Detection of DeepFake Audio with Convolutional Neural Network

Halit Erdogan[1]

[1] Istanbul Technical University, Istanbul, Turkey
erdoganh16@itu.edu.tr

Abstract

With the development of machine learning (ML) and deep learning (DL) algorithms, a phenomenon known as deep generative modeling, commonly known as DeepFake, has emerged in the recent years. Even though there are some beneficial use cases of synthetic media generated by these deep generative models, there are also various ways that deepfakes can be threatening to society. Most of the research in this focuses on image detection. However, unauthentic audios can be just as harmful, especially if the person in the fake speech file is a politician. Therefore this study aims to investigate the classification between real speech audio and fake ones. To this end, we utilized some traditional audio processing methods such as mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCC).

1. Introduction

According to a 2019 story by the Wall Street Journal, a the CEO of a UK-based company was called by someone he thought was the CEO of the parent company, demanding a £220,000 transfer and said it was urgent. After the transfer it was revealed that the caller was not actually the CEO, but criminals using AI-generated fake speech. [1] This is a very concrete example of how damaging deep generative models can be. However, deep fakes can be harmful in terms of social reputation also. In a 2018 video posted by a major internet media company and viewed by more than 8 million on YouTube, former US President Barack Obama is shown saying things including insulting another former US President Donald Trump. Even though in this video it is openly explained that the speech is fake, it might not have been the case and these type of content can be used for social manipulation.

In order to prevent this situation, we utilize a dataset consisting of both real and fake speeches of ten politicians, Barack Obama, Bill Clinton, John F. Kennedy, Ronald Reagan, Arnold Schwarzenegger, Bernie Sanders, George W. Bush, Donald Trump, Boris Johnson and Winston Churchill.

2. Dataset and Preprocessing

The dataset consists of 35-40 clips of fake speech and 45-50 clips of real speech of each politician, with each clip lasting 10 seconds. Raw videos were obtained from YouTube in the .wav format and divided to clips of 10 seconds with a Python Script. At the end we used a total of 835 clips.

After that, each clip was analyzed using audio processing techniques like bispectral analysis, mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCC). However, it is important to note here that the sampling frequency of each video was inspected and all of them were found to be equal to exactly 48 kHz.

Mel-Spectrogram: A spectrogram visually represents frequency components of a signal over the time domain. To calculate the mel-spectrogram, the signal is divided into frames with an overlap between the frames. Then, one windows function such as Hann, Hamming, Blackman is applied to avoid spectral leakage. Afterwards, each individual frame is transformed by the Discrete Fourier Transform (DFT) to represent the signal in the frequency domain $X(t,k)$. where $t = 1, ..., T$ is the frame index of the signal and $k = 0, ..., K-1$ are the DFT coefficients. Finally, the squared magnitude $|X(t,k)|^2$ of the complex-valued signal is calculated and to obtain the spectrogram itself [2]. Mel-spectrogram is an alteration of spectrogram which parallels the fact that humans do not perceive all sound frequencies in a linear manner, but are able to differentiate lower frequency sounds with a higher resolution [3]. The mel-transform is shown in Equation 5, where f is the frequency and f_{mel} is the mel frequency.

$$f_{mel} = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \tag{1}$$

Mel Frequency Cepstral Coefficients: Mel Frequency Cepstral Coefficients (MFCC) are derived from a Mel-scaled spectrogram by applying a Discrete Cosine Transform (DCT) to the logarithm of the Mel-filtered signal.

$$c(t,r) = \sum_{s=0}^{S} \log\left[X_{mel}(t,s)\right] \cdot \cos\left[\frac{\pi \cdot r \cdot (s+0.5)}{S}\right] \quad \forall r = 0, ..., R-1 \tag{2}$$

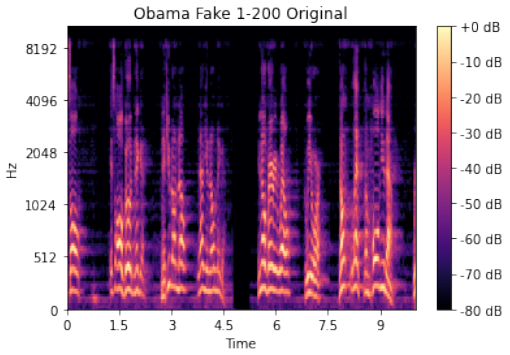
where R is the number of DCT coefficients.

In order to gain a visual intuitive on the audio files, graphs that show the mel-spectrograms of each clip were obtained. In Figure 1, two samples, one real and one fake, mel-spectrogram graph for each politicianâs clips are shown in the decibel order. In these graphs, it is seen that the silent parts of fake clips have absolute silence as opposed to the noise from the real speech clips which would create a bias in the model. To prevent this, an artificial Gaussian noise was added to each clip. Figure 2 shows the mel-spectrograms of the same clips.

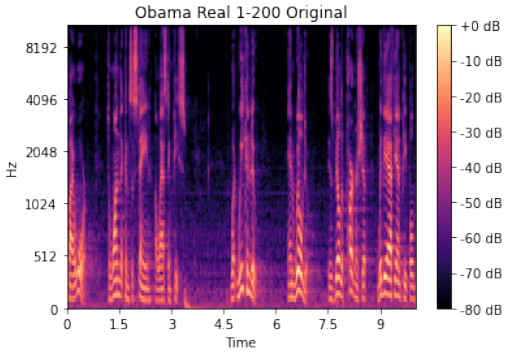
After adding the noise, the bispectrums, mel-spectrograms and MFCCs of all clips were calculated and labeled again. A train-test split of 80/20 was applied to both mel-spectrograms and MFCCs.

3. Models

The data containing mel-spectrograms and MFCCs were used to train a Convolutional Neural Network (CNN) model. CNNs are a class of neural networks generally used for visual pattern detection. [4]

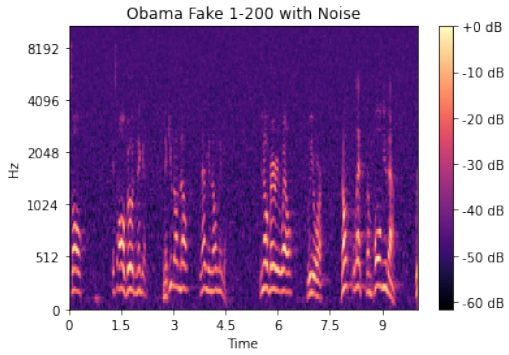


(a) Fake Speech Clip

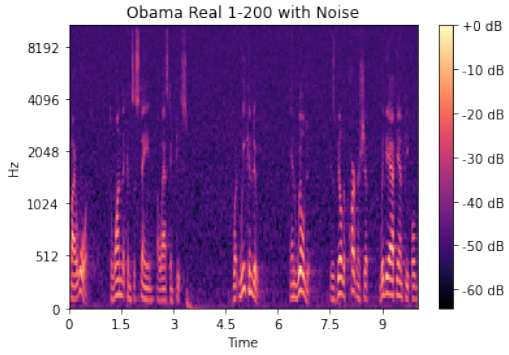


(b) Real Speech Clip

Figure 1. Mel-spectrograms of Obama's Fake and Real Speech Clips before adding the artificial noise



(a) Fake Speech Clip



(b) Real Speech Clip

Figure 2. Mel-spectrograms of Obama's Fake and Real Speech Clips after adding the artificial noise

In our dataset each data sample corresponds to an audio clip. Therefore, we adjusted the input layer shape of the network according to the mel-spectrogram or MFCC shape of one video, which corresponds to 128x920 and 20x938, respectively. here 938 is due to the time duration of each clip. 128 and 20 are results of the nature of mel-spectrogram and MFCC functions.

After the input layer, both CNNs were constructed in the same way. The first layer included 16 kernels of size 2x2, the maxpool layer following the first layer consists of pooling size of 2Ã2. The second layer included 32 kernels of size 2x2, with the following maxpool layer also having size 2x2. Then, the network is flattened. Afterwards, a dense layer with unit 32 is added, followed by a dropout layer with a rate of 0.3. Finally the output layer is set to a dense layer with unit 1. Activation functions of all layers are set to relu function, with the exception being the output layer, whose activation function is sigmoid function. The models were compiled with batch size equals to 4 and 100 epochs.

3.1. Results of Mel-spectrogram Model

The model trained with mel-spectrograms provided near perfect results with a 99.8% accuracy. The training of the model took 17-18 seconds for each epoch, which added up to approximately 30 minutes in total.

Figure 3 shows the loss and accuracy graph with each epoch, while Table 1 contains the scores of the model. Figure 4 shows the confusion matrix for this model.

Table 1. Scores of the Mel-Spectrogram Model

	Precision	Recall	F1-Score	Support
Fake	0.99	0.99	0.99	68
Real	0.99	0.99	0.99	98

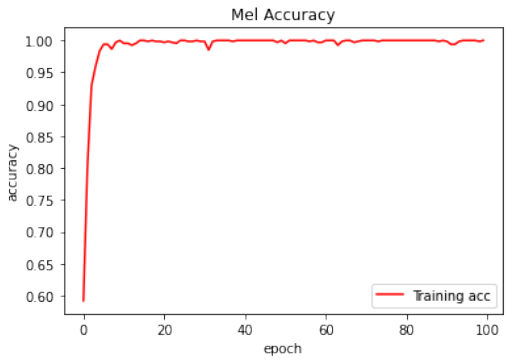
3.2. Results of MFCC Model

The model trained with MFCCs also provided near perfect results with a 98% accuracy, although still lower than the former. However, the training of the model took only 3-5 seconds for each epoch, which added up to approximately 5 minutes in total. This makes MFCCs considerably advantageous in comparison to mel-spectrograms if the dataset is expanded to astronomical sizes.

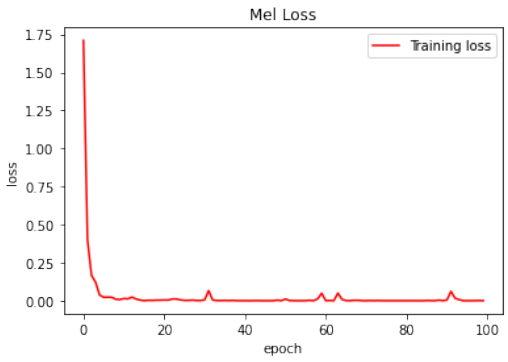
Figure 5 shows the loss and accuracy graph with each epoch, while Table 2 contains the scores of the model. Figure 6 shows the confusion matrix for this model.

Table 2. Scores of the MFCC Model

	Precision	Recall	F1-Score	Support
Fake	0.98	0.97	0.98	65
Real	0.98	0.99	0.99	101

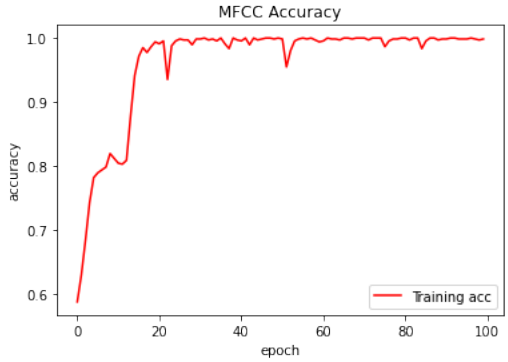


(a) Accuracy

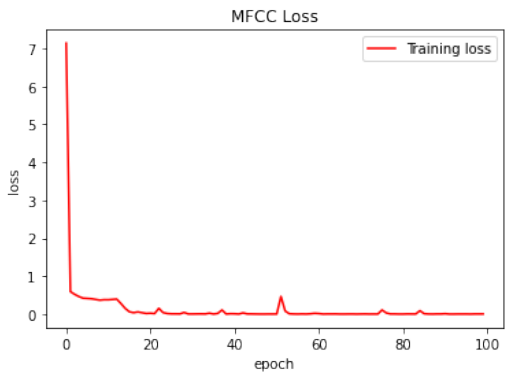


(b) Loss

Figure 3. Graphs showing accuracy and loss values over each epoch for mel-spectrogram model



(a) Accuracy



(b) Loss

Figure 5. Graphs showing accuracy and loss values over each epoch for MFCC model

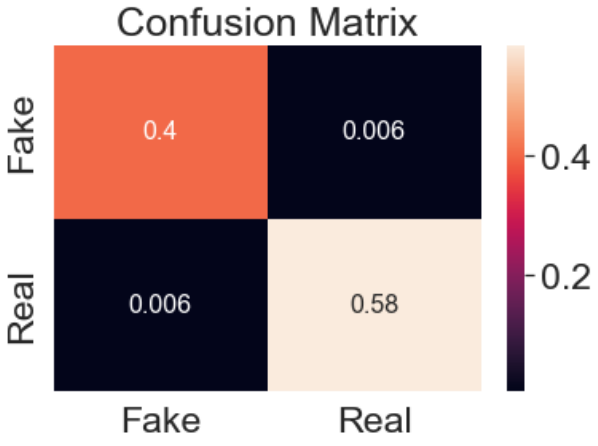


Figure 4. Confusion matrix of mel-spectrogram model

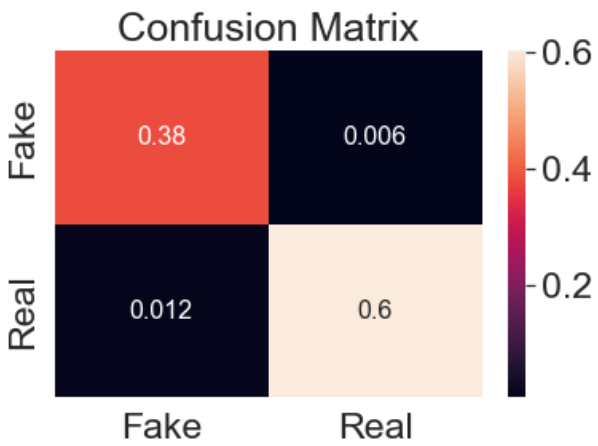


Figure 6. Confusion matrix of MFCC model

4. References

[1] C. Stupp, "Fraudsters used ai to mimic ceoâs voice in unusual cybercrime case," *The Wall Street Journal*. [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[2] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," 2021.

[3] H. Fastl and E. Zwicker, "Psychoacoustics: Facts and models. springer-verlag," *Germany.[Google Scholar]*, 2007.

[4] M. Valueva, N. Nagornov, P. Lyakhov, G. Valuev, and N. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Mathematics and Computers in Simulation*, vol. 177, pp. 232–243, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378475420301580