Grenze International Journal of Engineering and Technology, Jan Issue

Effectiveness of different Spectral Features in Replay Attack Detection - An Experimental Study

Bhaskar Jyoti Chutia[1] and Utpal Bhattacharjee[2]
[1-2] Rajiv Gandhi University, Arunachal Pradesh, India
Email: {bhaskar.chutia, utpal.bhattacharjee}@rgu.ac.in

***Abstract*—The Spoofing attacks poses serious threats to Automatic Speaker verification (ASV) systems. Replay attack is one of such spoofing attacks which is a low effort yet high-risk attack due to ease of availability as well as widespread usage of smart phones, recording and playback devices. In replay spoofing attack, the attacker replays a pre-recorded speech samples of a genuine speaker to get unauthorized access. This study is focused on experimental analysis of different features with Gaussian Mixture Model (GMM) for development of countermeasures system for detection of replay spoofing attacks on ASV systems. GMM is a standard widely used as well as effective classifier for baseline countermeasures in ASV systems. The results are compared based on Equal Error Rate (EER) metric. This work presents a comparison of various spectral features for the most common spoofing attack i.e. replay spoofing attack detection based on its performance over the released ASVspoof challenge dataset of 2017, which is a benchmark dataset dedicated to address replay spoofing attacks only.**

***Index Terms*— Replay Attack, Spoofing Attacks, Spectral Features, ASV system.**

I. INTRODUCTION

The purpose of an ASV system is to determine the veracity of an identity claim made by a speaker. This system faces a two-class classification challenge, wherein it needs to establish whether the claimed identity matches that of the target speaker or not. While a typical ASV system can withstand rudimentary impostors with minimal effort, it remains susceptible to more sophisticated attacks referred to as presentation attacks or spoofing attacks. In spoofing attacks, an adversary (attacker) impersonates the target speaker in order to obtain unauthorized access in to a protected system [6-8]. Given the ease with which a person's biometric information can be acquired, it is inevitable that spoofing attacks will occur across various domains [9]. Some of the common spoofing attacks are Speech Synthesis attack, Voice Conversion attack, Replay attack etc. These spoofing attacks present major security concerns for ASV systems and therefore a need of countermeasure (CM) system has arisen to differentiate genuine speech from spoofed speech. A brief introduction of different spoofing attacks is given below.

A. Impersonation

It is one of the most primitive approaches used to spoof ASV systems. In simple words it uses human mimicking to bypass the verification process. Before the development of necessary software tools and technology to launch other sophisticated spoofing attacks, human-altered voices were the only way to penetrate authentication of an ASV system. In this type of scenario, attacker attempts to imitate voice timbre and prosody of a specific target

Grenze ID: 01.GIJET.10.1.556_2

speaker using manual techniques, without relying on computer-assisted technologies. Amateur impersonators can easily modify their speech to deceive ASV systems, however this works if and only if their speech already bears natural resemblance to that of the target speaker. It has been observed that linguistic expertise is generally not advantageous, except in cases where the target speaker's voice significantly differs from that of the impersonator [1-2]. Impersonation primarily involves mimicking prosodic as well as stylistic cues. Impersonation is potentially better in deceiving human listeners compared to current advanced ASV systems [3].

B. Speech synthesis (SS)

Please Speech synthesis attacks on automatic speaker authentication systems are a type of security threat in which attackers use artificial voice technology to imitate the voice of legitimate users and obtain unauthorized access permission to system. This is generally done by utilizing a recording of the legitimate user's voice or by using text-to-speech technology to generate speech in the style of the legitimate user. The attacker's goal is to mislead the ASV system into anticipating that the synthesized speech is actually from a legitimate user.

C. Voice conversion attacks (VC)

The objective of voice conversion attacks is to alter speech of a particular speaker to bear a resemblance, to some extent, to that of another targeted speaker [4]. Unlike speech synthesis systems that rely on text input, a voice conversion system utilizes speech signal from a natural source as its input. Generally, voice conversion entails both spectral mapping, which deals with the voice's timbre, and prosody conversion, which handles prosodic attributes like fundamental frequency and duration. This process involves manipulating original speech's spectral characteristics as well as prosody to match those of the desired target speaker.

D. Replay attack

This refers to a method of deceiving where an enemy targets an ASV system by employing a previously recorded vocal sample obtained from an authentic individual. The vocal sample can be any secretly obtained recording, and even a combination of different speech samples taken from shorter segments, to overcome ASV systems that rely on specific texts. [5]. Replay is a straightforward nature of spoofing. It does not demand any particular expertise and knowledge in speech processing. Furthermore, thanks to the widespread availability of affordable and top-notch recording gadgets like smartphones, replay attacks are definitely the most common spoofing attack and consequently pose a notable danger. "Fig.1" illustrates an instance of a real-world replay attack.

Researchers have been working to protect these systems by developing various countermeasures. The countermeasures against spoofing attacks consist of two components: a front-end responsible for analyzing the speech fetched to the system, and a back-end that assesses whether the fetched speech is from the genuine speaker or artificially generated. The front-end, also known as the feature extraction unit, is designed to identify relevant information from the speech signal that indicates signs of conversion or synthesis. On the other hand, the back-end employs a modelling technique to accurately represent these speech features. Numerous counter measure techniques have been suggested for enhancing the performance of spoofing attack detection in both the front-end and back-end components. In this paper, we are focusing only on the comparison of different spectral features for replay attack on ASV system.

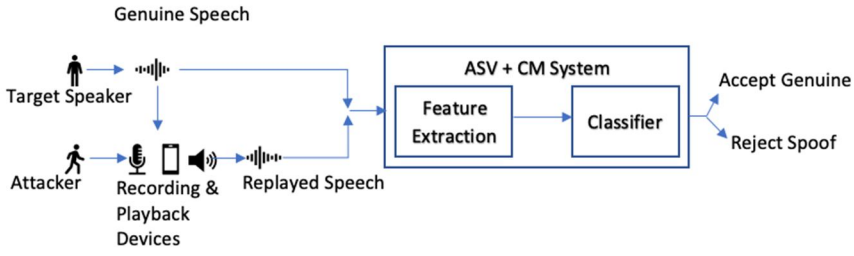

Figure 1. A replay attack scenario along with target speaker and Automatic Speaker Verification System

II. Description of ASVspoof 2017 Challenge dataset

This experimental study uses speech data from ASVspoof challenge dataset of 2017, which includes three sets of data: training set, development set and evaluation set. "Table 1" shows the instances of speech samples in each set. Each set includes a combination of three factors: the environment where the recording was made, the device

used to replay the recording, and the device used to make the original recording. The training set includes four different conditions, while the development set has ten conditions, one of which is also present in the training set. The ASVspoof challenge of 2017 focuses on different types of replay attacks that have not been seen before, so the development set includes some conditions that do not match those in the training set.

TABLE I. DESCRIPTION OF ASVSPOOF 2017 CHALLENGE DATASET

Dataset	No. of Speakers	No. of Sessions	No. of Utterances		
			Genuine Speech (G)	Replay Speech (R)	Total (G + R)
Training (T)	10	6	1508	1508	3016
Development (D)	8	10	760	950	1710
Evaluation (E)	24	161	1298	12008	14220
Total (T+D+E)	42	177	3566	14466	18946

The ASVspoof 2017 training set is typically employed to train the replay spoofing detector, particularly focusing on training the acoustic models (such as GMMs or DNNs) for both genuine speech and spoof speech categories. The ASVspoof 2017 development set, on the other hand, is utilized to optimize the system's hyper-parameters and refine its performance. To assess the actual performance of system, the ASVspoof 2017 evaluation set is utilized, consisting of both genuine and replay spoof speech. The evaluation set contains replay spoof signals that are unfamiliar to the detector, which allows for testing the system's ability to generalize. For more details regarding the database, refer to the comprehensive study [12]. Additionally, a subsequent version of the database has been released, featuring corrected recording environment labels on the development and evaluation protocol files. Further information can be found in [13].

III. EXPERIMENTAL SETUP

A. Baseline System

The system for replay attack detection, utilized in this experimental study is based on the Baseline system, which bears resemblance to the system put forth by the ASVspoof 2017 challenge organizers [10]. The ASV system employs constant Q cepstral coefficients (CQCC) as the front-end feature representation, while the back-end employs a GMM, which is trained with the maximum likelihood criterion.

The speech samples undergo analysis employing the CQCC feature extraction technique, which retrieves features from the power spectrum of speech frames. However, it deviates from the conventional method by employing a distinct approach. Instead of utilizing the short-term Fourier transform (STFT), the power spectrum is determined using a constant-Q transform (CQT) as a substitute for the discrete Fourier transform (DFT). By employing the CQT, a more detailed examination of the speech is accomplished, providing superior temporal resolution for higher frequencies and enhanced spectral resolution for lower frequencies. Subsequently, the logarithmic CQT power spectrum, which has been uniformly re-sampled, undergoes a discrete cosine transform to extract the CQCC features. For additional elaboration on the extraction process of CQCC, refer to [11].

After extracting the CQCC features, the initial system uses a GMM classifier to identify replay attacks. A GMM is assigned to each class, namely genuine and replay, and the training features from each class are employed to train their respective GMMs using the expectation maximization (EM) algorithm.

During the testing phase, a set of feature vectors, $X = \{x_1, x_2, \ldots, x_T\}$, obtained from a test speech sample, along with Gaussian Mixture Models of each class, are employed to calculate the log-likelihood ratio (LLR) score, also known as the decision score, as depicted in "(1)".

$$\text{LLRGMM}(X) = \log p(X|\lambda_{\text{genuine}}) - \log p(X|\lambda_{\text{replay}}) \quad (1)$$

In this context, p represents the likelihood function, and $p(X|\lambda)$ represents the GMM likelihood of X in relation to model λ. Specifically, λ_{genuine} and λ_{replay} refer to the parameters corresponding to genuine and spoof models, respectively.

B. Performance Criteria

The equal error rate (EER) is widely utilized as a performance metric for biometric systems. The ASV spoof 2017 challenge organizers used EER as performance metric for replay attack detection [10]. The Equal Error

Rate is the error rate that occurs at the threshold point where the false acceptance rate (FAR) as well as false rejection rate (FRR) are equal. In other words, the EER is the point at which the rates of falsely accepting genuine claims and falsely rejecting valid claims are the same. When spoofed speech of a replay attacker is accepted as genuine target speaker, it is considered as FAR. On the other hand, when a genuine speech of a target speaker is rejected as spoofed speech of an attacker, it is called as FRR. There is a trade-off between both of these error rates. FAR and FRR becomes equal at a threshold value. It is known as EER. The lower EER value represents superior verification accuracy of spoofing detection system. Therefore, higher EER value is not desirable in performance evaluation of biometric systems.

IV. Spectral Feature Extraction

A While features based on the magnitude of speech signals have been shown to be productive and are commonly used, it is also crucial to investigate the use of features based on the phase of speech signals, since they may contain additional information that is not present in the magnitude-based features. Magnitude based features such as Mel Frequency Cepstral Coefficients (MFCC), Rectangular Frequency Cepstral Coefficients (RFCC), Linear Prediction Cepstral Coefficients (LPCC) etc. In following subsections, different spectral features or their extraction techniques are discussed in brief:

A. Constant Q Cepstral Coefficients (CQCC)

The extraction process of MFCC feature starts by applying a pre-emphasis filter to the speech signal, and then dividing it into frames of 20ms, with a 10ms shift between each of the frame. The Hamming window function is applied to each frame individually, multiplying them. Then, the fast Fourier transform (FFT) is utilized to calculate the power spectrum of the frames. The power spectrum undergoes processing by a Mel-filter bank consisting of 27 channels. Subsequently, the logarithm of the filter bank outputs is computed. Finally, the logarithmic filter bank outputs are subjected to the Discrete Cosine Transform (DCT), yielding the MFCC features. MFCC-GMM model showed EER of 13.74% and 34.39% for development and evaluation dataset respectively [21].

B. Mel Frequency Cepstral Coefficients (MFCC)[15-18]

The extraction process of MFCC feature starts by applying a pre-emphasis filter to the speech signal, and then dividing it into frames of 20ms, with a 10ms shift between each of the frame. The Hamming window function is applied to each frame individually, multiplying them. Then, the fast Fourier transform (FFT) is utilized to calculate the power spectrum of the frames. The power spectrum undergoes processing by a Mel-filter bank consisting of 27 channels. Subsequently, the logarithm of the filter bank outputs is computed. Finally, the logarithmic filter bank outputs are subjected to the Discrete Cosine Transform (DCT), yielding the MFCC features. MFCC-GMM model showed EER of 13.74% and 34.39% for development and evaluation dataset respectively [21].

C. Inverted Mel Frequency Cepstral Coefficients (IMFCC)[20]

In the low-frequency region, the mel-scale exhibits closely spaced intervals, while in the inverse mel-scale, the spacing is reversed. According to certain researchers, there is an argument that the mel-scale is densely packed in the low-frequency range, which implies that the MFCC might overlook certain discriminatory pieces of evidence related to high-frequency characteristics in distinguishing between genuine and replay speech signals. As such the IMFCC feature are explored for detection of replay signals. The computational procedure of IMFCC feature is exactly similar to traditional MFCC feature except the use of inverse mel-filter. The IMFCC-GMM [21] system showed EER of 4.83% on development dataset, which is better than all other spectral features performance. However, the same system unable to do better than baseline CQCC-GMM in terms of evaluation dataset. It EER stands at 28.59% which is comparatively higher than that of baseline model.

D. Linear Frequency Cepstral Coefficients (LFCC)

In search of good features for detection of replay signals, people explored cepstral processing with linearly spaced triangular and rectangular filters. In [23], the potential of LFCC features is evaluated on ASVspoof2017 dataset. GMM classifier along with LFCC features (40 static, 40 delta, 40 delta-delta) provide an EER of 16.62% on evaluation dataset. The baseline system performed better than LFCC-GMM with respect to development dataset. The LFCC-GMM shows 10.58% EER on development dataset.

E. Rectangular Filter Cepstral Coefficients (RFCC) [24]

The potential of RFCC features is evaluated on ASVspoof2017 database by considering GMM as backend. RFCC features (30 static, 30 delta, 30 delta-delta) with GMM classifier provide 9.32% and 11.90% EER on ASVspoof 2017 development dataset and evaluation dataset respectively. Previous works suggest the use of linearly spaced triangular and rectangular filter-banks are more effective for detection of replay signals.

F. Linear Prediction Cepstral Coefficients (LPCC)[25]

The LPCC features have been found very promising in terms its results on development dataset of ASVspoof 2017 challenge database in the work of [19]. EER of LPCC-GMM model on development set is 5.94% as shown in table 2. Their work reported an EER of 25.20% by using LPCC features along with GMM for evaluation dataset.

G. Subband Spectral Flux Coefficients (SSFC) [28]

These features serve as a substitute representation of modulation characteristics in order to detect replay signals. The Spectral Subband Centroids (SSFC) algorithm focuses on identifying the harmonic consistency within the signal, which can be potentially compromised in spoofed speech caused by replay techniques. When combined with a GMM classifier, the SSFC [22] yields improved outcomes than the baseline CQCC-GMM approach. Specifically, on the ASVspoof 2017 evaluation dataset of the ASVspoof2017 database, the SSFC-GMM achieves an EER of 22.38%. However, it falls short of surpassing the baseline model on the development dataset, as indicated by a higher EER of 12.81%.

H. High-Frequency Cepstral Coefficients(HFCC)

As mentioned in [27], the authors use a window length of 30 ms and a hop-size of 15 ms for their framing step. The final feature vector includes the first 30 coefficients of the discrete cosine transform along with their first and second derivatives. To compare the performance of the HFCC feature to the baseline, they use a 512-component GMM that is trained for both the genuine and spoof classes. They use the log-likelihoods of the two models to generate scores for test utterances. HFCC-GMM outperforms the CQCC baseline on both Dev and Eval sets with EERs of 5.9% and 23.9%, respectively as shown in “Table II”.

I. Subband Spectral Centroid Frequency Coefficients (SCFCs) and Subband Spectral Centroid Magnitude Coefficients (SCMCs) [26, 29-30]

The researchers mentioned in [19] observed that replay speech exhibits higher spectral centroid deviation (SCD) scores compared to genuine speech. To enhance the detec-tion of replay signals, they propose incorporating SCD alongside SCFC and SCMC as supplementary evidence. The authors utilize SCD as the initial step in feature ex-traction and then apply the same parameters to extract SCMC and SCFC. The EERs for SCFC and SCMC on the evaluation dataset of ASVspoof are reported as 24.83% and 11.49%, respectively [19].

V. RESULTS

In this work, the experiments are carried out for individual spectral features like CQCC and MFCC, which are again compared with results of other spectral features carried out in recent works as mentioned in previous section 4. Experiments are carried out on both development as well as evaluation dataset of ASVspoof 2017 challenge dataset. GMM is taken as default backend in all the cases to achieve an unbiased comparison. In the “Table II”, we can see performance of each of the spectral features in terms of EER.

TABLE II. PERFORMANCE STATISTICS BASED ON EER (%) RESULTS FOR DIFFERENT SPECTRAL FEATURES ON ASV SPOOF 2017 DATASET

Feature	Development	Evaluation
CQCC (Baseline)	10.42	28.48
MFCC	13.74	34.39
IMFCC	**4.83**	28.59
LFCC	10.58	16.62
LPCC	5.94	25.20
HFCC	5.90	23.90
SCFC	24.51	24.83
SSFC	12.81	22.38
SCMC	9.32	**11.49**
RFCC	6.91	11.90

VI. CONCLUSION

We have presented results from an experimental comparison of several well-known spectral features for replay spoofing attack detection in Automatic Speaker Verifica-tion systems. We have conducted our experiments using ASV spoof 2017 database. This comparative experimental study led to following findings regarding spectral feature extraction for replay attack detection:

SCMCs found to be most effective features for replay speech detection on evalua-tion dataset. It shows ERR of 11.49%. The SCMC-GMM system had the lowest error rates and the most consistent performance across different experiments. However, the best performance on ASVspoof 2017 development dataset is achieved by IMFCC-GMM model with EER 4.83%.

Performance of modulation features like SSFC and SCMC are found to be better than baseline CQCC-GMM on both the datasets, whereas performance of SCFC is constant irrespective of the dataset considered. Future work may focus on, a score-level fusion by a weighted sum of all the scores of these systems, which can definite-ly give promising results.

REFERENCES

[1] Lau, Yee Wah, Michael Wagner, and Dat Tran. "Vulnerability of speaker verification to voice mimicking." In Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004., pp. 145-148. IEEE, (2004).

[2] Lau, Yee W., Dat Tran, and Michael Wagner. "Testing voice mimicry with the YOHO speaker verification corpus." In International conference on knowledge-based and intelligent information and engineering systems, pp. 15-21. Springer, Berlin, Heidelberg, (2005).

[3] Hautamäki, Rosa González, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. "Comparison of human listeners and speaker verification systems using voice mimicry data." Target 4000 (2014): 5000.

[4] Evans, Nicholas WD, Federico Alegre, Zhizheng Wu, and Tomi Kinnunen. "Anti-spoofing, Voice Conversion." (2015): 115-122.

[5] Villalba, Jesus, and Eduardo Lleida. "Preventing replay attacks on speaker verification systems." In 2011 Carnahan Conference on Security Technology, pp. 1-8. IEEE, (2011).

[6] Ergünay, Serife Kucur, Elie Khoury, Alexandros Lazaridis, and Sebastien Marcel. "On the vulnerability of speaker verification to realistic voice spoofing." In 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1-6. IEEE, (2015).

[7] Evans, Nicholas WD, Tomi Kinnunen, and Junichi Yamagishi. "Spoofing and countermeasures for automatic speaker verification." In Interspeech, pp. 925-929. (2013).

[8] Marcel, Sébastien, Mark S. Nixon, Julian Fierrez, and Nicholas Evans, eds. Handbook of biometric anti-spoofing: Presentation attack detection. Vol. 2. Cham, Switzerland: Springer, (2019).

[9] Galbally, Javier, Sébastien Marcel, and Julian Fierrez. "Biometric antispoofing methods: A survey in face recognition." IEEE Access 2 (2014): 1530-1552.

[10] Kinnunen, Tomi, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection." (2017).

[11] Todisco, Massimiliano, Héctor Delgado, and Nicholas WD Evans. "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients." In Odyssey, vol. 2016, pp. 283-290. (2016).

[12] Kinnunen, Tomi, Md Sahidullah, Mauro Falcone, Luca Costantini, Rosa González Hautamäki, Dennis Thomsen, Achintya Sarkar et al. "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research." In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp. 5395-5399. IEEE, (2017).

[13] Delgado, Héctor, Massimiliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, and Junichi Yamagishi. "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements." In Odyssey 2018-The Speaker and Language Recognition Workshop. (2018).

[14] Todisco, Massimiliano, Héctor Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification." Computer Speech & Language 45 (2017): 516-535.

[15] Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28, no. 4 (1980): 357-366.

[16] Gupta, Shikha, Jafreezal Jaafar, WF Wan Ahmad, and Arpit Bansal. "Feature extraction using MFCC." Signal & Image Processing: An International Journal 4, no. 4 (2013): 101-108.

[17] Jelil, Sarfaraz, Rohan Kumar Das, SR Mahadeva Prasanna, and Rohit Sinha. "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features." In Interspeech, pp. 22-26. (2017).

[18] Chen, Zhuxin, Zhifeng Xie, Weibin Zhang, and Xiangmin Xu. "ResNet and Model Fusion for Automatic Spoofing Detection." In Interspeech, pp. 102-106. (2017).

[19] Font, Roberto, Juan M. Espín, and María José Cano. "Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 Challenge." In Interspeech, pp. 7-11. (2017).
[20] Chakroborty, Sandipan, Anindya Roy, and Goutam Saha. "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks." International Journal of Electronics and Communication Engineering 2, no. 11 (2008): 2554-2561.
[21] Liu, Meng, Longbiao Wang, Jianwu Dang, Seiichi Nakagawa, Haotian Guan, and Xiangang Li. "Replay attack detection using magnitude and phase information with attention-based adaptive filters." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6201-6205. IEEE, (2019).
[22] Patil, Hemant A., and Madhu R. Kamble. "A survey on replay attack detection for automatic speaker verification (ASV) system." In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1047-1053. IEEE, (2018).
[23] Tak, Hemlata, and Hemant A. Patil. "Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection." In INTERSPEECH, pp. 726-730. (2018).