Automatic Speech Verification Against Replay Attacks

Yu Chu **An-Chen Li** **Yuan Tang** **Yi-Yu Zheng**
Electrical and Computer Engineering
Carnegie Mellon University
{yuchu, anchenl, yuantang, yiyuz}@andrew.cmu.edu

Abstract

Automatic Speech Verification system serves for the security of these voice biometric systems for real-world applications, which suffers from various attacks including Text-to-speech, voice conversion and replay. The ASVSpoof 2017 dataset focuses on detection of replay spoofing audio from non-replay ones. The previous speech verification system works heavily on SVM and deep neural networks, with a benchmark of 30.6% EER given by ASVSpoof 2017 organizer. The spoofing detection task has two stages, where feature extraction is performed on input signal first. Features are then fed into training models, and evaluate through the test data. We have designed a series of experiments to achieve better performance with different feature extraction methods, models and parameters and achieved 27.8% EER using MFCC feature extraction and GMM classifier.

1 Introduction

Upon application, the speech recognition is divided into two specific stages: identification and verification. Verification is to determine from a voice sample if a person is who he or she claims to be. As biometric authentication technology reinvents itself, it is concerning that speech is susceptible to spoofing attacks. As a result, the speech verification system serves for the security of these voice biometric systems for real-world applications. In application, recording and replaying the audio is a common and cheap attack as it requires no expertise in performing such attacks. Therefore it is important for an ASV system to detect such replay attacks, which is the focus of this project, automatic speech verification against replay spoofing attacks.

In replay attack, noises exist due to the recording environment, playback and recording devices in playback and rerecording phase. These noises mark the key difference of a replay and non-replay audio. As the noises are hard to detect by human ear, machine learning models can be a good detection method. In replay detection, the model is trained to detect a replay audio, which makes the task a binary classification.

In the project, we utilized the ASVspoof 2017 dataset [1]. Based on the dataset, we investigated common feature extraction method, MFCC, LFCC and CQCC; as well as different classifiers, including SVM and GMM, to find out the best training setup in detecting replay attacks.

2 Related Work

Previous speech verification systems mainly work on SVM, CNN, and other neural networks. Various analysis illustrated the idea to train and test models within direct usage of data in waveform format. As end-to-end learning treats the entire system as a black box, sufficient data is a prerequisite for it

to work well on target task. Multiple end-to-end neural networks embedded systems were studied. The best performing model submitted to ASVspoof challenge [1] achieved 6.73% EER while the CQCC-GMM method achieved 30.6% EER, which could be used as a benchmark for our research.

3 Dataset

ASVSpoof 2017 dataset is used for the automatic speech verification task. It contains speech data collected from 179 replay sessions in 61 unique replay configurations, with label distribution shown in the following table. A replay configuration refers to a unique combination of room, replay device and recording device, while a replay session refers to a set of source files, which share the same replay configuration. The sampling rate of audio files are 16kHz, and stored in 16-bit format.

Table 1: Dataset Distribution

	Total	non-replay corpus	replay corpus
Training Set	3014	1507	1507
Validation Set	1710	760	950
Evaluation Set	13306	1298	12008

4 Methodology

4.1 Feature Extraction

4.1.1 Mel-frequency cepstral coefficient (MFCC)

For speech recognition, MFCC is the most commonly used acoustic features. It takes human perception sensitivity with respect to frequencies into consideration, and therefore regarded to be the greatest possible human ear approximation. MFCC processes the data in the following step: window the signal, apply the Discrete Fourier Transform (DFT), take the log of the magnitude, and then warp the frequencies on a Mel scale, followed by applying the inverse discrete cosine transform (DCT). In our experiments, we extract MFCC features with dimensions 13.

4.1.2 Linear Frequency Cepstral Coefficient (LFCC)

LFCC and MFCC are nearly equivalent [3]. The sole difference is in the filter bank, because the LFCC filter bank coefficients cover all speech frequency ranges evenly and value them equally. We extract LFCC features with dimensions 13 in the experiments.

4.1.3 Constant Q Cepstral Coefficient (CQCC)

M. Todisco proposed constant Q cepstral coefficients (CQCCs) [4] that couples constant Q transform (CQT) with cepstral analysis, and obtains excellent performance for both known and unknown spoofing attacks.The CQT was initially proposed in music processing [5]. It processed different frequencies with variable resolution which means higher resolution in higher frequencies while lower resolution in lower frequencies.
In our experiment, CQCC features of dimension 60 were extracted.

4.2 Classification Model

4.2.1 Gaussian Mixture Model (GMM)

A Gaussian mixture model [6] is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions that has no known parameters. It is a universally used model for generative unsupervised learning or clustering based on the optimization strategy. The advantage of Mixture models is that they can learn the sub-populations automatically.

In our experiments, we trained a GMM for each positive and negative samples respectively, where positive is the real person speaking sample and negative refers to the replay sample pre-recorded. In

the testing step, we applied the trained GMMs to the testing audio features to get similarity scores. Comparing the 2 similarity scores, if the scores from the GMM trained on positive samples is higher, we would inference that the testing audio is positive; if, on the contrary, we would inference that the testing audio is negative.

In our experiment, We used 144 Gaussian components to model audio features. We found that increasing the Gaussian components would increase the system accuracy. However, there is a trade-off between accuracy and the computational complexity. More components would lead to a higher computational complexity.

4.2.2 Support Vector Machine (SVM)

The support vector machine is a binary classification model to optimize the maximum margin hyperplane. Kernels could utilize on linear, nonlinear, polynomial. SVM tends to be resistant to overfitting, and has been applied to text classification, speaker identification successfully.
In our experiments, we utilized a SVM with radial basis function (RBF) kernel for training.

4.2.3 Hidden Markov Model (HMM)

The HMM is based on augmenting the Markov chain. A Markov chain is a model that tells us something about the probabilities of sequences of random variables, states. HMM is a straightforward application of the Bayesian classification framework, with the HMM being used as the probabilistic model describing data. HMM can be used in an unsupervised fashion too, to achieve something akin to clustering. In our experiments, we trained HMM with 1000 iterations and 4 states, but achieved unsatisfactory results.

5 Result

5.1 Evaluation Metric

Accuracy Accuracy can be used as a metric as the automatic speech verification task is a binary classification problemm.

$$Accuracy = \frac{Correctly\ Classifed\ Attempt}{Number\ of\ Identification\ Attempt} \tag{1}$$

Area Under Curve (AUC) AUC refers to the area under the ROC curve, indicating how well the probabilities from the positive classes are separated from the negative classes.

Equal Error Rate (EER) EER is a measure to evaluate the performance of biometric security systems. EER is the point where false acceptance rate (FAR) and false rejection rate (FRR) are minimal and optimal, that is, the intersection of two curves plotted with respect to different thresholds.

$$FAR = \frac{Number\ of\ False\ Acceptance}{Number\ of\ Identification\ Attempt} \tag{2}$$

$$FRR = \frac{Number\ of\ False\ Rejection}{Number\ of\ Identification\ Attempt} \tag{3}$$

5.2 Performance

Since speech verification tasks have two stages, feature extraction and classification. We experimented different combinations between feature extraction methods and classifiers.

To better investigate the effect of Gaussian components on the GMM performance, we have also conducted experiments with different number of Gaussian components on data of length 50 frames for training efficiency. As the best performing GMM model was trained using all frames (around 300 frames for each audio), the performance in below table is slightly worse. We used MFCC feature extraction in the following experiments.

Table 2: Experiment Performance with Different Feature Extraction Method

Feature Extraction	Classifier	Accuracy	AUC	EER
MFCC	SVM	0.702	0.762	0.297
	GMM	**0.741**	**0.821**	**0.278**
LFCC	SVM	0.690	0.751	0.309
	GMM	0.644	0.702	0.355
CQCC	SVM	0.581	0.616	0.418
	GMM	0.593	0.612	0.406

Table 3: Experiment Performance with Different Gaussian Components

Number of Gaussian Component	Accuracy	AUC	EER
2	0.682	0.743	0.318
5	**0.691**	**0.750**	**0.309**
8	0.671	0.730	0.329
16	0.657	0.705	0.343
32	0.639	0.691	0.361
72	0.639	0.704	0.361
144	0.648	0.709	0.352

6 Discussion and Analysis

6.1 Performance Comparison

Through the series of experiments, the combination of MFCC feature extraction method and GMM has achieved the best performance, outperforming the benchmark of 30.6% on the dataset.
As compared with the best performance of 6.72% from the challenge, we found there are salient modeling limitation on machine learning algorithms. In nature of speech signals, a model that utilizes continuous and sequential behaviour would possible lead to better performance. Side information would also be valuable add-on to speech verification task and it provides text information in support of the speech modeling process.

6.2 Speed Comparison

Despite the performance, latency is important in real-world applications. SVM took 1.6 seconds to train on the training set and 5.9 seconds to evaluate, while GMM took 227.5 seconds to train and 0.7 seconds to evaluate. Therefore when applied in real-world scenario where a speech verification model is pre-trained and only does inference in real time, GMM is a more suitable model. However, SVM might be in consideration if frequent model updates are required.
Compared to other deep neural network method, time can be saved in both training and inference stage.

6.3 Effect of Number of Gaussian Components

As we have anticipated, the performance of GMM trained on data truncated into 50 frames is worse than the best performance on full frame dataset. We can still observed the effect of number of Gaussian components, in this particular experiment setup, the performance is best when there are 5 Gaussian components, and decreases when there are more or less components.

6.4 Choice of Feature Extraction Method Matters

Another finding from the experiments is that choice of feature extraction method has a greater impact on classification performance rather than the choice of classifier. It can be observed that the performance of SVM and GMM using the same feature extraction method vary around 5% in AUC, while there is a significant drop of AUC when CQCC feature extraction is used.

References

[1] Kinnunen, Tomi; Sahidullah, Md; Delgado, Héctor; Todisco, Massimiliano; Evans, Nicholas; Yamagishi, Junichi; Lee, Kong Aik. (2018). The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database, Version 2, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2332.

[2] Kinnunen, Tomi Sahidullah, Md Delgado, Héctor Todisco, Massimiliano Evans, Nicholas Yamagishi, Junichi Lee, Kong Aik. (2017). The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. https://doi.org/10.21437/Interspeech.2017-1111.

[3] Zhou, Xinhui, Garcia-Romero, Daniel; Duraiswami, Ramani; Espy-Wilson, Carol ; Shamma, Shihab. (2011). Linear Versus Mel Frequency Cepstral Coefficients for Speaker Recognition. 2011 IEEE Workshop on Automatic Speech Recognition Understanding

[4] M. Todisco H. Delgado and N. Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. in Odyssey 2016 - The Speaker and Language Recognition Workshop, 2016.

[5] J. C. Brown, Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. Journal of the Acoustical Society of America, vol. 105, no. 3, p.

[6] Reynolds, Douglas A. (2009). "Gaussian Mixture Models." Encyclopedia of Biometrics.

[7] Senoussaoui, Mohammed Cardinal, Patrick Dehak, Najim Koerich, Alessandro. (2016). Native Language Detection Using the I-Vector Framework. 2398-2402. https://doi.org/10.21437/Interspeech.2016-1473.