It was conceived based on insights from the principles governing human auditory perception. By adopting the Mel frequency scale, the method extracts salient acoustic character- is tics, leading to the derivation of cepstral coefficients. These coefficients bear substantial significance in formant estimation,which is the process of identifying the resonant frequencies in human vocalization

Futures

When discussing models for detecting deep fake voices, it's essential to consider various approaches that have been explored in the literature. Here are some key models used in this context:

Convolutional Neural Network (CNNs)

CNNs have been widely used for detection tasks due to their ability to capture spatial dependencies in audio spectrograms. Researchers have designed CNN architectures tailored for audio processing, leveraging techniques such as 1D convolutions and pooling layers to extract discriminative features from audio signals.

Recurrent Neural Networks (RNNs)

RNNs, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are effective for capturing temporal dependencies in sequential data, making them suitable for analysing audio signals over time.

These models are adept at learning long-range dependencies and have been applied to tasks such as voice activity detection and emotion recognition, which are relevant for deep fake detection.

Generative Adversarial Networks (GAN)s

Gags consist of a generator and a discriminator network trained in an adversarial fashion. In the context of deep fake detection, GANs have been used to generate synthetic audio samples for

training detection models, as well as to create adversarial attacks against existing detection systems to assess their robustness.

Auto Encoders

Auto encoders are unsupervised learning models that aim to reconstruct input data at the output layer.

In deep fake detection, auto encoders can be trained to reconstruct genuine audio samples, enabling them to identify anomalies in manipulated audio signals that deviate from learned representations of authentic audio.

Attention Mechanism

Attention mechanisms allow models to focus on relevant parts of the input data while ignoring irrelevant information. These mechanisms have been incorporated into deep fake detection models to improve their interpretability and focus on discriminative features in audio signals

Proposed Audio Deep Fake Detection Model

Frequency is a scale that shows how the human ear perceives the change in sound frequencies. These obtained features are then classified using machine and deep learning based models. The decision is made based on the majority of the results obtained.

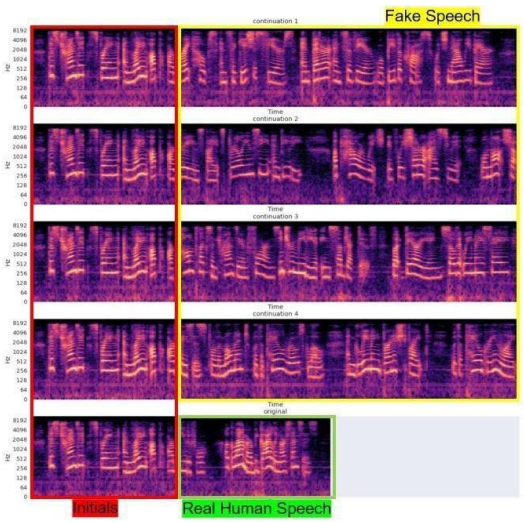


Figure 2 shows the solution for obtaining features using MFCC and detecting fake and real audios. In the Audio Object Transformation Module, fake and real audio files are convertedinto audio objects using the Librosa library [16], and each audio recording is divided into small 2-3 second slices.

In the MFCC Attribute Extraction Module, the MFCC method is used to extract the attributes of audio objects. At this stage, using windowing methods such as Hamming, Blackman,Gaussian, rectangular and triangular windowing methods, the audio streams are reshaped into smaller windows and divided into frames. In the next step, MFCC coefficients are calculated using fast Fourier transform, logarithm and discrete cosine transform and 40 MFCC feature data are obtained from each audio object.

In the Audio Deep fake Analysis and Detection Module,deep learning algorithms such as Convolutional Neural Net- work (CNN) and Multiple Layer Perceptron (MLP) and ma- chine learning algorithms such as Support Vector Classifier (SVC), Decision Tree, Ada Boost, Random Forest, Extra Trees, Gradient Boosting, K- Neighbors, Logistic Regression, Linear Discriminant Analysis are used by majority voting, fake or real voice detection is made.

The Audio Deep Fake Notification Module manages the re- sults from the Audio Deep Fake Analysis and Detection Moduleand notifies different platforms such as cloud and mobile in different ways.

Experimentation

Dataset

As dataset, we used 28 different speech data [3], [4]. Eachspeech consists of a 6-second original speech file, a 3-second prompt, and 4 10-second produced continuation audio files. There are 168 raw sound file samples in total.

Figure 3 shows the audio sample structure we used. Audio files are converted into Librosa sound objects. beginningsof the continuation speeches are removed and the remaining are used as fake class. All recordings are split into 2-second chunks. MFFC features are generated. After all, we enriched them to 448 samples and used these files in our tests. Figure shows an example of enriched data for speech data.

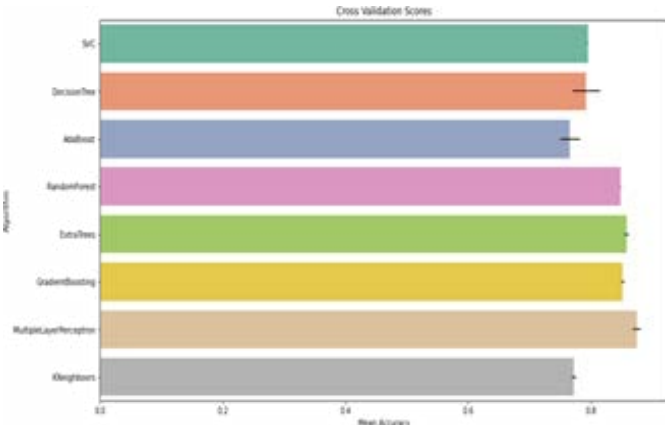


Figure 3 Cross Validation Scores

A. Testing Results And Evaluation

Noteworthy were the ExtraTreesClassifier and GradientBoostingClassifier, which demonstrated robust and steady performances, with accuracy values of 0.85 to 0.86 and 0.84 to 0.85, respectively. Ad- dictionally, the MLPClassifier emerged as the top-performing model, showcasing impressive accuracy between 0.86 and 0.88, highlighting its proficiency in discerning intricate data patterns. All models were trained using their default parameters with Sci-kit learn package in Python.

Conclusion

While synthetic speech technologies allow useful scenarios such as voice assistant and voice device control, it can also lead to abuse scenarios of these systems such as fraud and manipulation of information. Robust audio deep fake detection systems are needed to prevent abuse of artificial speech generators and to protect social benefits.

In this study, an MFCC-based detection system is proposed for audio deep fake detection. Using MFCC feature data, real and fake speeches were tried to be detected using 8 different classification methods, and results were obtained in the accuracy range between 75% and 88%.

In future studies, we will focus on the optimization of algorithm parameters and usage areas in the field of cyber security. Also, we aim to test and develop the model under various distortions such as echo, background noise, clipping, and speaker environment effects.

Acknowledgment

This study was supported in part by EUREKA cluster in this study, eight distinct machine-learning models were evaluated through 10 iterations to assess their accuracy in a supervised learning context. The models analyzed encom- passed a diverse set of algorithms, including Support Vector Classifier (SVC) DecisionTreeClassifier,AdaBoost Classifier,RandomForestClassifier,ExtraTreesClassifier,Gradi entBoost- ingClassifier,MLPClassifier(MultiLayPerceptron),and KNeighborsClassifiers. As shown in Table I, among the models scrutinized, the SVC demonstrated consistent accuracy, yielding 0.79.

	data	data_len	folder_id	type	duration
0	[0.0023753168, 0.0030640871, 0.0024172754, 0.0...	44100	1	continuation	2.0
1	[-0.0271832, -0.014532235, 0.023155987, 0.0004...	44100	1	continuation	2.0
2	[-0.012091764, -0.0068453695, -0.0051356293, -...	44100	1	continuation	2.0
3	[-0.025188593, -0.060285963, -0.053796634, -0....	44100	1	continuation	2.0
4	[0.0015387082, -0.0023384176, -0.00046688275, ...	44100	1	continuation	2.0
5	[-0.08238074, 0.096537635, 0.09141776, 0.00068...	44100	1	continuation	2.0
6	[0.051445264, 0.011930261, 0.0055687227, 0.014...	44100	1	continuation	2.0
7	[-0.096897, -0.068347156, -0.04082616, -0.0205...	44100	1	continuation	2.0
8	[-0.0077326, -0.004442761, -0.0011035203, -0.0...	44100	1	continuation	2.0
9	[-0.008052021, -0.006326882, 0.0028287238, 0.0...	44100	1	continuation	2.0
10	[-0.10079528, -0.11480869, -0.11515833, -0.104...	44100	1	continuation	2.0
11	[-0.0004452407, -0.001405702, -0.0034516295, -...	44100	1	continuation	2.0
12	[-0.0009499488, -0.0013900853, -0.0015294905, ...	44100	1	original	2.0
13	[-0.015314762, -0.01616377, -0.016596604, -0.0...	44100	1	original	2.0
14	[0.0023035882, 0.0009683671, -5.8044283e-05, -...	44100	1	original	2.0
15	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	44100	1	prompt	2.0

Table Accuracy Ranges of Classification Results

Classifier	Accuracy
SVC	Ranges
Decision Tree	[0.79 - 0.79]
Classifier hjbsh	[0.76 - 0.81]
AdaBoost Classifier	[0.75 - 0.78]
Random Forest Classifier	[0.84 - 0.84]
Extra Trees Classifier	[0.85 - 0.86]
Gradient Boosting Classifier	[0.84 - 0.85]
MLP Classifier	[0.86 - 0.88]
KNeighbors Classifier	[0.77 - 0.76]

ITEAproject VESTA which is also supported by TUBITAK (The Scientific and Technological Research Council.

References

1. Korshunov, Pavel, and S. Marcel. "Deepfake detection using inverse contrastive loss." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 3920-3929. 2020.
2. Mukhopadhyay Rudrabhaetal. "A comprehensive survey of voice conversion and deep fake techniques." arXivpreprintarXiv:2103.03230 (2021).
3. Xie, Jin, et al. "Voice Deep Guard: Towards Intelligent "VoiceDeepfakeDetection." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 4471-
4. Prakash, Shreya, et al. "A comprehensive study on deep fake audio detection." In Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security, pp. 81-90. 2020.
5. Li, Yang, et al. "Universal voice conversion." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3526-3535. 2019.
6. Pasupathi Panupong, and Taxing Li. "Detecting AI- Generated Text with BERT." In Proceedings of the 2021 Conference oftheNorthAmerican Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 672-684. 2021.