MFCCs, Chroma Features, and Spectrogram Images for Deepfake Audio Classification Using Machine Learning

Nora Bakken, Shavil Singh, Makwana Prashant Lnu, and Tapadhir Das
Department of Computer Science, University of the Pacific
Stockton, California, USA
n_bakken@u.pacific.edu,tdas@pacific.edu

ABSTRACT

CCS CONCEPTS

• **Mel-frequency cepstral coefficients (MFCCs)** → *Audio representation*; • **Convolutional Neural Networks** → *Machine Learning*; • **Adversarial attack protection** → *Cybersecurity*.

KEYWORDS

MFCCs, Spectrogram, VGG16, ResNet50, Chroma Features, SVM, Gradient Boosting, Deepfakes

ACM Reference Format:
Nora Bakken, Shavil Singh, Makwana Prashant Lnu, and Tapadhir Das. 2024. MFCCs, Chroma Features, and Spectrogram Images for Deepfake Audio Classification Using Machine Learning. In *Proceedings of Conference name and year (Conf 2024).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

2 MOTIVATION

With the rise of artificial intelligence (AI), deepfakes are more prevalent than ever, bringing with them a slough of potential dangers in a variety of areas. Likely the most well-known effect of deepfakes in everyday life is the rise of false media, especially targeting individuals.

The American Bar Association highlighted a targeted defamation attack involving an audio recording with the voice of a high school principal making racist and antisemitic comments. After the recording spread throughout the school, the principal was in danger of losing his livelihood. He denied making these comments. After a thorough investigation, the local police deemed the recording to have been manipulated using AI [Jr. 2024].

The less talked about consequence of accessible and easy to create deepfakes is the rise of non-consensual explicit deepfake attacks on individuals. These attacks, along with being traumatic to the individual, are expanding the ever-present gender gap at the global level, inflicting consequences at the societal level [Kim 2024].

In the recent United States election, nearly half of American voters stated that deepfakes had an influence on their ballots [Genovese 2024]. Looking at this problem from a purely monetary perspective, CFO states that 92 percent of companies have experienced financial loss due to a deepfake [Zaki 2024]. This is a significant economic impact.

These are just a few of many examples of individuals who have been hurt by deepfake attacks. There is no question about the negative impact that deepfakes actively have on our lives. It is now imperative that an application is developed to reliably identify deepfakes so that fewer individuals are harmed.

3 LITERATURE REVIEW

4 SYSTEM MODEL / BACKGROUND

Numerous technologies and algorithms were explored to gather insights on the most effective methods for deepfake audio detection. In this section we will explore the complex features we utilized in our data preprocessing pipeline and discuss the sophisticated machine learning models we employed in our experiments.

4.1 MFCCs

Mel-Frequency Cepstral Coefficients, or MFCCs, are a widely used feature set for speech recognition and other audio processing applications. They are a competitive feature set because they mimic what the human ear perceives [Hamza et al. 2022]. The process of extracting MFCCs involves several steps that convert raw audio into a compact representation of its spectral characteristics. One of the key stages in this process is the de-correlation of the log energies from a filter bank. The following quote provides a more detailed explanation of how MFCC coefficients are derived:

> "MFCC coefficients are obtained by de-correlating the output log energies of a filter bank which consists of triangular filters, linearly spaced on the Mel frequency scale. Conventionally an implementation of discrete cosine transform (DCT) known as distributed DCT (DCT - II) is used to de-correlate the speech as it is the best available approximation of the Karhunen-Loeve Transform (KLT)" [Hossan et al. 2010].

The resulting vector represents spectral acoustic features as floating-point numbers. These values capture essential characteristics of the audio that aid the model in distinguishing between real and fake samples. The vector's size is flexible and depends on user configuration.

Generating a vector of MFCC features was an essential step in our preprocessing efforts. A sample rate of 44100 Hz corresponding to a maximum sound frequency of 22050 Hz is usually used for recording sound, because a human can hear sounds ranging from 20 Hz to 20000 Hz [Tran and Lundgren 2020]. In our experiments, we

Conf 2024, Date, Place

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX
https://doi.org/XXXXXXX.XXXXXXX

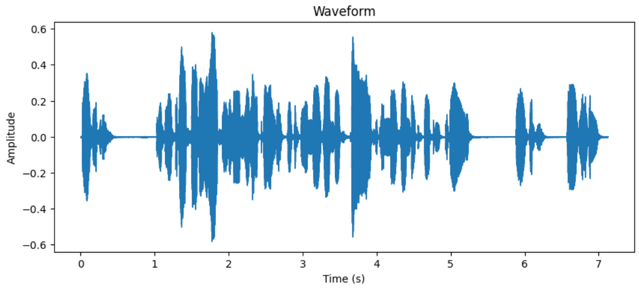

Figure 1: Waveform representation of a Real Audio Sample.

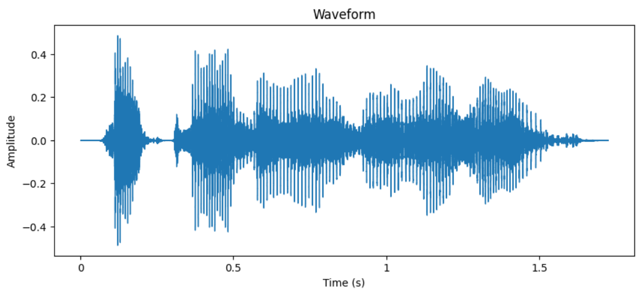

Figure 2: Waveform representation of a Fake Audio Sample.

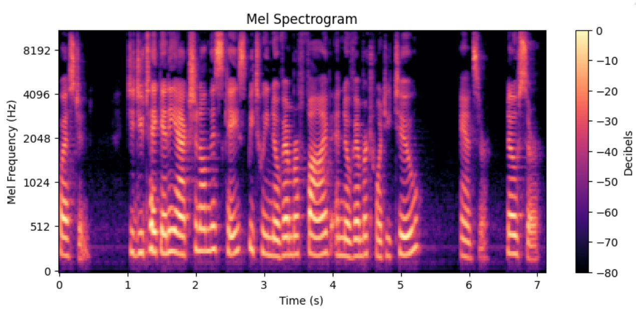

Figure 3: Spectrogram representation of a Real Audio Sample.

used a sampling rate of 22050 Hz. This is half of the common sample rate for recording sound, and is a common value for sampling rate when generating MFCCs. This resulted in 20 features for each audio sample. These features were then saved to a CSV file, making them available for later model training. These features played a pivotal role in the model's ability to make accurate decisions.

4.2 Mel-Spectrogram Images

A Mel-Spectrogram image is an alternative method for visually representing sound. Generally, we see sound visualized as a two-dimensional waveform showing amplitude over time. Figure 1 shows an example of a "real" audio sample as a commonly seen waveform. Conversely, figure 2 shows an example of a "fake" audio sample as a waveform. Similar to MFCCs, Mel-Spectrograms are generated by performing a number of transformations on an audio sample, then representing the results as a graph.

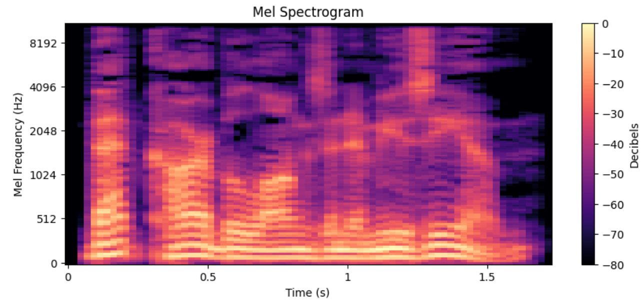

Figure 4: Spectrogram representation of a Fake Audio Sample.

4.3 Convolutional Neural Networks

Nora TODO

4.3.1 VGG16. Nora TODO

4.3.2 ResNet50. Nora TODO

4.4 Support Vector Machine (SVM)

4.5 Chroma Features

4.6 Gradient Boosting

5 METHODOLOGY

5.1 Probably a section about the data here would be good

6 RESULTS

6.1 for-original dataset

6.2 for-norm dataset

6.3 for-2sec dataset

6.4 for-rerec dataset

7 FUTURE WORK

Nora TODO

8 CONCLUSION

ACKNOWLEDGMENTS

REFERENCES

Daniella Genovese. 2024. Nearly 50% of voters said deepfakes had some influence on election decision: survey. https://www.foxbusiness.com/politics/nearly-50-voters-said-deepfakes-had-some-influence-election-decision. Accessed: 2024-12-03.

Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S. Almadhor, Zunera Jalil, and Rouba Borghol. 2022. Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access* 10 (2022), 134018–134028. https://doi.org/10.1109/ACCESS.2022.3231480

Md. Afzal Hossan, Sheeraz Memon, and Mark A Gregory. 2010. A novel approach for MFCC feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*. 1–5. https://doi.org/10.1109/ICSPCS.2010.5709752

Judge Herbert B. Dixon Jr. 2024. The "Deepfake Defense": An Evidentiary Conundrum. https://www.americanbar.org/groups/judicial/publications/judges_journal/2024/spring/deepfake-defense-evidentiary-conundrum/. Accessed: 2024-12-03.

Hyung-Jin Kim. 2024. In South Korea, rise of explicit deepfakes wrecks women's lives and deepens gender divide. https://www.pbs.org/newshour/world/in-south-korea-rise-of-explicit-deepfakes-wrecks-womens-lives-and-deepens-gender-divide. Accessed: 2024-12-03.

Thanh Tran and Jan Lundgren. 2020. Drill Fault Diagnosis Based on the Scalogram and Mel Spectrogram of Sound Signals Using Artificial Intelligence. *IEEE Access* 8 (2020), 203655–203666. https://doi.org/10.1109/ACCESS.2020.3036769

Adam Zaki. 2024. 92 percent of companies have experienced financial loss due to a deepfake. https://www.cfo.com/news/most-companies-have-experienced-financial-loss-due-to-a-deepfake-regula-report/732094/?utm_campaign=Yahoo-Licensed-Content&utm_source=yahoo&utm_medium=referral. Accessed: 2024-12-03.

Received XX December 2024; revised XX X XXXX; accepted X X XXXX