A Conv1D-LSTM Approach to the ASVspoof 2019 Challenge

By Daniel Evert

YouTube Link: https://youtu.be/uGR2TjqpNK8

MOTIVATION

The rise and accessibility of (deep) fake audio poses a modern risk to financial institutions as it grapples with a variety of social engineering attack vectors, specifically one in which cybercriminals disguise their voice using synthetic speech to access consumer accounts.

If a financial institution was equipped with knowledge at the point when fraudsters disguise their voice, it can potentially mitigate this attack vector which attempts to acquire credentials from unsuspecting consumers saving these companies millions of dollars.

The ASVspoof 2019 Challenge was set up to deal with this challenge of detecting synthetic audio by providing variety of spoofing scenarios injected into the datasets.

Text-to-Speech (TT) and Voice Conversion (VC) were two types of spoofing attack vectors that this project focused on tackling using the ASVspoof's Logical Access (LA) dataset.

By combining an automated speech verification (ASV) system with the scores produced by the spoofing classifier of this project, it is possible to provide a solution to a financial institution which not only verifies a customer's identity but determines whether that voice be bonafide or spoofed.

CONTACT

Daniel Evert
Stanford Univesrity
Email: Dje334@Stanford.edu
Phone: (512) – 680 – 2638

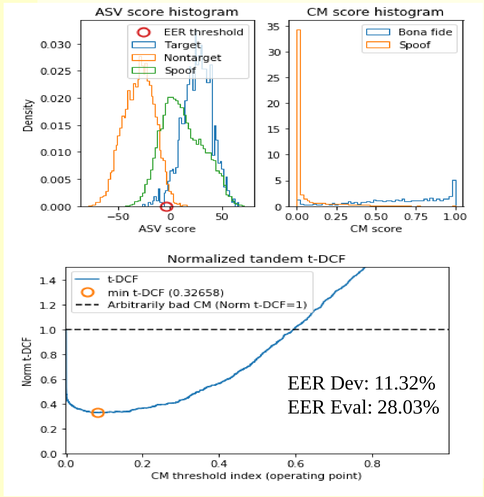
PROBLEM DEFINITION

The Automatic Speaker Verification (ASV) 2019 Challenge came about as a need to classify "bonafide" and "spoof" audio which then integrates into an ASV system that verifies individuals.

The purpose is to thus combine the spoofing classifier with the ASV-system using and decisioning on whether to believe the individual to be who they are (spoof, bonafide target, bonafide non-target).

Thus, develop an algorithm that can reduce the error rate on individual misclassification, using the tandom decision cost function (t-DCF).

RESULTS



REFERENCES

[1] Lavrentyeva1, G., et al. 2017. Audio replay attack detection with deep learning frameworks. ITMO University, St.Petersburg, Russia. INTERSPEECH2017

[2] Automatic Speaker Verification (ASV): Spoofing And Countermeasures Challenge. 2019.

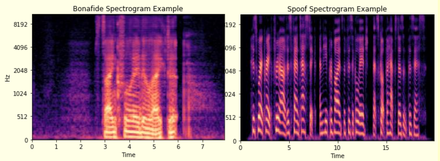
[3] M. Sahidullah, T. Kinnunen and C. Hanilci.A comparison of features for synthetic speech detection.Proc. INTERSPEECH 2015, pp. 2087–2091

DATA

To train the spoofing classifier, the data consisted of three datasets: training (25,380 audio files), development (24,844 audio files) and evaluation (71,237).

Converted audio files used the Mel-frequency cepstral coefficients (MFCC) as the feature extractor.

Inputs into the dataset were then determined by a sliding window of the spectrogram, yielding 787,967 records each 128X48 in dimensions of the frequency and time domains, respectively.



All partitions ~ 89% spoof to bonafide examples.

CHALLENGES

Challenge One: *Offsetting Sliding Windows*

Ideally, offsetting of size 1 per window would yield the most data points for training. Unfortunately, the AWS cloud could not support all data into memory at one time, so an offset parameter of size "k" was chosen.

Challenge Two: *Feature Extractor*

Whether to use grayscale MFCC's or a combination of multiple feature extractors so to have multiple channels for the Convolutions was considered but yielded too much in-memory costs.

Challenge Three: *Hyper-Parameter Tuning*

The hyper-parameter space consisted of several thousand possibilities, ranging from learning rate, batch-size, number of convolutions, offset size, number of time windows, number of kernels and filters per convolution, dropout, type of activation, etc. The final model took 8 hours to run.

APPROACHES

Approach – Use Holdout methodology. This mitigates the time required to build the Conv1D-LSTM at the risk of less generalization.

Approach – Use a Deep Conv1D with Max Pooling at each timestep to be used as the input of the LSTM's time-step.

Approach – Used Batch-Normalization after each Max Pooling Layer and between Fully Connected Nodes

Approach – Used Dropout (0.1-0.5) for regularization to reduce overfitting.

Approach – Models prioritized by F1-Score.

Discussion Point – If the ASVspoof challenge provided more "bonafide" examples, a likelier case, would that help with model generalization?

Discussion Point – What if one wanted to make this a real-time scoring model? What architecture and sliding window approach would fit this problem description?

CONCLUSIONS & FUTURE WORK

Conclusion – Model was able to successfully pick up on features and generalize synthetic speech from training to the evaluation set.

Conclusion – Model was slow to score and generate samples. Real-time scoring applications are not foreseeable with current approach.

Future Work – Attempt Conv2D methodology with a specified set of outputs as inputs into the LSTM (equal in time steps). This approach has potential of allowing deeper convolutional networks.

Future Work – Configure a customized validation function that conforms to the avg(preds) for each dev example

Future Work – Attempt a softmax() for each of the spoof-attacks and discover if it outperform by learning the weights differently

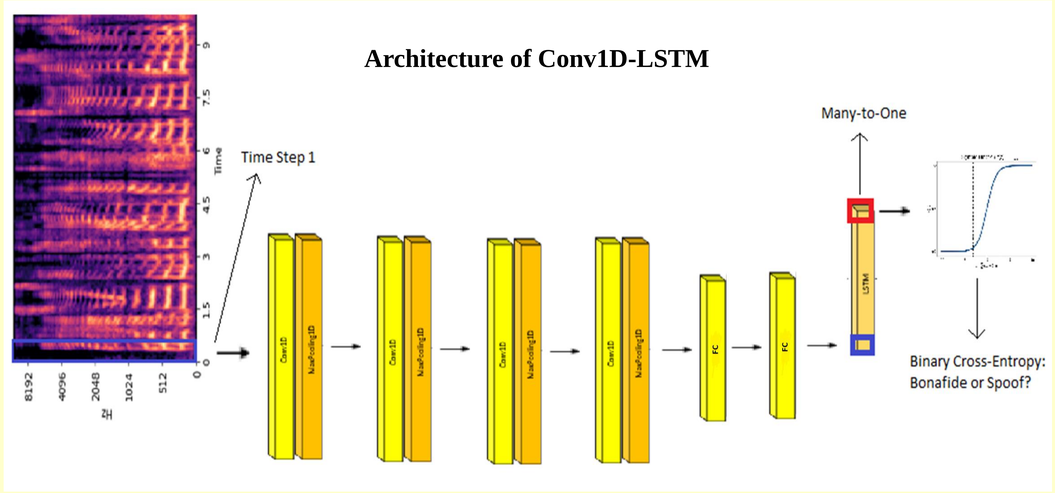


Figure 1