A Deeper Look At Gaussian Mixture Model Based Anti-Spoofing Systems

Queen Mary University of London

School of EECS
Queen Mary University of London, UK

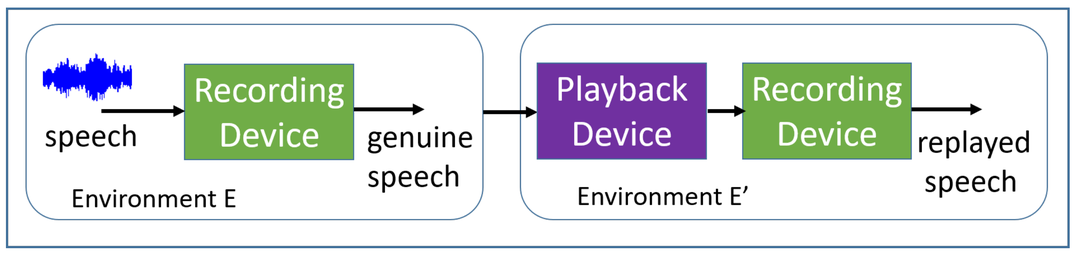
Bhusan Chettri and Bob L. Sturm
{b.chettri, b.sturm}@qmul.ac.uk

1 Introduction

We analyze several Gaussian Mixture Model (GMM) based replay spoofing detection systems and investigate key factors that influence model prediction. What have these GMM systems learned about genuine vs spoof signals in this dataset? Our analysis shows how system performance can depend on a class-dependent cue in the dataset: initial silence frames of zeros appear in the genuine signals but missing in the replayed signals. We demonstrate how system predictions can be fooled using this cue. Finally, we show that pre-processing the dataset helps mitigate the problem.

2 ASVspoof 2017 challenge [1]

Automatically determine if a speech utterance is genuine speech or a recording.

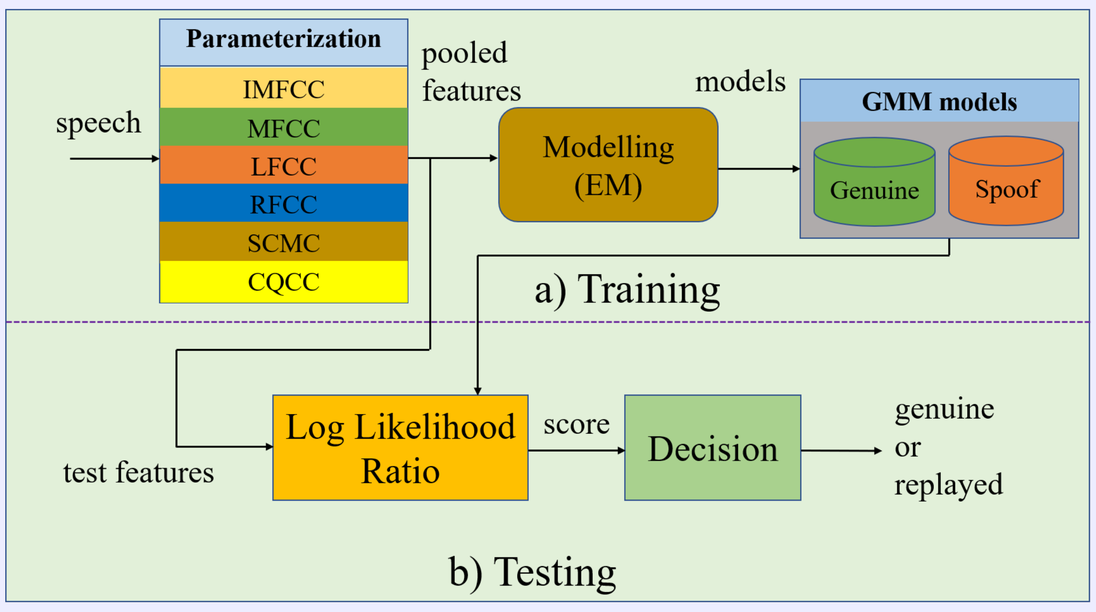


The ASVspoof 2017 database [2] comes from a subset of the RedDots corpus. Table 1 shows the statistics.

Subset	# Speakers	# Genuine	# Replayed	Duration(hr)
Train	10	1508	1508	2.22
Development	08	760	950	1.44
Evaluation	24	1298	12922	11.95

3 System description

All our systems use 40-dimensional features (20 delta and 20 acceleration coefficients). No voice activity detection, pre or post processing. Pooled data is used for GMM training.



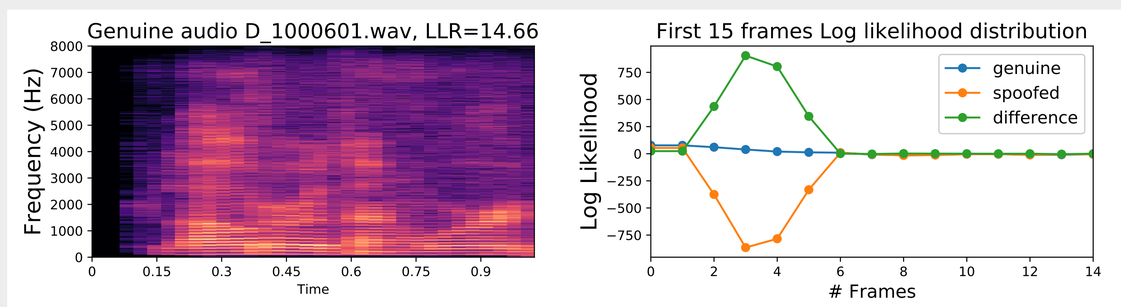
4 Results

Table 2 below shows the **performance (EER%)** on the evaluation set. * depict performance after adding frames from D_1000601.wav to the test files.

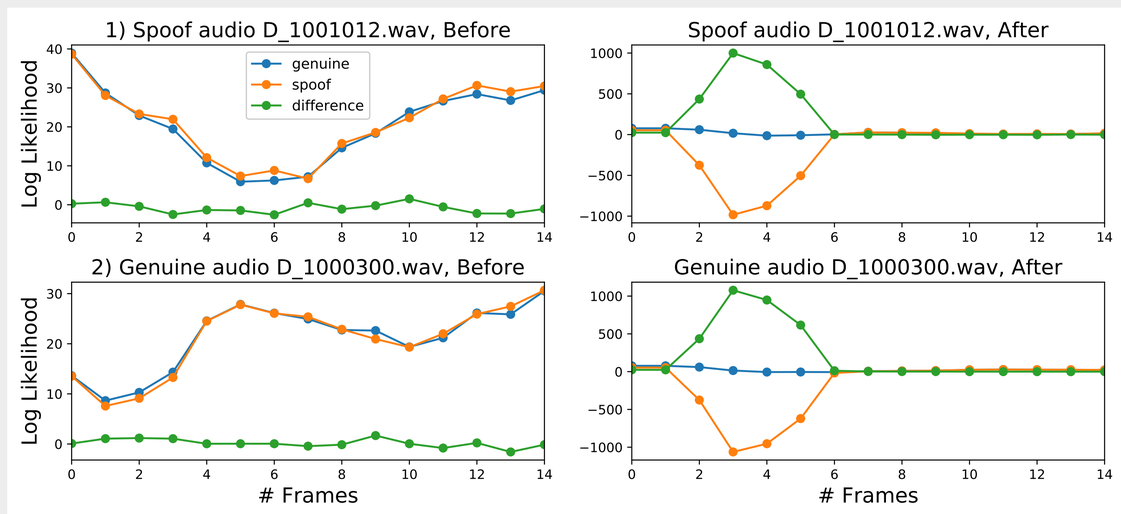
	IMFCC	MFCC	LFCC	RFCC	SCMC	CQCC
	17.43	26.02	17.61	16.67	14.82	17.78
*	34.46	35.95	38.23	34.22	44.44	18.71

5 Analysis

We look at our best GMM system that use SCMC-features. How is the log likelihood distributed over the frames of a test signal? The figure below shows **frames 3-6 of a genuine file D_1000601.wav** have a very low conditional probability in the spoof model.



If we add those particular frames of D_1000601.wav to any test file, the model will always choose 'genuine'. We demonstrate this on two examples: **1) Spoof file** D_1001012.wav correctly classified as spoof. **2) Genuine file** D_1000300.wav misclassified as spoof.



Next, we apply this across all the test files and observe a **dramatic change in the EER**. The results are shown in the second row of Table 2.

6 Intervention

Two approaches to mitigate this problem: 1) remove initial 60ms from any test file; 2) remove initial 60ms from all files and retrain model. Table 3 below show the results.

Approach	IMFCC	MFCC	LFCC	RFCC	SCMC	CQCC
1	19.18	31.79	21.46	19.85	17.98	19.79
2	19.10	31.9	21.06	20.1	17.7	19.35

7 Conclusion and future work

The ASVspoof 2017 dataset [1] contains a **cue in its 'genuine' files:** low energy frames at the beginning. Knowledge of such cues can lead to manipulate class predictions. An updated version of the database has been released [3] that has fixed this problem. We are now working in the following directions: 1) Design and analysis of replay countermeasures in an end-to-end setting. 2) Improve the benchmark on ASVspoof 2017 dataset.

[1] Kinnunen et. al. The ASVspoof 2017 Challenge: Assessing the Limits of Audio Replay Attack Detection in the Wild. In *Proc. Interspeech 2017.*

[2] Kinnunen et. al. RedDots Replayed: A New Replay Spoofing Attack Corpus for Text-dependent Speaker Verification Research. In *ICASSP 2017*.

[3] H. Delgado et. al. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. In *Speaker Odyssey 2018*.