A Survey of Machine Learning Methods on Spoof Detection

Allison John[1], Calvin Sokk[2], Sheena Lai[1]

[1]Department of Computer Science [2]Department of Electrical Engineering

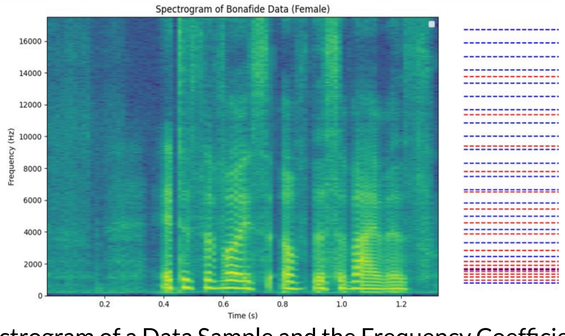
Project Overview

With the emergence of AI capable of generating deepfake audio and video content, there is an increasing need to differentiate between genuine and computer-generated media due to its potential use in misinformation, defamation, scams, etc. We built, tested, and compared several machine learning models with various feature sets that predict whether a speech audio sample is a computer-generated voice sample or a human-generated voice sample. The model that achieved the best overall test accuracy is a Recurrent Neural Network (RNN) with Mel Frequency Cepstral Coefficient (MFCC) features.

Dataset

- Third Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof) dataset from 2019 [1].
- Features "bonafide" (human-generated) or "spoof" (computer-generated) voice audio samples.

Features



Spectrogram of a Data Sample and the Frequency Coefficient Axes of LFCC (Blue) and MFCC (Red)

Frequency analyses are valuable in audio detection as voices are characterized by distinct frequency components.

1. **Raw Fourier Transform (FT)**, samples the Fourier Transform magnitude at linear intervals.
2. **Linear Frequency Cepstral Coefficients (LFCC)** samples the power density spectrum of the Fourier Transform at linear intervals.
3. **Mel Frequency Cepstral Coefficients (MFCC)** samples the power density spectrum of the Fourier Transform on intervals of the Mel scale to approximate human audio perception.
4. **Constant-Q Cepstral Coefficients (CQCC)** samples the power density spectrum of the Constant-Q transform instead of the Fourier Transform on a logarithmic scale similar to MFCC.

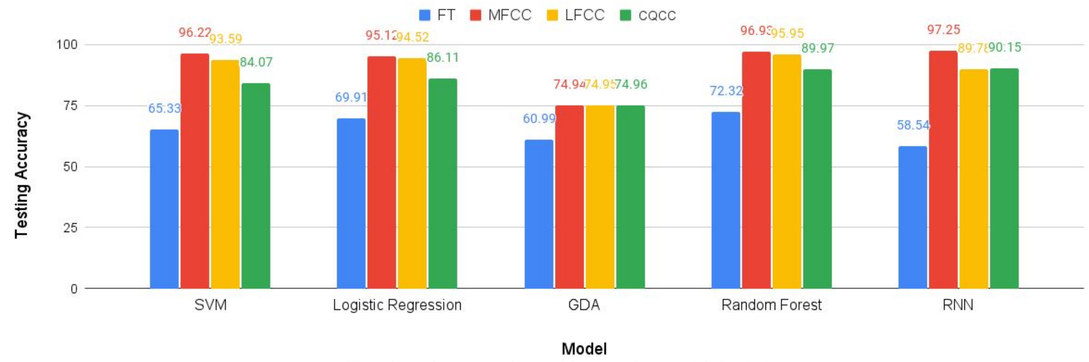
Models

1. **Support Vector Machine (SVM)**: Our baseline model is SVM, commonly used for audio deepfake detection. SVM aims to find a hyperplane that maximizes the margin between the two classes with the objective function $L_w(x, y) = \Sigma_i max(|w^T x_i - y_i|, 0)$
2. **Logistic Regression**: Logistic regression is a simple model that can classify as 0 or 1, so this matches our bonafide or spoof problem. The logistic regression objective function is
   $\text{Loss} = \sum_i -y_i \log(\theta^T x_i) - (1 - y_i) \log(1 - \theta^T x_i)$
3. **Gaussian Discriminant Analysis (GDA)**: Since Gaussian Mixture Models are often used for spoof-detection tasks, we explored if GDA could characterize the spoof and bonafide classes well with approximations of Gaussian distributions. GDA is also similar to logistic regression, but it is a generative learning algorithm instead of discriminative. The log likelihood objective function is $\log \prod_i p(x_i|y_i; \mu_0, \mu_1, \Sigma)p(y_i|\phi)$
4. **Random Forest**: Random forests use the average result of a set of decision tree classifiers for a more controlled estimate. Random forests have a tendency to overfit, so we hypothesized that if the model overfits the bonafide class, it will be more likely to correctly classify unseen spoof generation techniques. We tuned all 100 trees to a maximum depth of 8. The function to determine each split is the Gini impurity: $1 - p_0^2 - p_1^2$
5. **Recurrent Neural Network (RNN):** RNNs work well on sequential data such as audio data due to how the recurrent structure assumes that subsequent data depends on previous data. Our RNN was constructed with 2 hidden layers, each of width 40 nodes. The loss function is cross entropy loss: $-\log\left(\frac{\exp(x_c)}{\sum_j \exp(x_j)}\right)$

For logistic regression and RNN, we added regularization to the loss to prevent overfitting: $\lambda||w||^2$ where λ is the regularization constant.

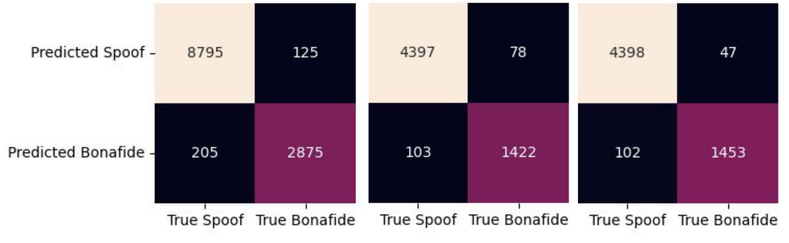
[1] Yamagishi, Junichi, et. al.. (2019). ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2555.

Results & Discussion



Testing Accuracies of Investigated Models

- **GDA:** Worst performance shows Gaussian distribution is not sufficient in representing spoof or bonafide classes.
- **Random forest:** Random forest generally performed better than the SVM baseline marginally.
- **RNN**: RNN was able to achieve the highest evaluation accuracy trained on MFCC features, with an accuracy of 97.25%.
- **MFCC feature set performed the best:** For each model, training on the MFCC feature set was able to achieve the highest evaluation accuracies.
- **Failure analysis:** Failure analysis revealed inconsistencies in the dataset between bonafide and spoof classes, such as audio length and amount of silence.



RNN confusion matrices for all examples (left), female (middle), and for male (right). The RNN produced more false positives in the female set. This could be due to the female samples being shorter on average than male samples.

Conclusion & Future Research

Conclusions:

- RNN and random forest models are effective in distinguishing spoof versus bonafide.
- MFCC and LFCC characterize audio samples well for determining spoof versus bonafide, although MFCC performs slightly better.

Future Work:

- Preprocess and normalize audio to get more accurate representation of the model's spoof detection performance.
- Explore random forest by iterating on more random seeds to find a more optimal forest or combine with different feature forests to find a more optimal composite forest.
- Improve our RNN model by employing gradient clipping and initialize RNN at different states to prevent training loss from settling in non-optimal minima.