Can We Trust Our Ears?
Detecting DeepFake Audio with Targeted Feature Extraction and Machine Learning

Gokul Dharan
Department of Computer Science
Stanford University

Nikka Mofid
Department of Electrical Engineering
Stanford University

David Whisler
Department of Electrical Engineering
Stanford University

Abstract

Synthesized video or audio, often referred to as "deepfakes", have increased in sophistication to the point that they are often indistinguishable from genuine media. This technology poses a significant risk to security and can be a tool for identity theft and social manipulation. We present a number of classifiers that achieve excellent performance in detecting spoofed audio, with a focus on the impact of feature extraction methods on model performance. We find that Constant Q Cepstral Coefficients (CQCC) outperform all other tested features across model architectures, and that the best classifier architecture is a Convolutional Neural Network (CNN).

1 Introduction

An audio deepfake, or "spoof", can sound just like a real voice and has the potential to fool voice verification systems in banks and even put words into the mouths of politicians [4]. Deepfakes are often generated by utilizing machine learning methods like auto-encoders or generative adversarial networks (GANs) in order to synthesize speech. Thus, methods for accurate audio spoofing detection are essential. In this work, we build a classifier that can distinguish between authentic and synthetic speech, utilizing several different signal-processing based feature extraction methods and machine learning models. Particularly, we explore whether more advanced feature extraction techniques can improve the performance of standard classification algorithms. The input to our algorithms is an audio clip which may or may not be a deepfake. We use several different feature extraction methods to extract features from the input audio file, which are then fed to one of our three machine learning models (logistic regression, GMM, and CNN) in order to classify whether or not the audio is a deepfake.

2 Related Work

The biennial ASVspoof Challenge is a competition that most recently occurred in 2019 [7]. Its results reveal that the classification of spoofed audio is as dependent on the features used as it is on the model architecture. In [5], the authors present an analysis of different feature extraction pipelines, including Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, on the performance of Gaussian Mixture Models (GMMs) for synthesized audio classification. Deep Learning has also been utilized for this classification problem, with [2] showing promising results with a CNN+RNN architecture combined with Constant Q Cepstral Coefficient (CQCC) features. In [8], the authors use high-frequency versions of features including CQCCs and MFCCs along with GMMs as they show that sufficient evidence for classification of spoofing lies solely in high-frequency bands of the signal. CQCC's in particular are a powerful method for detecting utterances and synthesized speech, as shown in a number of studies [3][6]. The common theme among these works is that they tend to couple one or more feature extraction methods to a particular model architecture. In contrast, we evaluate a number of the most promising feature extractors across three different classification models to evaluate their relative performance and value independent of model architecture.

3 Dataset

We will be using the Automatic Speaker Verification dataset [9] released by Google for the ASVspoof Challenge. This dataset contains 25,000 short ($\sim$5 second) audio files labeled as either real or spoofed. From the full ASVspoof Challenge data, a subset of 5160 total audio files was used for training, which was balanced between the two classes (authentic and spoofed).

The ASVspoof dataset provides training data with a selection of spoof methods, notated A01-A06. In addition to evaluating each model on a balanced test set of 1806 audio clips using the same spoof methods as in the training set, we also evaluated each model on a second test set of 505 audio files generated using *different* spoof methods, notated A07-A19. This helped indicate how well each spoof-detection model generalized to other spoof methods that may be used in the real world, rather than just those it was explicitly trained on. The different spoof methods are described in the ASV spoof dataset [7].

4 Methods

The project code is publicly available on Github [1].

4.1 Features

We extracted several features from the audio data in both the time and frequency domains. These included the RMS energy, spectrogram, Mel-spectrogram, MFCCs, and CQCCs. The RMS energy is a time domain feature which calculates the power in the signal. This is useful as a baseline time domain feature, since it cuts down the dimensionality of audio data to a manageable level.

The spectrogram of an audio signal is a frequency domain versus time domain representation. To calculate it, a time domain signal is divided into a series of windows over time with a given size. The Fourier transform is calculated for each window to move it into the frequency domain, creating a plot of frequencies versus time. The mel-spectrogram is variant of the spectrogram where instead of plotting using a linear scale for the frequency domain, the Mel scale is used, which is an empirically derived nonlinear scale of frequencies judged by human listeners to be of equal distance.

Mel frequency cepstral coefficients are features that are commonly used for audio analysis. They are calculated by taking a windowed Fourier transform of an audio signal, mapping it to the Mel frequency scale, taking the logarithm of the signal power at each of the Mel frequencies, and taking the Discrete Cosine Transform of the resulting powers treating them as a new signal. The MFCCs are then taken as the magnitude of the resulting DCT coefficients, and are plotted versus time in the same way as a spectrogram.

CQCCs are first presented in [6] as an improvement over all previous feature extraction methods for spoof-detection tasks. Rather than the Fourier transform used by MFCCs, this method uses the constant Q transform (CQT), which uses geometrically spaced frequency bins which are intended to closer approximate human perception. They were used as a baseline feature extraction method in the most recent ASVspoof challenge [7], and this open-sourced implementation of CQCCs is used in this project, as well.

4.2 Models

LR - Logistic Regression is very simple yet very powerful and is one of the most popular classical machine learning models used for classification tasks. It utilizes the logistic function to model a binary variable dependent on the input features.

GMM - A Gaussian Mixture Model approach is also evaluated. The flexibility of GMMs as a form of generative classification as well as its use in related baselines [5] motivated its use in this setting. Since we are binning frequencies for the spectrogram into 512 bins, we use a 512-component GMM. Two GMMs are trained, one on the spoofed clips and the other on genuine audio, and then predictions are made by scoring the probability that a new example clip is generated by each of the two GMMs.

CNN - The third model used was a Convolutional Neural Network. The spectrogram, mel-spectrogram, MFCC, and CQCC feature extraction methods all produce two-dimensional data, with time on one axis, and frequency on the other axis. Because of this, and since spoof detection should be time-invariant, a convolutional neural network (CNN) may effectively learn spatial patterns that are useful in classifying spoofed audio. A simple network with the the architecture shown in Figure 1 was used for the classification task, utilizing two convolution layers with 8 and 16 3x3 filters respectively, as well as dropout layers with a dropout rate of 0.1 to promote regularization.

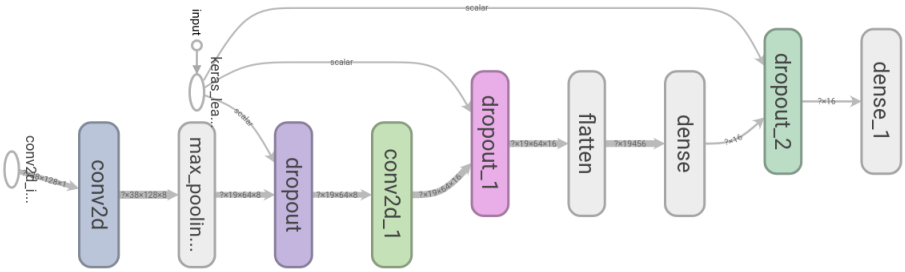

Figure 1: CNN Model Architecture

5 Results

All models were run with a balanced training set of 5160 audio files sampled at 16kHz. As described in the Dataset section, in addition to evaluating each model on a test set using the same spoof methods as in the training set (A01-A06), we also evaluated on another test set using different spoof methods (A07-A19) to see how well the models generalized. For the CNN, a learning rate of 1e-3 was used, with a batch size of 1 and 10 epochs. Table 1 displays the evaluation results.

Model	Features	Spoof Methods	Accuracy	Precision	Recall	F1
Logistic Regression	RMS Energy	A01-A06	0.576	0.997	0.507	0.672
		A07-A19	0.877	0.979	0.765	0.859
	Spectrogram	A01-A06	0.379	0.971	0.764	0.355
		A07-A19	0.762	0.847	0.627	0.620
	Mel-Spectrogram	A01-A06	0.306	1.000	0.191	0.321
		A07-A19	0.770	0.985	0.538	0.696
	MFCC	A01-A06	0.866	0.980	0.860	0.916
		A07-A19	0.786	0.867	0.663	0.752
	CQCC*	A01-A06	0.943	0.999	0.935	**0.966**
		A07-A19	0.838	0.956	0.700	0.808
GMM	RMS Energy	A01-A06	0.587	0.998	0.519	0.683
		A07-A19	0.867	0.950	0.769	0.850
	Spectrogram	A01-A06	0.820	0.998	0.792	0.883
		A07-A19	0.869	0.955	0.769	0.852
	Mel-Spectrogram	A01-A06	0.575	0.996	0.506	0.671
		A07-A19	0.873	0.974	0.761	0.855
	MFCC	A01-A06	0.896	0.998	0.880	0.935
		A07-A19	0.822	0.954	0.668	0.786
	CQCC*	A01-A06	0.934	0.999	0.924	**0.960**
		A07-A19	0.846	0.972	0.704	0.817
CNN	Spectrogram	A01-A06	0.903	0.999	0.888	0.940
		A07-A19	0.871	0.969	0.761	0.853
	Mel-Spectrogram	A01-A06	0.526	0.997	0.448	0.619
		A07-A19	0.832	0.982	0.668	0.795
	MFCC	A01-A06	0.817	1.000	0.787	0.881
		A07-A19	0.848	0.983	0.700	0.818
	CQCC*	A01-A06	0.953	0.997	0.948	**0.972**
		A07-A19	0.909	0.967	0.842	0.900

Table 1: Full Test Results, with best features (by average F1 across both test sets) highlighted in bold

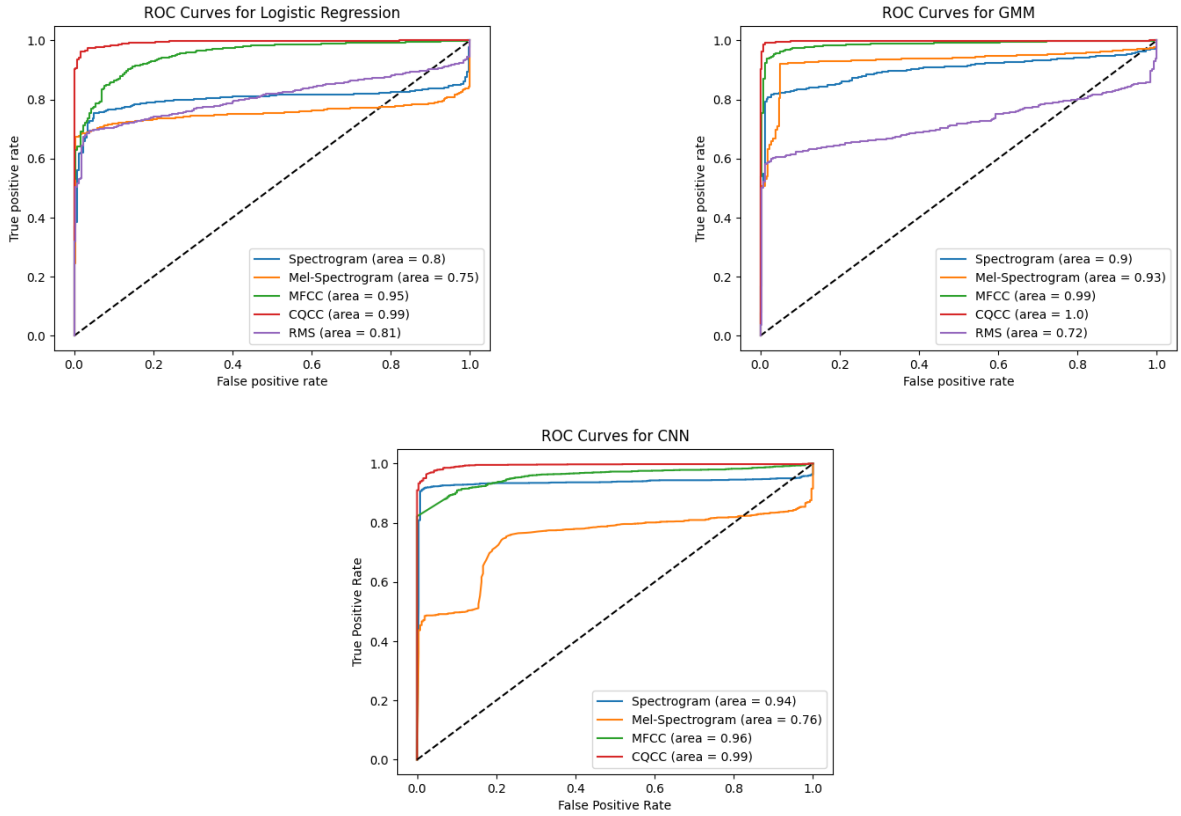

Figure 2: ROC curves for each model and feature on the test set (A01-A06).

6 Discussion

On the whole, the classifiers show positive results, consistently achieving F1 > 0.9 on the A01-A06 test set. On this set, LR-CQCC, GMM-CQCC, and CNN-CQCC achieve an F1 of 0.966, 0.960, and 0.972, respectively, demonstrating excellent classification performance. When evaluating the model on unseen spoof methods A07-A19, the performance does suffer, but regardless of the model, there is at least one feature that achieves F1 > 0.8 on the unseen spoofing methods, proving that these methods can generalize well. Additionally, CNN-CQCC generalizes better than any other model-feature pair, achieving an F1 of 0.900 on the unknown spoofing methods.

We also note that for every model-feature pairing except CNN-CQCC, precision is significantly higher than recall and remains high across features. Spoof detection may be a scenario in which Type II errors are less desirable than Type I errors. Specifically, we may prefer to have some false positives when detecting spoofs rather than letting a spoofed clip be classified as genuine (false negative), in which case we ought to evaluate these methods by their recall. In terms of recall, there is much more variance across model-feature pairs than with pure F1 or precision scores. We note that LR-CQCC, GMM-CQCC, and CNN-CQCC have the best recall scores of 0.935, 0.924, and 0.948, respectively. This suggests room for improvement since a 5% classification rate of spoofs as genuine audio may be too high for real-world applications. The ROC curves in Figure 2 show that in general, adjusting the classification threshold to increase the true positive rate above 0.95 will significantly increase the false positive rate. This indicates that more sophisticated models or features may be required to boost the true positive rate without compromising the overall performance of the model.

6.1 Model Evaluation

Table 1 shows that GMM performed similarly to standard logistic regression, while having an advantage with some features like the mel-spectrogram. In order to make a more deliberate, holistic assessment of the relative performance of these models and features, we use the area under the ROC curves (AUC), presented in Figure 2. This metric reveals that GMM outperforms logistic regression across all features except for RMS Energy, which significantly underperforms on the GMM model with an AUC of 0.72. This suggests that the GMM may not be suited to time-domain features;

the modelling assumption that the input can be modeled by a sum of Gaussians may not be as accurate as in the case when the audio has been transformed to the frequency domain, as is the case with all other features. Figure 2 shows that GMMs are better suited for this task with all other features, as the AUC is consistently higher with these features. Interestingly, we see a soft upper-limit of 0.8 for the true positive rate of logistic regression for spectrogram and mel-spectrogram features that does not seem to exist for GMMs. It appears that for both these features, GMM exploits features of the data that standard logistic regression is not able to. On the whole, the GMM classifier achieves higher AUC across more features than logistic regression, leading us to conclude that GMMs are overall better-suited for this task. However, in terms of test set performance, neither model performed as well as CNN-CQCC.

For the CNN, deep methods generally are more difficult to analyze. However, one area of note is that the CNN tended to perform better with dense data - the CQCC and MFCC features tended to outperform the spectrogram and mel-spectrogram features, the latter of which tended to be sparse (since much of their frequency spectrum were close to zero for large frequencies). The MFCC and CQCC features were more densely packed with data, giving more degrees of freedom to the deep model to learn with. This shows that a deep model may improve its performance with less binning, and allowing it to work with as many degrees of freedom as possible.

6.2 Feature Evaluation

MFCCs and CQCCs are excellent features for all three models, since they achieve AUC ≥ 0.95 (Figure 2) across model architectures. As a result, we conclude that MFCCs and CQCCs are strong feature extraction methods for this task regardless of the choice of classifier model, with CQCCs being preferred since they outperform MFCCs across model architectures. Our results in Table 1 - for example, an F1 of 0.355 for LR-Spectrogram and 0.966 for LR-CQCC - demonstrate the importance of deliberate feature selection when it comes to classification performance for this task, as the right choice of feature can elevate simple models to highly accurate classifiers.

As hypothesized, CNNs were able to take advantage of the 2D spatial patterns in a spectrogram, significantly outperforming GMM-Spectrogram and LR-Spectrogram. Interestingly, this did not translate to the mel-spectrogram. Visualizing some of the spectrogram results between real and spoofed audio files, one observes that the real audio tended to be more band-limited in frequency than the fake audio files (see Figure 3). This makes sense, because real human voices are more limited in the range of frequencies they can produce, compared to algorithms which can introduce frequency artifacts outside of the spoken frequency range. When using a spectrogram, we suspect the CNN was better suited to detecting this change in the bandwidth, leading to its higher performance compared to the GMM and LR equivalents.

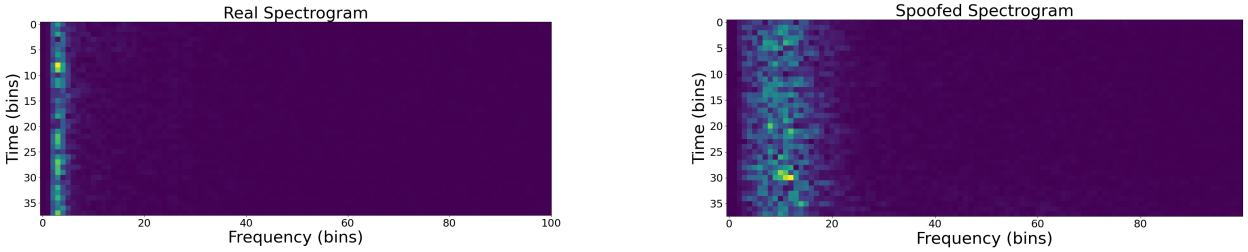

Figure 3: Example Spectrogram Results

7 Conclusion & Future Work

We showed that though several frequency domain features were able to successfully identify deepfake audio from real audio. CQCC features performed the best among the logistic regression, GMM, and CNN models, and between models the CNN performed the best of all, especially when evaluating performance in the general case with spoof methods not seen during training. This was likely due to the representational power of the deep network, and its ability to learn relationships among the data with fewer assumptions than the classical logistic regression and GMM models. The most significant limitation of the tested methods is that the recall rate may not be high enough for real-world applications in which identifying every single instance of spoofed audio may be critical. Going forward, spending more time refining the hyperpameters and architecture of the CNN model would likely yield improvements in the accuracy for generalized spoof methods, and more sophisticated architectures combined with CQCC features may boost the recall rate. In addition, exploring other deep methods that can work better in an end-to-end design from raw audio data, such as LSTM RNNs, may yield further improvements and efficiency.

8 Contributions

- **David Whisler:** Data Preprocessing, RMS, Spectrogram, Mel-Spectrogram, and MFCC Feature Extraction, and CNN Model
- **Nikka Mofid:** Logistic regression model implementation, Logistic regression feature experiments (RMS, Spectrogram, Mel-Spectrogram, MFCC), ROC curves logistic regression
- **Gokul Dharan:** Gaussian Mixture Model implementation, CQCC Feature Extraction, GMM experiments, LR-CQCC experiments

References

[1] Gokul Dharan, Nikka Mofid, and David Whisler. *Audio Deepfake Detection*. URL: https://github.com/gokuldharan/audio_deepfake_detection.

[2] G. Lavrentyeva et al. "Audio Replay Attack Detection with Deep Learning Frameworks". In: *INTERSPEECH*. 2017.

[3] Dipjyoti Paul, Md Sahidullah, and Goutam Saha. "Generalization of Spoofing Countermeasures: a Case Study with ASVspoof 2015 and BTAS 2016 Corpora". In: (2019). eprint: arXiv:1901.08025.

[4] J.M. Porup. *Deepfake videos: How and why they work - and what is at risk*. Apr. 2019. URL: https://www.csoonline.com/article/3293002/deepfake-videos-how-and-why-they-work.html.

[5] Balamurali B. T. et al. "Towards robust audio spoofing detection: a detailed comparison of traditional and learned features". In: *CoRR* abs/1905.12439 (2019). arXiv: 1905.12439. URL: http://arxiv.org/abs/1905.12439.

[6] Massimiliano Todisco, Hector Delgado, and Nicholas Evans. "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients". In: *ODYSSEY 2016, The Speaker and Language Recognition Workshop, June 21-24, 2016, Bilbao, Spain*. B, June 2016. URL: http://www.eurecom.fr/publication/4855.

[7] Massimiliano Todisco et al. *ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection*. 2019. arXiv: 1904.05441 [eess.AS].

[8] Marcin Witkowski et al. "Audio Replay Attack Detection Using High-Frequency Features." In: *Interspeech*. 2017, pp. 27–31.

[9] Junichi Yamagishi et al. *ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database*. URL: https://datashare.is.ed.ac.uk/handle/10283/3336.