International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

A REVIEW ON REPLAY SPOOF DETECTION IN AUTOMATIC SPEAKER VERIFICATION SYSTEM

Ajila A[1], Smitha K S[2]

[1]M. Tech Student, Dept. of ECE, LBS Institute for Women, Kerala, India
[2]Assistant Professor, Dept. of ECE, LBS Institute for Women, Kerala, India

---

Abstract - *Spoof detection in the Automatic Speaker Verification (ASV) system is an essential problem nowadays. Among spoofing, replay possesses a greater threat to the ASV system. This paper presents a survey on spoofing detection under the case of replay. Replay attacks in the ASV system lead to the performance degradation of the entire system. There have been many methods developed for detecting replay spoof in past works. This paper reviews the performance of the best anti-spoofing techniques used in ASV systems.*

Key Words: *Automatic Speaker Verification (ASV), Spoof detection, Replay, Countermeasures.*

1. INTRODUCTION

Voice is one of the most important human biometrics used in everyday communication. Its unique characteristics play a major role in conveying the identity of an individual. Voice biometrics is considered as a behavioral characteristic. The ASV system consists of two major parts namely, speaker verification and spoof detection system. Speaker verification accepts or rejects the claimed identity based on speech sample and spoof detection system checks whether the speech sample is genuine or spoofed. Like any other biometrics, ASV is also vulnerable to spoofing attacks [1]. In an ASV system, nine possible attack points are classified as direct attacks, also known as spoofing attacks and indirect attacks. For an indirect attack, the attacker needs access to the inside of the ASV system. There are mainly five spoofing attacks namely, impersonation, Voice Conversion (VC), Speech Synthesis (SS), twins and replay. Impersonation attacks are made by performing human-altered voices, where attacker tries to imitate exactly like the target speaker. In VC attacks, the attacker tries to replicate the target speaker's voice by using any computer-aided technologies. SS, often referred to as Text-To-Speech (TTS) uses a technique where the speech is produced from the input text. In the twin's attack, the attacker tries to fools the system by providing a speech sample of his/her twin. The twin attacks are relatively low when compared with other spoof attacks. A replay attack is the easiest and simplest among spoofing since it does not need any computer expertise or complex algorithms.

The main blocks of an ASV system are pre-processing, feature extraction, classifier, and decision. In pre-processing, the input raw signals are processed to increase the efficiency in upcoming stages. The pre-processing techniques commonly used are noise removal, pre-emphasis, etc. The raw signals are transformed into some sort of parametric representation in the feature extraction stage. Feature extraction provides an understandable representation of an input signal. Commonly used feature extraction methods are Constant Q Cepstral Coefficients (CQCC), Mel-Frequency Cepstral Coefficients (MFCC), etc. When the features are extracted the next step is to decide whether the input speech is genuine or spoofed. A classifier helps to do this task. Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Convolutional Neural Networks (CNN) are some of the classifiers employed. Finally, the decision stage, where the input signal is accepted or rejected by the ASV system. This paper studies the works related to the ASVspoof 2017 challenge [2]. The baseline system implemented was the CQCC with GMM [3].

2. EXISTING METHODS OF SPEECH SPOOF DETECTION

In [4], Tharshini Gunendradasan, Buddhi Wickramasinghe, Phu Ngoc Le, Eliathamby Ambikairajah and Julien Epps propose a work which explains the use of spectral centroid based Frequency Modulation (FM) features which they called as Spectral Centroid Deviation (SCD) for the replay attack detection. They also extracted the Spectral Centroid Magnitude Coefficient (SCMC) features from the front-end of SCD along with Spectral Centroid Features (SCF). The work employs GMM as the back-end classifier. They introduced an FM feature extraction based on Linear Predictive Coefficients (LPC) model and the feature characteristics for genuine and spoofed speech were examined. An Equal Error Rate (ERR) of 15.68%, 12.34%, and 11.45% was obtained for SCMC with GMM, SCF with GMM and SCD with GMM systems respectively. The fusion score of the above three systems

produced an EER of 9.20%. This work provides an EER improvement of 60% than the CQCC baseline system.

Prasad A. Tapkir, Ankur T. Patil, Neil Shah, and Hemant A. Patil in [5], proposed new feature sets called Magnitude based Spectral Root Cepstral Coefficients (MSRCC) and Phase based Spectral Root Cepstral Coefficients (PSRCC). The classifiers they opted was GMM along with CNN. They conducted a study on both development set and evaluation set with MSRCC and PSRCC with GMM classifiers. An EER of 8.53% and 18.61% is obtained for MSRCC-GMM in the development set and evaluation set respectively. An EER of 35.53% is obtained for PSRCC-GMM in the development set and 24.35% is obtained for PSRCC-GMM in the evaluation set. The fused system MSRCC+PSRCC with GMM gave an EER of 6.58% and 10.65% in the development set and evaluation set respectively. When used CNN as a classifier, MSRCC-CNN gave an EER of 3.05% in the development set and 24.84% in the evaluation set. For PSRC-CNN the EER for development set and evaluation set were 36.21% and 26.81% respectively. The fused system, MSRCC+PSRCC with CNN gave an EER of 2.63% in development set and 17.76% in the evaluation set.

Sarfaraz Jelil, Rohan Kumar Das, S. R. M. Prasanna, and Rohit Sinha in Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features [6] uses a combination of features like glottal closure instants, epoch strength and the peak to sidelobe ratio of Hilbert envelope of linear prediction residual along with Instantaneous Frequency Cosine Coefficients (IFCC), CQCC and MFCC. This system used GMM as a classifier. First, they performed the five individual feature extraction methods. System 1 (S1) was based on Epoch Features (EF) calculated from the glottal activity regions. System 2 (S2) used the features of Peak to Side Lobe Ratio consisting of Mean and Skewness (PSRMS). System 3 (S3), System 4 (S4) and System 5 (S5) are based on IFCC features, CQCC features, and MFCC features respectively. The EER scores of the evaluation set in the systems S1, S2, S3, S4, and S5 are 28.66%, 28.90%, 35.19%, 19.58%, and 23.55% respectively. They also done various fusions of the above five systems. The fusion score of the combined systems (S1 + S2 + S3 + S4 + S5) gave an EER of 5.31% in development set and 13.95% in evaluation set.

In Audio Replay Attack Detection Using High-Frequency Features [7], Marcin Witkowski, Stanisław Kacprzak, Piotr Zelasko, Konrad Kowalczyk, Jakub Gałka proposed a system by detecting the replay attacks that was found in the high-frequency band of the replay recordings. Their work was based on modeling the sub-band spectrum and also deriving features from the linear prediction analysis. The high-frequency features like Inverse Mel Frequency Cepstral Coefficients (IMFCC), Linear Prediction Cepstral Coefficients (LPCC), and Linear Prediction Cepstral Coefficients residual (LPCCres) are selected for feature extraction and GMM as a classifier. The work was conducted in various frequency ranges, ranging from 16 to 8000 Hz. An EER of 4.48% was obtained for IMFCC, 3.38% for cepstrum and 6.37% for LPCCres in the evaluation set. A relative reduction in EER of 30% was obtained for the evaluation set.

Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev and Vadim Shchemelinin in their work [8] proposed an anti-spoofing system. They investigated the efficiency of the deep learning approaches like CNN and Residual Neural Network (RNN). The study was based on the SVM i-vector, Light Convolution Neural Network (LCNN) and the fusion of CNN and RNN systems. LCNN was conducted in three systems namely, $LCNN_{FFT}$ which is the truncated Fast Fourier Transform (FFT) system, $LCNN_{CQT}$ which is the Constant Q Transform (CQT) and $LCNN^{SW}_{FFT}$ which is the sliding window of the FFT system. These systems were used to estimate the GMM likelihood ratio scores. Among these, LCNN with truncated FFT features shows the best result with 7.37% EER and the fusion set system provided 6.73% EER in the evaluation set. These results show that there is a relative improvement of about 72% of the baseline system.

In [9], Weicheng Cai, Danwei Cai, Wenbo Liu, Gang Li, and Ming Li proposed a multiple replay spoofing countermeasure system. With the help of parametric sound reverberator and phase shifter, they converted the genuine speech signal into a replay speech signal then they replaced the general CQCC input with the spectrogram and this spectrogram is fed as the input to the deep residual network (ResNet). Fully-connected Deep Neural Network (FDNN) and Bi-directional Long-Short Term Memory (BLSTM) are employed as the classifiers. The BLSTM got an EER of 40.08% and the fusion score of CQCC-GMM (baseline), DA-CQCC-GMM (augmented CQCC-GMM) and ResNet gave an EER of 16.39%. This system shows an increment of 26% from the baseline system.

3. CONCLUSIONS

Voice biometrics are used for applications like telephone banking where security is the key. Since it is vulnerable to various attacks, it is important to maintain efficient countermeasures. The ASVspoof 2017 challenge mainly focuses on the replay attacks in speech. This work aims to provide a detailed description of various replay spoof detection methods. The researches show that introducing

efficient feature extraction and classifier techniques can make the spoof detection a lot effective.

REFERENCES

[1] Singh, Madhusudan, and Debadatta Pati. "Usefulness of linear prediction residual for replay attack detection." *AEU-International Journal of Electronics and Communications* 110: 152837, 2019.

[2] T. Kinnunen et al., "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," Proc. in INTERSPEECH, pp.2-6, 2017.

[3] T. Kinnunen et al., "ASVspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," Training, vol. 10, pp. 1508, 2017.

[4] Tharshini Gunendradasan, Buddhi Wickramasinghe, Phu Ngoc Le, Eliathamby Ambikairajah and Julien Epps, "Detection of Replay-Spoofing Attacks using Frequency Modulation Features", in INTERSPEECH, Hyderabad, pp. 636–640, 2018.

[5] Prasad A Tapkir, Ankur T. Patil, Neil Shah, Hemant A. Patil, "Novel Spectral Root Cepstral Features for Replay Spoof Detection, "APSIPA Annual Summit and Conference, Honolulu, Hawaii, USA, pp. 1945-1950, 2018.

[6] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in INTERSPEECH, Stockholm, Sweden, pp. 22–26, 2017.

[7] M. Witkowski, S. Kacprzak, P. Å˙zelasko, K. Kowalczyk, and J. GaÅ´Cka, "Audio replay attack detection using high-frequency features," in INTERSPEECH, Stockholm, Sweden, pp. 27–31, 2017.

[8] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev and Vadim Shchemelinin, "Audio replay attack detection with deep learning frameworks", in INTERSPEECH, Stockholm, Sweden, 2017.

[9] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in INTERSPEECH, Stockholm, Sweden, pp. 17–21, 2017.