Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Voice spoofing countermeasure for voice replay attacks using deep learning

Jincheng Zhou[1,2], Tao Hai[1,2,3*], Dayang N. A. Jawawi[3], Dan Wang[2,4], Ebuka Ibeke[5] and Cresantus Biamba[6*]

Abstract

In our everyday lives, we communicate with each other using several means and channels of communication, as communication is crucial in the lives of humans. Listening and speaking are the primary forms of communication. For listening and speaking, the human voice is indispensable. Voice communication is the simplest type of communication. The Automatic Speaker Verification (ASV) system verifies users with their voices. These systems are susceptible to voice spoofing attacks - logical and physical access attacks. Recently, there has been a notable development in the detection of these attacks. Attackers use enhanced gadgets to record users' voices, replay them for the ASV system, and be granted access for harmful purposes. In this work, we propose a secure voice spoofing countermeasure to detect voice replay attacks. We enhanced the ASV system security by building a spoofing countermeasure dependent on the decomposed signals that consist of prominent information. We used two main features— the Gammatone Cepstral Coefficients and Mel-Frequency Cepstral Coefficients— for the audio representation. For the classification of the features, we used Bi-directional Long-Short Term Memory Network in the cloud, a deep learning classifier. We investigated numerous audio features and examined each feature's capability to obtain the most vital details from the audio for it to be labelled genuine or a spoof speech. Furthermore, we use various machine learning algorithms to illustrate the superiority of our system compared to the traditional classifiers. The results of the experiments were classified according to the parameters of accuracy, precision rate, recall, F1-score, and Equal Error Rate (EER). The results were 97%, 100%, 90.19% and 94.84%, and 2.95%, respectively.

Keywords Automatic Speaker Verification (ASV) spoofing voice biometrics deep learning neural network machine learning

Introduction

The voice is considered a form of human biometrics and is a medium of communication. The characteristics of a person's voice is unique, so the voice can used for authentication and biometric identification purposes [1]. Voice biometrics is a simple way of authenticating users that doesn't require any unique sensor device or equipment [2, 3]. A regular smartphone or microphone can be used for this. Voice biometrics are used in the process of verification or recognition of the speaker. Voice biometrics is the technology which uses one-to-one processing to compare the speeches of two individuals. If both speeches originate from the same individual, it is referred to as speaker verification. On the other hand, speaker identification is where an unknown individual

*Correspondence:
Tao Hai
haitao@bjwlxy.edu.cn
Cresantus Biamba
cresantus.biamba@hig.se
[1] School of Computer and Information, Qiannan Normal University for Nationalities, Duyun, Guizhou 558000, China
[2] Key Laboratory of Complex Systems and Intelligent Optimization of Guizhou Province, Duyun, Guizhou, China
[3] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), 81310 UTM Skudai, Johor Bahru, Johor, Malaysia
[4] School of Mathematics and Statistics, Qiannan Normal University for Nationalities, Duyun, Guizhou 558000, People's Republic of China
[5] School of Creative and Cultural Business, Robert Gordon University, Garthdee, Aberdeen AB10 7AQ, UK
[6] Department of Educational Sciences, University of Gavle, Gävle 801 76, Sweden

© The Author(s) 2022, corrected publication 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

is identified with his voice. It is a one-to-many process, although this could result in numerous repetitions of 1-1 comparisons. According to [4], there are two attributes of a person referred to as biometrics. There is the more natural authentication technique such as the iris, face, fingerprint, etc., and the behavioral technique such as signature, voice, gait, etc.

Our primary goal is to create a security layer for the protection of customers from voice replay spoofing attacks to the ASV systems. In this system, the speech signals are used to do a 1-1 comparison of the user's voice and voice prints stored in the database. Naika [5] called this the Automatic Speaker Verification. The ASV system has two main types of attacks— the spoofing and zero effort attack. In the latter attack, a speaker who isn't registered speaks an 'authentic' speech to be granted access as if he was actually registered. This kind of attack is simple to detect because a 1-1 comparison will fail to match with the registered user. In the spoofing attack, the attacker attempts to gain access by playing a speech that was previously recorded, and is similar to the speech of registered speakers. In recent times, the susceptibility of ASV systems to voice spoofing attacks is increasing [6]. The voice spoofing attacks are further divided into logical and physical access attacks. These attacks reduce the efficiency of ASV systems [7]. Building of state-of-the-art ASV systems is an emerging topic. Over the past few years, several authors have organized numerous evaluation challenges [8]. This challenge primarily deals with the logical access attacks, where the samples are created with voice conversion or text-to-speech or algorithms. Challenges have also been organized to address both logical and physical access attacks, and just physical access attacks [9–11]. For greater security of the ASV systems, the spoofing countermeasure should be able to detect the kinds of attacks in the training set, and the system must be capable and robust enough to identify unseen attacks with a reduced EER rate. The anti-spoofing system should be able to generalize. [8] addressed this issue by employing some deep audio features extraction processes. In [12], several embeddings were obtained from an inner layer of the deep neural network in order to indicate the whole audio or the frame of audio signals. The anti-spoofing system tries to ascertain the genuineness of the input speech signal. After the process of extracting the features, a classifier is used to classify the speech into a genuine or fake. Researchers make use of deep learning or ML classifiers, but deep learning classifiers have proven to be more effective; this is also evident in several studies using deep learning approaches [13–16]. The following are the contributions of this research:

1. We propose a state-of-the-art strong voice spoofing countermeasure based on the decomposed signals through the process of empirical mode decomposition for the detection of voice replay attacks.
2. We examine the features of the decomposed acoustic and create a hybrid features-based architecture to detect fake speech.
3. We study the existing ML classifiers for the performance of detecting spoofing.
4. We evaluate the deep learning classifiers for their effectiveness against voice replay attacks.

The remainder of this paper is organized as such: "Literature review" section critically presents related literature. The methodology of the proposed study is discussed in "Proposed method" section. "Results" section discusses the experimental results of the proposed system and compares it with traditional methods. "Conclusion" section concludes this paper.

Literature review

Hanilci et al. [17] presented a method for detecting voice replay spoofing using a high-frequency and glottal excitation band. The information of glottal excitation was extracted using a method of Iterative Adaptive Inverse Filtering which illustrates the unique specifics of fake and genuine speech. They observed a decrease of 3.68% and 8.32% in the Equal Error Rate (EER) in the evaluation and development set. In [18], the authors explored the improved Enhanced Teager Energy Operator (ETEO) and cepstral coefficient, and signal mass to detect replay attacks. The EER of the evaluation and development sets were 10.75% and 5.55%, respectively. The authors in [19] evaluated a spectrum analyzer referred to as 'cochlear' which comprises of a level-dependent compression and a sharp frequency tuning. They then created a method by using an adaptive notch and resonant filter for the cochlear model. This technique showed advancements in the EER by 60.8% and 51.9%. In [20], the authors proposed a framework to detect replayed audio for the security of the ASVs from fraudulent purposes. The framework could also detect the fake audio created by spoofing algorithms.

Aljasem et al. [21] presented a system to detect replay attacks based on the GMM and SVM classifiers, and the analysis of the linear prediction. The evaluation set generated an EER of 4.8%. In [22], the authors proposed a framework for the detection of replay audio and the security of voice assistants like Alexa and Siri based on the difference in the locations of phonemes between a live human's voice and the replay audio. The authors in [23] introduced a technique to detect voice spoofing based on the Gammatone Cepstral Coefficients features and

LTP for the security of the Internet of Things (IoT) and cyber-physical systems. These features were merged and fed into the SVM for differentiating purposes. In [24], the authors proposed a voice replay detection framework by employing the spectral and spatial of signals. They emphasized on the non-speech segments and used spatial features based on the generalized cross correlation to identify the difference. Yaguchi et al. [25] investigated the logSpec and cepstral coefficients to enhance the identification of attacks. The first feature is based on a ratio of the noise and harmonic sub-band. The two features were extracted with the linear prediction signals. The development and evaluation datasets had a reduced EER of 7% and 51.7% respectively. In [26], the authors proposed a spoofing detection technique for replay attacks based on an energy separation algorithm [27, 28]. They also examined a Teager energy operator as a result of it being robust to the noise. They observed EER improvements of 66.34% and 21.88% in noisy and clean environments respectively.

The authors in [29] designed a technique to detect live audio signals based on constant-Q transform which uses distributed frequency bins geometrically. In [30], the authors created spectral based features, i.e., shifted-CQCC and the Glottal Mel-Frequency Cepstral Coefficient (GMFCC), and integrated them for the detection of replay signals. The shifted-CQCC generated an EER of 11.34%, while CQCC produced an EER value of 7.94%. Meng et al. [31] proposed an anti-spoofing measure for smart home systems called ARRAYID which detects the passive liveness that uses the collated speech to distinguish between a live human and the replayed speech. Mittal and Dua [32], explored the deep learning models, CNN and LSTM, and the CQCC spectral feature. Two levels were used to detect spoofs. In the first level, LSTM and CNN were utilized. The next level used time distributed wrappers and LSTM. The authors in [33] presented a system which distinguishes between the genuine and a replayed audio by exploring the manipulations generated by the recording device using an SVM. In [34], the authors proposed a system to detect replay attacks which places prime importance on replaying and recording devices. They also introduced a countermeasure system using the evaluation of a regular audio spoofing tool.

Garg, Bhilare, and Kanhangad [35] introduced an anti-spoofing measure based on MFCC and CQCC feature. Sub-band analysis was performed on these features. The baseline system had an improved EER value of 36.33%. In [36], the authors proposed an anti-spoofing measure for replayed speech based on the human cochlear as compared to the filter bank models. They also designed two features for extracting features from modulation. The experimental results showed that the integration of these features outperformed the filter bank. The authors in [37] presented a linear prediction signal for the detection of replayed speech. The residual-MFCC and excitation source obtained from the linear residual audio signals were integrated for the detection of the replay audio. The residual-MFCC had better outcomes compared to the baseline systems. In [38], the authors evaluated a detection system for spoof speech based on linear frequency residual cepstral coefficient. Two classifiers called the GMM and CNN were used to distinguish between the replayed and genuine audio signals. A decline of 28.78% and 42.72% in the EER values of the development and evaluation sets respectively was reported.

Proposed method

The primary objective of this research is the detection of voice replay attacks against ASV systems. Our proposed system comprises two basic stages: the extraction of features and classification stage. In the first stage, the audio signals are decomposed and two features of 7, 7-dim, i.e., MFCCs and GTCCs are obtained. Next, we used the Bi-directional long short-term memory network as a deep learning classifier. Our proposed spoofing countermeasure determines the user's authenticity based on the voice provided. We used the ASVspoof2019 PA dataset for all tests. Figure 1 illustrates the proposed countermeasure. For the implementation of this work, we used a MATLABR2022a. The Matlab2022a has several tools for audio processing. We also performed extraction of features and classification in a single tool.

Empirical mode decomposition

This process was built in 1998. It deals with the audio signals such that fast oscillations are covered on the slow oscillations which can further be decomposed into meek and intrinsic oscillations in a unique way making use of a dynamic scale without the vital earlier machine specifics [39]. This kind of decomposition is implemented with a restricted single information time scale for it to be suitable for non-linear and non-stationary processes which produces Intrinsic Mode Functions (IMFs) [40]. This has been applied in system detection challenges and health monitoring [41]. Although several applications have proven the legitimacy and robustness of the EMD, it has not been used as a countermeasure to optimize the system and for the purposes of voice spoofing [42]. Before now, the EMD has been studied in two forms: varying the process of sifting and configurations that are empirically stated. It has been used in several applications including

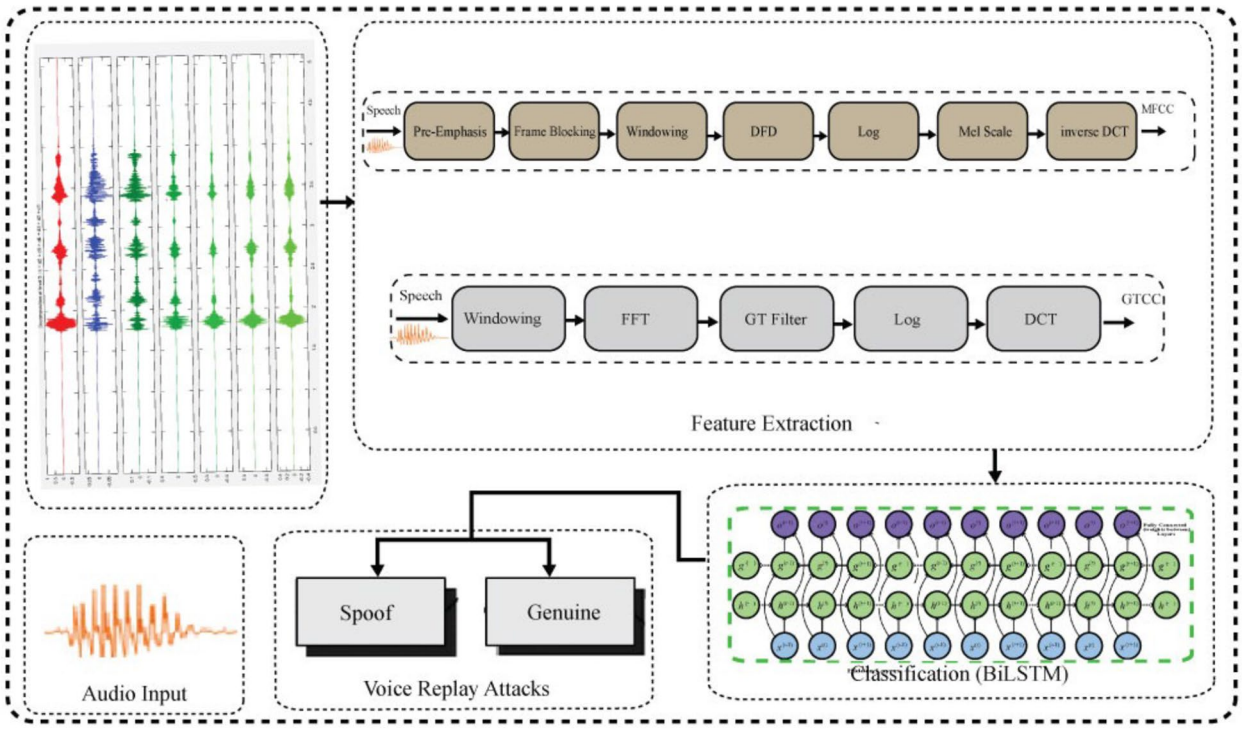

Fig. 1 The proposed system

fault detection, evaluation of biomedical data, and analysis of power and seismic signals. The EMD elements are called IMFs. By using EMD, the noise in signals can be eradicated and the signals reconstructed again. We acquire the significant elements for assessing the audio signals. In this paper, we initially empirically decomposed the signals and obtained the two prinicpal coefficients as presented in Fig. 2.

Methods of extracting features

Mel-Frequency cepstrum coefficient

The MFCC is a popular technique of obtaining features from audio signals [43]. It is referred to as the filter banks-based cepstral domain features obtaining technique. The Mel-scaled filter bank and the Fast Fourier Transform (FFT) is used in the audio signals. The filter bank divided the spectrum non-linearly by adhering to the mel-scale. The lower zones' frequency filters have lower bandwidth than their counterparts. The mel-scale has the spacing of the frequency below 1kHz, in contrast to the logarithmic spacing. The final stage consists of the ranges of coefficients according to their significance. Their importance is obtained through the computation of the discrete cosine transform of the filter bank's logarithmic output. The signals were decomposed, and the 7-dim MFCCs features were extracted from the audio signals. Figure 3 below shows the details.

Gammatone cepstral coefficients

In the next phase, the audio signals were decomposed and the 7-dim features of the GTCC were extracted for more evaluation. It is another technique for obtaining features originally created in [44]. Gammatone's function presents several characteristics that make the GTCC filters suitable to imitate the auditory of the system's spectral and human response [45–47]. Gammatone's function is computed by the multiplication of the Gamma distribution function with the sinusoidal tone. It is illustrated as follows:

$$gt(t) = Kt^{(n-1)}e^{(-2\pi Bt)}cos(2\pi f_c t + \phi) \qquad t \geq 0 \tag{1}$$

The K, B, n, ϕ, and f_c represent the amplitude factor, bandwidth parameter, filter order, phase shift and filter central frequency, respectively. The filter impulse response period is directly connected to the equal rectangular bandwidth, i.e., is a metric used to approximate the bandwidth of human audio filters in the cochlea, a part of the ear. There is a connection between the *ERB* and B. Equation 2 shows the computation of the ERB as:

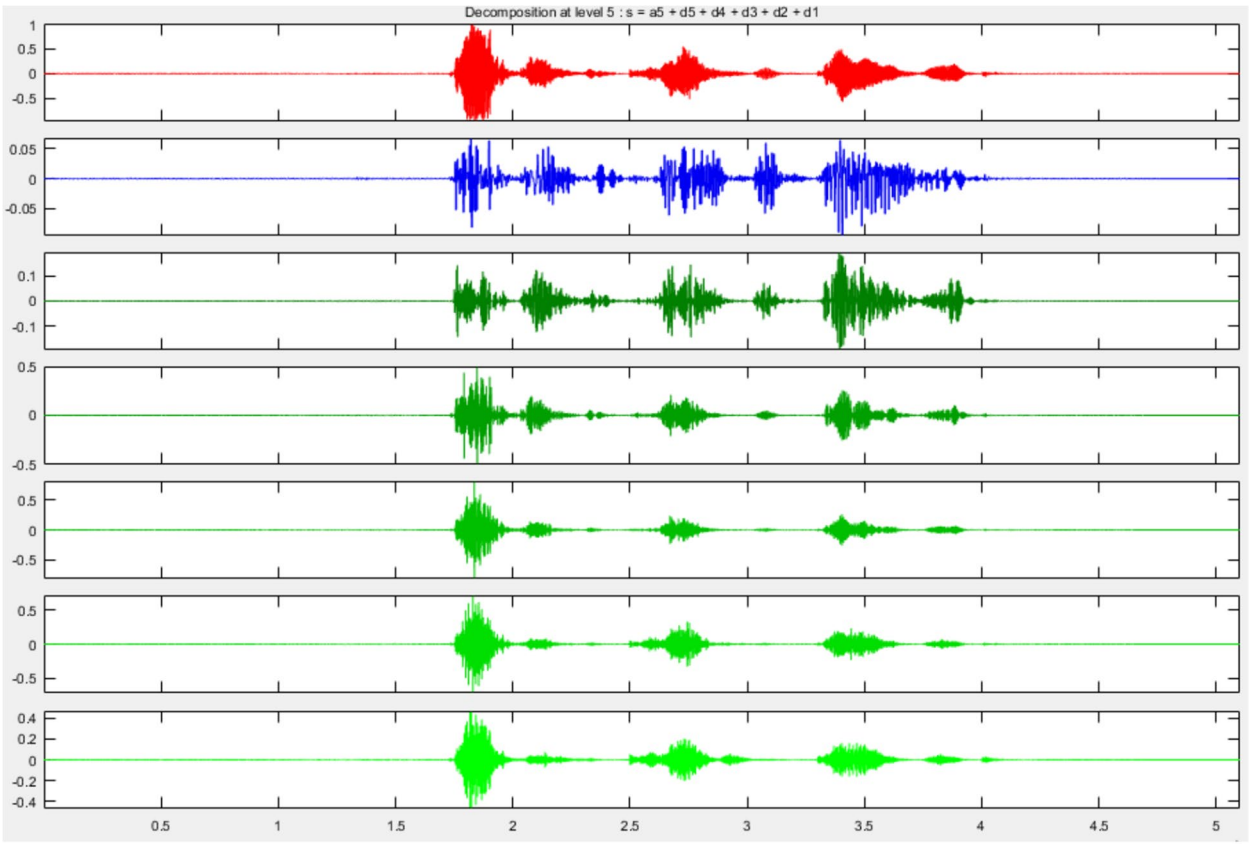

Fig. 2 Decomposed Signals

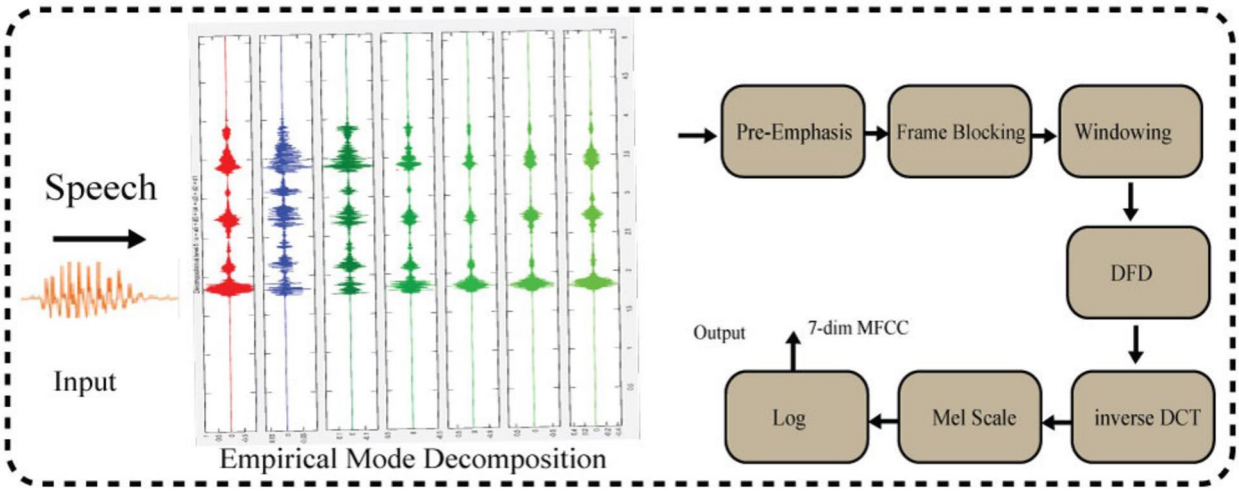

Fig. 3 Extraction of the Decomposed Features from MFCC

$$ERB = [(\frac{fc}{EarQ})^n + minBW^n]^{\frac{1}{n}} \quad (2)$$

The *fc*, *minBW*, *EarQ*, and *n* represent the filter central frequency, lowest bandwidth at zones of lower frequencies, asymptotic quality at higher frequency zones, and

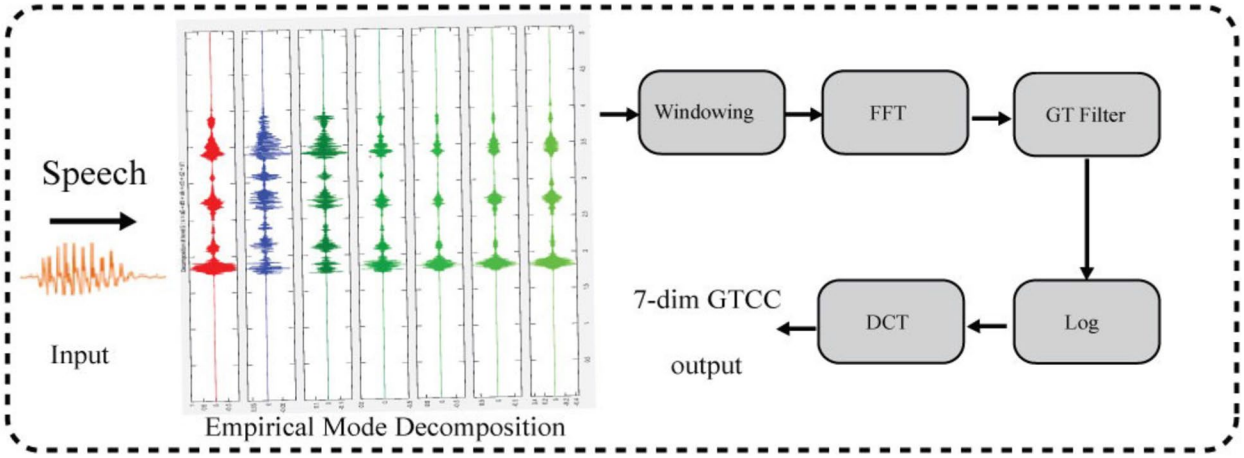

Fig. 4 Decomposed GTCC Features Extraction

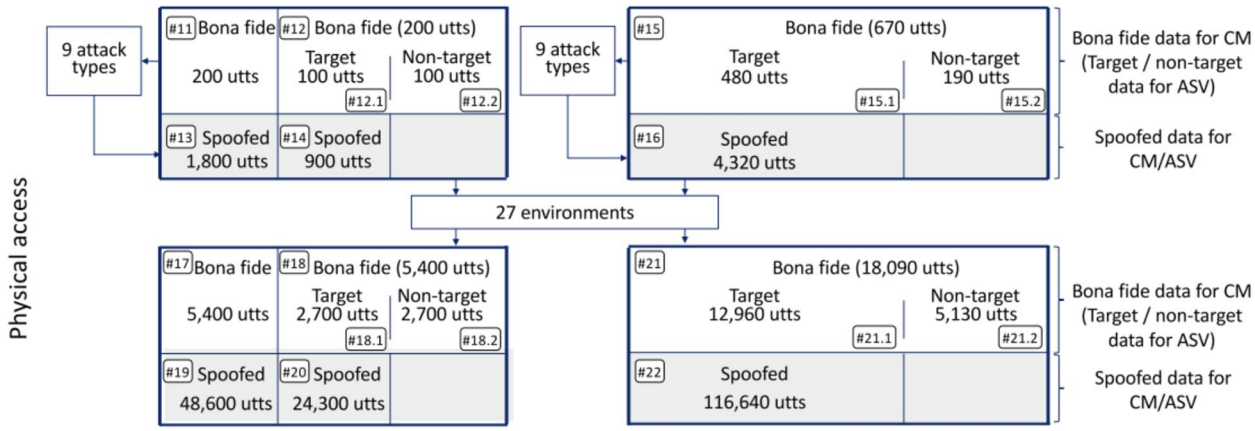

Fig. 5 Statistics of the Dataset

order of approximation, respectively. Equation 3 shows the computation of the middle frequency of every filter *fci* below:

$$f_{ci} = (f_h + EarQminBW)e^{-\frac{i\ step}{EarQ}} - EarQminBW \tag{3}$$

The f_h, $EarQ$, and $minBW$ illustrate the increased frequency and ERB parameters, while i is the GT filter index. The stage is computed by employing Eq. 4 below:

$$Stage = \frac{EarQ}{N}\ln(\frac{f_h + EarQminBW}{f_l + EarQminBW}) \tag{4}$$

In Eq. 3 above, the N illustrates the amount of filters. The GTCC extraction of feature process is similar to that of the MFCCs, but GTCCs use gammatone filter bank instead of a mel-filter bank. 7-dim decomposed GTCCs features were procured from the audio. Figure 4 below gives the details.

Dataset

The ASVspoof2019 [48] PA dataset we used for experimentations is publicly available and the statistics are shown in Fig. 5. The sub-folders are three in number: the training, development, and evaluation folder. They all contain bonafide and replayed speech samples. The genuine data comprises 200 samples collated from 20 various speakers as illustrated in (#11). In a single environment produces voice replay in accordance to nine different attacks, resulting in 1,800 generated samples as illustrated in (#13). A matching method is used for the development partition samples. It is however only for the 10 different speakers as shown in (#12), therefore, 900 samples as depicted in (#14). This process is repeated

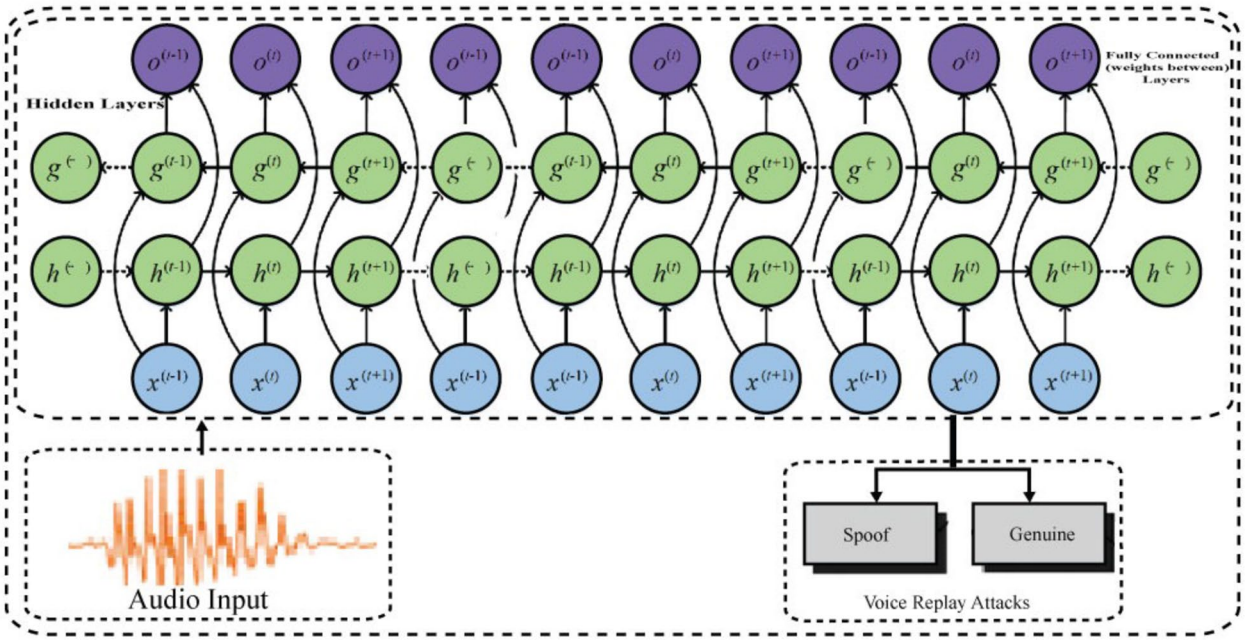

Fig. 6 Classification with input and output

for the entire set of the acoustic surroundings producing the same number of samples as depicted in Fig. 5. There are 24,300 and 48,600 samples for training and development, respectively. An evaluation partition is created in a similar manner, with 48 and 19 target and non-target speakers. This produced 4,320 samples for 9 various voice spoofing assaults in a single environment as illustrated in (#16). For the entire set of 27 varying environments, 116,640 samples are there as illustrated in (#22). The known and unknown attack types are different in the scenario of physical access. If the samples in all the three folders are created with a particular setting of voice replay categories, the impulse responses in every set would be different. In this sense, the evaluation folder's samples are considered unknown attack types.

Classification

Figure 6 illustrates the suggested spoofing countermeasure classification. The audio is processed, and the extracted features are passed into the BiLSTM network to be classified into bonafide or spoofed audio. BiLSTM has continuously been utilized in several approaches [49]. BiLSTM is a Recurrent Neural Network [50] used

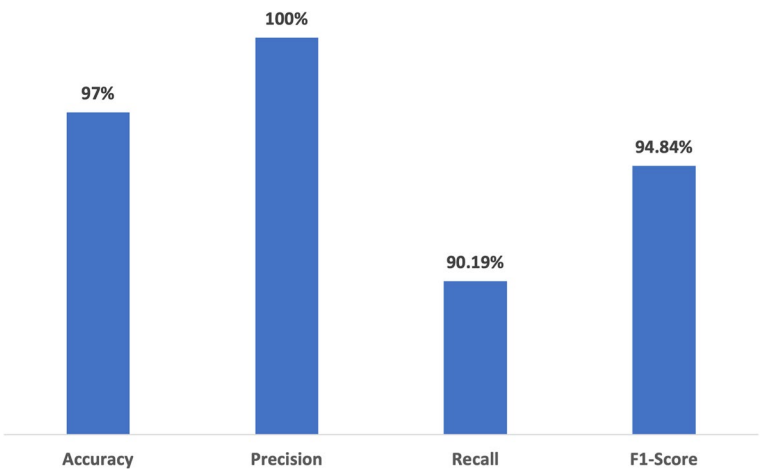

Fig. 7 Our proposed method's results for detecting replay attacks

for Natural Language Processing and the prediction of the time series. The audio signal is also data in the time series. The input moves in a single direction in the LSTM network. In the BiLSTM network, on the other hand, the input flows in both forward and backward directions. This allows the BiLSTM network to use prominent details from both directions. BiLSTM has an additional LSTM layer that varies the movement of details. This means that the input moves in the opposite direction in the additional LSTM layer. Afterwards, the output obtained from the two layers are then merged. Figure 6 illustrates the specifics of the BiLSTM framework below.

Results

This section discusses the comprehensive performance evaluation of the proposed system to detect voice replay attacks. Our technique's performance was evaluated using the Accuracy, Recall, Precision, F1-score, and Equal Error Rate (EER) performance parameters. However, the comparison with other systems will be based on the EER. This experiment was conducted to evaluate our technique (emd-GTCC+emd-MFCC-BiLSTM) using the ASVspoof2019 PA dataset. This dataset has three sets: the training, evaluation, and development set. The training set is used for training, while the evaluation set is used to test the model that has been trained. The samples of the development set cannot be used to evaluate spoofing countermeasures. We empirically decomposed the audio signals and extracted the 7-dim features of the MFCC and GTCC from the evaluation and training set. As far as we know, this is the earliest effort of the signals being decomposed and evaluation of the efficiency of the detectors of spoofing. We used the 14-dim (emd-MFCC and emd-GTCC) features and fed it into the BiLSTM classifier to classify the audio into authentic or spoofed. There are various algorithms which depict improved performance on the classification of the time series data. The audio is a data in the time-series and the proposed BiLSTM framework has shown impressive results. Figure 7

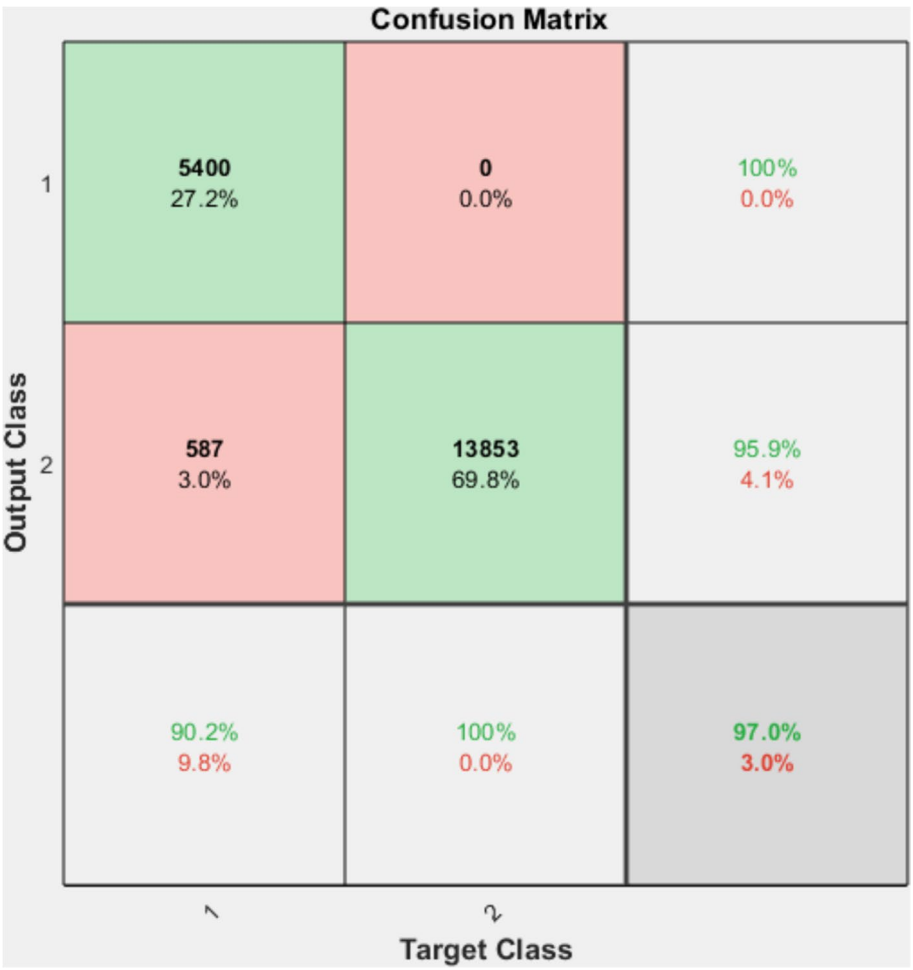

Fig. 8 Voice replay attacks confusion matrix

illustrates the outcome of our spoofing countermeasure. Our proposed method obtained a remarkable accuracy of 97% for binary classification of spoofed and bona-fide audio. The 100% precision rate of our technique signifies that the proposed countermeasure is effective in detecting replay signals. It had recall and F1-score of 94.84% and 90.19%, respectively. The ASVspoof organizers' baseline used Constant Q Cepstral Coefficient (CQCC) and GMM as a form of classifier. Also, the baseline used GMM and Linear Cepstral Coefficients (LFCC) to classify. The resultant systems however are not effective enough to be used in a real-time environment as a result of the features' inability to obtain maximum information. The 2.95% EER value of our method is significantly lower than the baseline methods. The voice replay detection baseline methods obtained an EER of 13.54% and 1q.04% using LFCC-GMM and CQCC-GMM, respectively, in comparison to our system which obtained 10.59% and 8.09%, respectively. The ASVspoof2019 PA dataset contains audio samples recorded making use of several recording gadgets of different qualities: perfect, high or low. The sizes of the room used for the replay attacks

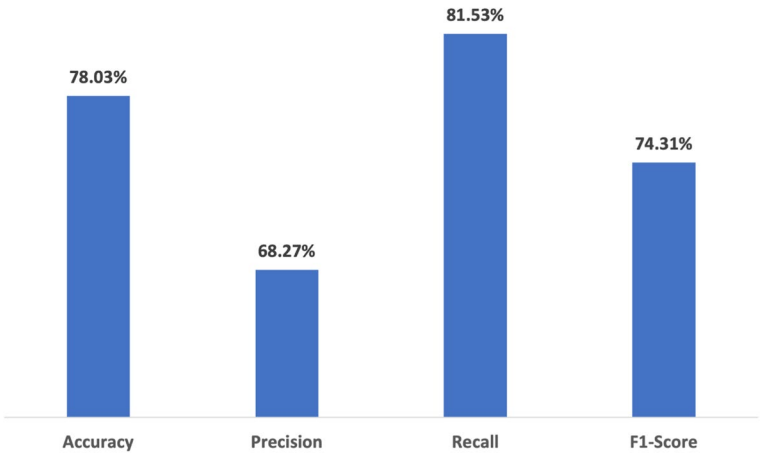

Fig. 9 The SVM results for detecting replay attacks

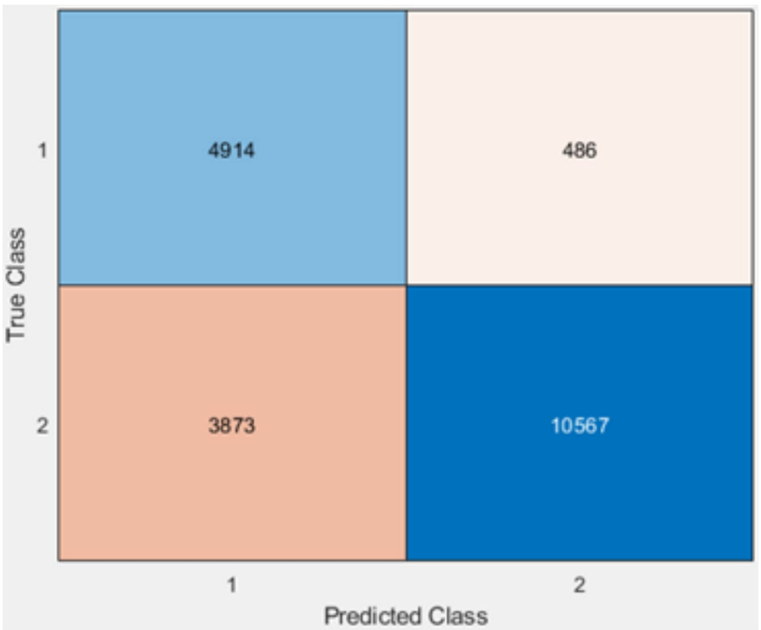

Fig. 10 The SVM's Confusion Matrix for Voice Replay Attacks

recordings are also of different sizes (10-20m, 5-10m and 2-5m). The PA dataset is assorted. The proposed method had an accuracy of 97%, indicating it is effective in detecting voice replay spoofing attacks.

Confusion matrix of the system

This section gives a comprehensive evaluation of the results of classification of our proposed system as depicted in Fig. 8. The confusion matrix was developed for classification challenges, with four kinds of values: True Positive and True Negative (TP and TN), and False Positive and False Negative (FP and FN). The True Positive shows the accurate positive class prediction, and the True Negative shows the accurate negative class prediction. Conversely, the FP shows the inaccurate positive class prediction, and the FN indicates the inaccurate negative class prediction class. As depicted in our system's confusion matrix, the TP, FP, FN, and TN values are 5,400, 13,853, 0 and 597, respectively. These values indicate that the proposed system has accurately classified all the bona-fide samples. The system accurately detected 13,853 samples that were spoofed, and 587 samples that

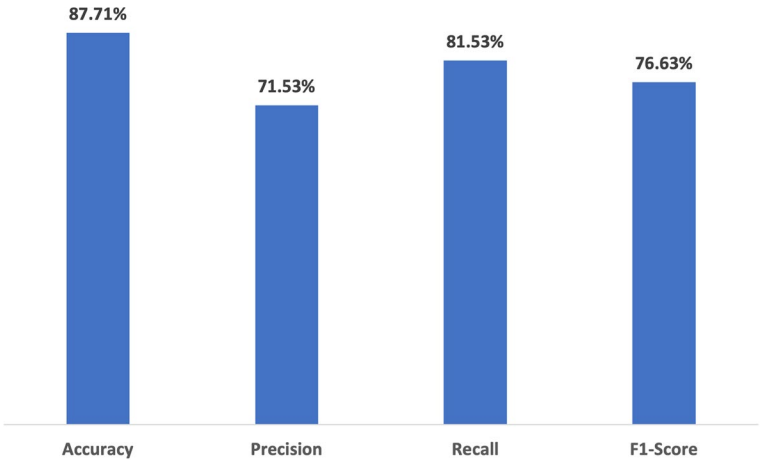

Fig. 11 The ENSEMBLE results for detecting replay attacks

Fig. 12 The ENSEMBLE classifier's confusion matrix for voice replay attacks

were spoofed were detected as bona-fide. 3% of the data are classified incorrectly, the rest are correctly classified. In the confusion matrix, 1 stands for the bona-fide class, and 2 stands for a spoofed class.

SVM performance

The SVM classifier is utilized in various applications. Firstly, 14-dim features were extracted for the training of the SVM classified. The SVM obtained 78.03% accuracy and 68.27% precision. The F1-score and recall attained by the emd-MFCC and emd-GTCC+SVM are 74.31% and 81.53%, respectively. Figure 9 shows the detailed results.

Confusion matrix of the sVM classifier

A confusion matrix was created for the SVM classifier to evaluate the performance in detecting replay attacks. Figure 10 shows the details of the four values. We observed that this method has obtained values of 4,914, 486, 3,873,

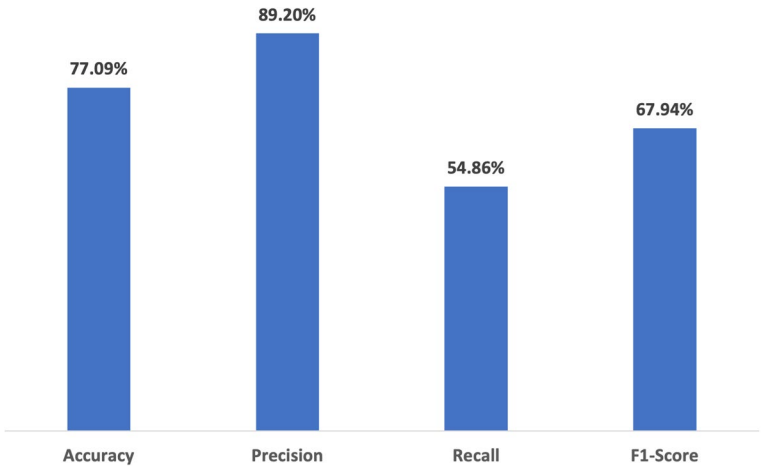

Fig. 13 The KNN results for detecting replay attacks

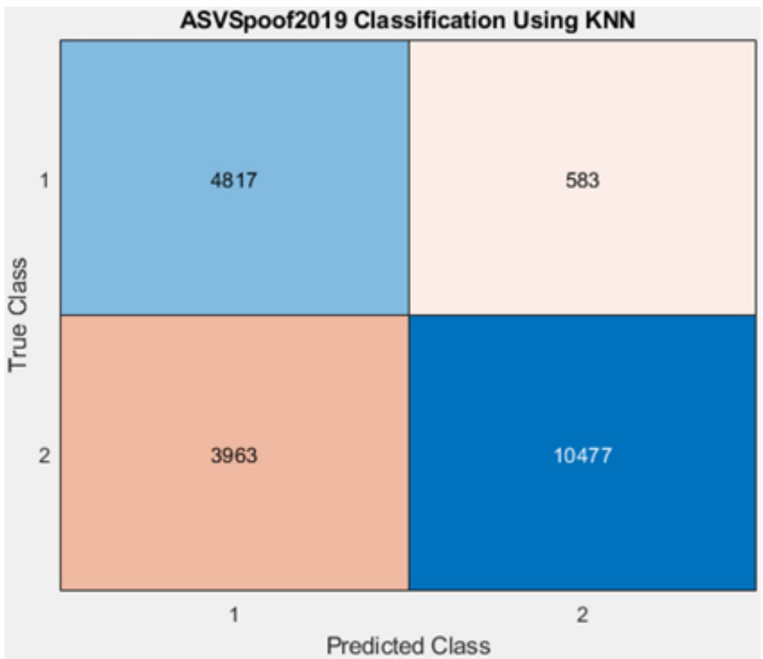

Fig. 14 Confusion Matrix of KNN classifier for voice replay attacks

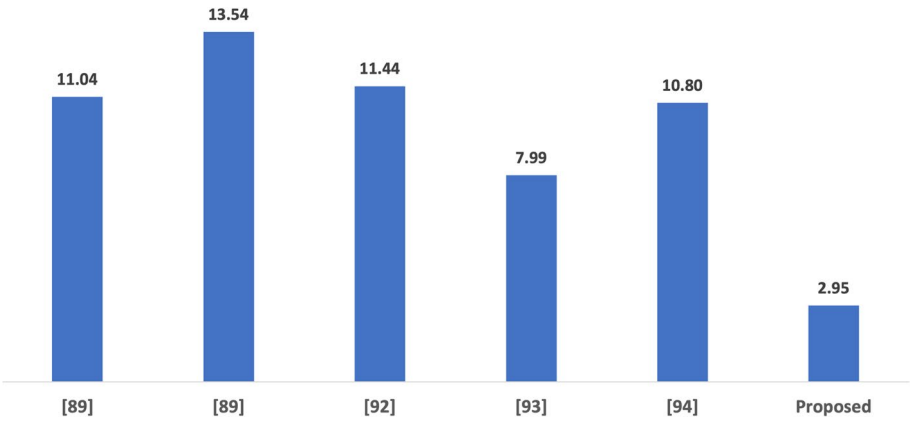

Fig. 15 Comparison with other approaches

and 10,567 of TP, FP, FN, and TN respectively. The values of the FP and FN are optimum, proving that this technique is not effective in detecting replay attacks. 1 stands for the bona-fide class, and 2 depicts the spoofed class.

The ensemble classifier's performance

The second traditional classifier utilized is the ensemble classifier for the detection of replay attacks. This classifier is utilized in several applications. 14-dim features are obtained and passed into the classifier for the classification into spoofed and bona-fide audio. This method obtained an accuracy of 84.71%, 6.68% higher than that of the SVM classifier. The precision rate is 71.53%, higher than the 68.27% of the SVM. The F1-score and recall are 76.63% and 81.53%, respectively. Our technique had the impressive outcomes of precision of 100% and accuracy of 97%, 12.29% higher than that of the ensemble classifier. Figure 11 shows the comprehensive results of the classifier and our technique.

Confusion matrix of ensemble

Figure 12 shows the comprehensive classification performance outcomes of the TP, FP, FN, and TN values. Figure 12 shows that the ensemble classifier accurately classified 3,644 and 13,162 bona-fide and spoofed samples, and 1,756 and 1,278 audio samples are inaccurately classified.

Performance of kNN classifier

The performance of the KNN classifier in detecting voice replay attacks was checked. KNN classifier is utilized in several applications. The obtained 14-dim features are passed in to the KNN for the classification into a bona-fide or a replay voice. Figure 13 shows that the accuracy realized using our proposed method with KNN classifier is 77.09%. The precision rate of 89.2% is 18.49% less than that of an ensemble classifier. The F1-score and recall is 67.94% and 54.86%, respectively. These two parameters on the KNN-based method are smaller than that of the SVM-based technique and the ensemble classifier.

The kNN confusion matrix

A confusion matrix was created for the KNN-based method for the complete classification outcome. Figure 14 illustrates the TP, FP, FN, TN values of 4,817, 583, 3,963, and 10,477, respectively. The KNN-based method has accurately classified 4,817 and 10,477 speech samples, and 583 and 3,963 samples are classified inaccurately.

Performance comparison with existing systems

The performance of our proposed technique is likened to the other existing methods. The comparison is based on the obtained EER value. The most ineffective approach is the baseline with the EER value of 13.54% using LFCC-GMM, while the CQCC-GMM had an EER of 11.04%. The second most effective approach is [49] with an EER value of 7.99%. A Deep Neural Network and CQSPIC method was used. The DNN was used for the classification into authentic or replay speech. In comparison with other methods, our proposed method performed remarkably well with an EER of 2.95%, which is significantly smaller EER value than those of the other techniques. Figure 15 illustrates the comparison between our proposed approach and the others. The comprehensive experimental conclusions and comparison with traditional classifiers show that our proposed approach can encapsulate the unique features from the authentic audio and replay signals.

Conclusion

Attackers use enhanced gadgets to record the voices of bona-fide and registered speakers, replay it to ASV systems to obtain unlawful access for malicious purposes.

These kinds of attacks are serious menaces to the security of these systems. To secure the ASV systems from voice replay spoofing attacks, we proposed a method which uses the empirical mode decomposition of speech signals. GTCC and MFCC are used as features, and the BiLSTM is used to classify the audio into bona-fide or spoofed. The ASVspoof2019 PA dataset is used for the experiments carried out. An accuracy of 97% and precision rate of 100% is achieved by our approach. The F1-score and recall values are 94.84% and 90.19%, respectively. Our proposed approach obtained a significantly lower EER value of 2.95%, and is 8.09% and 10.59% less than the traditional baseline methods. The evaluation and conclusions indicate that our proposed system is reliable for the detection of replay attacks. Subsequently, we aim to explore the efficiency of our proposed approach on the algorithm-generated voice attacks.

Acknowledgments
The authors express their appreciation to the National Natural Science Foundation of China, the Science and Technology Foundation of Guizhou Province, the Top-notch Talent Program of Guizhou province, the program of Qiannan Normal University for Nationalities, the Natural Science Foundation of Education of Guizhou province.

Authors' contributions
Conceptualization by Jincheng Zhou; Methodology by Tao Hai; Software by Dan Wang; formal analysis by Ebuka Ibeke and Dayang Norhayati Abang Jawawi investigation by Tao Hai and Ebuka Ibeke Resources and data collection by Jincheng Zhou, Dan Wang and Dayang Norhayati Abang Jawawi Writing by: Jincheng Zhou and Tao Hai Validation by: Ebuka Ibeke and Cresantus Biamba Funding Acquisition by Jincheng Zhou and Cresantus Biamba. The author(s) read and approved the final manuscript.

Funding
The National Natural Science Foundation of China under Grant No. 61862051; the Science and Technology Foundation of Guizhou Province under Grant Nos. ([2019]1299, ZK[2022]549); the Top-notch Talent Program of Guizhou province under Grant No. KY[2018]080; the program of Qiannan Normal University for Nationalities under Grant Nos. (QNSY2018JS013, QNSYRC201715, QNSY2018003, QNSY2019RC09); the Natural Science Foundation of Education of Guizhou province, China ([2019]203). Open access funding provided by University of Gävle.

Availability of data and materials
The supporting data can be provided on request.

Declarations

Ethics approval and consent to participate
The research has consent for Ethical Approval and Consent to participate.

Consent for publication
The research has consent by all authors and there is no conflict.

Competing interests
There is no competing interest.

Received: 9 July 2022 Accepted: 29 July 2022
Published online: 24 September 2022

References

1 Xu Y, Zeng Q, Wang G, Zhang C, Ren J, Zhang Y (2020) An efficient privacy-enhanced attribute-based access control mechanism. Concurr Comput Pract Experience 32(5):5556
2 Mittal M, Iwendi C (2019) A survey on energy-aware wireless sensor routing protocols. EAI Endorsed Trans Energy Web 6(24). https://eudl.eu/doi/10.4108/eai.11-6-2019.160835
3 Ponnan S, Saravanan AK, Iwendi C, Ibeke E, Srivastava G (2021) An artificial intelligence-based quorum system for the improvement of the lifespan of sensor networks. IEEE Sensors J 21(15):17373–17385.
4 Jain AK, Ross A, Pankanti S (2006) Biometrics: a tool for information security. IEEE Trans Inf Forensic Secur 1(2):125–143.
5 Naika R (2018) An overview of automatic speaker verification system. Intell Comput Inf Commun:603–610. https://link.springer.com/chapter/10.1007/978-981-10-7245-1_59
6 Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2015) Spoofing and countermeasures for speaker verification: A survey. Speech Commun 66:130–153.
7 Korshunov P, Marcel S (2016) Cross-database evaluation of audio-based spoofing detection systems In: Interspeech. https://infoscience.epfl.ch/record/219837?ln=en
8 Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilçi C, Sahidullah M, Sizov A (2015) Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge In: Sixteenth Annual Conference of the International Speech Communication Association. https://www.eurecom.fr/publication/4573
9 Korshunov P, Marcel S, Muckenhirn H, Gonçalves AR, Mello AS, Violato RV, Simoes FO, Neto MU, de Assis Angeloni M, Stuchi JA, *et al* (2016) Overview of btas 2016 speaker anti-spoofing competition In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 1–6. IEEE. https://ieeexplore.ieee.org/abstract/document/7791200?casa_token=W9RbLt8WBD0AAAAA:b7UL3xnAGjtfvUxtocPZXg4YdSkVaPE4Ezy6KQsAuBYRiFlPVlLN4d6pubtUml1Q9ifpqYjKBgk
10 Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, Lee KA (2017) The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. https://www.isca-speech.org/archive/interspeech_2017/kinnunen17_interspeech.html
11 Palanivinayagam A, Gopal SS, Bhattacharya S, Anumbe N, Ibeke E, Biamba C (2021) An optimized machine learning and big data approach to crime detection. Wirel Commun Mob Comput 2021. https://www.hindawi.com/journals/wcmc/2021/5291528/
12 Kinnunen T, Delgado H, Evans N, Lee KA, Vestman V, Nautsch A, Todisco M, Wang X, Sahidullah M, Yamagishi J, *et al* (2020) Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. IEEE/ACM Trans Audio Speech Lang Process 28:2195–2210.
13 Mittal M, Saraswat LK, Iwendi C, Anajemba JH (2019) A neuro-fuzzy approach for intrusion detection in energy efficient sensor routing In: 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 1–5.. IEEE.
14 Latif SA, Wen FBX, Iwendi C, Li-li FW, Mohsin SM, Han Z, Band SS (2022) Ai-empowered, blockchain and sdn integrated security architecture for IoT network of cyber physical systems. Comput Commun 181:274–283.
15 Iwendi C, Srivastava G, Khan S, Maddikunta PKR (2020) Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems:1–14. https://link.springer.com/article/10.1007/s00530-020-00701-5
16 Iwendi C, Maddikunta PKR, Gadekallu TR, Lakshmanna K, Bashir AK, Piran MJ (2021) A metaheuristic optimization approach for energy efficiency in the IoT networks. Softw Pract Experience 51(12):2558–2571.
17 Hanilci C, Kinnunen T, Sahidullah M, Sizov A (2016) Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. Speech Comm 85:83–97.
18 Bharath K, Kumar MR (2022) New replay attack detection using iterative adaptive inverse filtering and high frequency band. Expert Syst Appl 195:116597.
19 Patil AT, Acharya R, Patil HA, Guido RC (2022) Improving the potential of enhanced teager energy cepstral coefficients (etecc) for replay attack detection. Comput Speech Lang 72:101281.

20 Gunendradasan T, Ambikairajah E, Epps J, Sethu V, Li H (2021) An adaptive transmission line cochlear model based front-end for replay attack detection. Speech Comm 132:114–122.
21 Aljasem M, Irtaza A, Malik H, Saba N, Javed A, Malik KM, Meharmohammadi M (2021) Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging. IEEE Trans Inf Forensic Secur 16:3524–3537.
22 Nasersharif B, Yazdani M (2021) Evolutionary fusion of classifiers trained on linear prediction based features for replay attack detection. Expert Syst 38(3):12670.
23 Yue L, Cao C, Li Y, Li J, Liu Q (2021) Liveear: An efficient and easy-to-use liveness detection system for voice assistants In: Journal of Physics: Conference Series, vol. 1871, 012046. IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-6596/1871/1/012046/meta
24 Javed A, Malik KM, Irtaza A, Malik H (2021) Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. Appl Acoust 183:108283.
25 Yaguchi R, Shiota S, Ono N, Kiya H (2021) Replay attack detection based on spatial and spectral features of stereo signal. J Inf Process 29:275–282.
26 Wei L, Long Y, Wei H, Li Y (2022) New acoustic features for synthetic and replay spoofing attack detection. Symmetry 14(2):274.
27 Xu Y, Yan X, Wu Y, Hu Y, Liang W, Zhang J (2021) Hierarchical bidirectional rnn for safety-enhanced b5g heterogeneous networks. IEEE Trans Netw Sci Eng 8(4):2946–2957.
28 Xu Y, Liu Z, Zhang C, Ren J, Zhang Y, Shen X (2021) Blockchain-based trustworthy energy dispatching approach for high renewable energy penetrated power systems. IEEE Internet Things J. https://ieeexplore.ieee.org/document/9560154
29 Prajapati GP, Kamble MR, Patil HA (2021) Energy separation based features for replay spoof detection for voice assistant In: 2020 28th European Signal Processing Conference (EUSIPCO), 386–390. IEEE. https://ieeexplore.ieee.org/abstract/document/9287577?casa_token=GZiV_1nQIJ8AAAAA:UYPT7IwwAXHErozDrXJERnHsCg63Ke43hc-btmjYAeEmTeU0ZTeJ8Rq2a73VXF4sknn0JnDg1K0
30 Dutta K, Singh M, Pati D (2021) Detection of replay signals using excitation source and shifted cqcc features. Int J Speech Technol 24(2):497–507.
31 Meng Y, Li J, Pillari M, Deopujari A, Brennan L, Shamsie H, Zhu H, Tian Y (2022) Your microphone array retains your identity: A robust voice liveness detection system for smart speaker In: USENIX Security. https://www.usenix.org/conference/usenixsecurity22/presentation/meng
32 Mittal A, Dua M (2022) Static–dynamic features and hybrid deep learning models based spoof detection system for asv. Compl Intell Syst 8(2):1153–1166.
33 Ren Y, Fang Z, Liu D, Chen C (2019) Replay attack detection based on distortion by loudspeaker for voice authentication. Multimed Tools Appl 78(7):8383–8396.
34 Yoon S-H, Koh M-S, Park J-H, Yu H-J (2020) A new replay attack against automatic speaker verification systems. IEEE Access 8:36080–36088.
35 Garg S, Bhilare S, Kanhangad V (2019) Subband analysis for performance improvement of replay attack detection in speaker verification systems In: 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), 1–7. IEEE. https://ieeexplore.ieee.org/abstract/document/8778535?casa_token=swFCpmqf1sgAAAAA:IMxyoJwsGipHVxdSa2_skF3CyDpsEhI74jQtQrGYwtVwAJKZuwQ1Ih_m9YeJOxZJz6urNsR97Q8
36 Gunendradasan T, Irtza S, Ambikairajah E, Epps J (2019) Transmission line cochlear model based am-fm features for replay attack detection In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6136–6140. IEEE. https://ieeexplore.ieee.org/abstract/document/8682771?casa_token=xwlzDD2oWzEAAAAA:5AuG-q43ii2y_mz5VGTn8TlSf1eMcXK0srIwfU1vX5ZE43wGDzzwcUHG2LWwATPZr7tNs4_F4G8
37 Singh M, Pati D (2019) Usefulness of linear prediction residual for replay attack detection. AEU-Int J Electron Commun 110:152837.
38 Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen N-C, Tung CC, Liu HH1998. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. https://www.jstor.org/stable/53161
39 Rilling G, Flandrin P, Goncalves P, *et al* (2003) On empirical mode decomposition and its algorithms In: IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, vol. 3, 8–11. IEEER Grado. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.586.812&rep=rep1&type=pdf
40 Lee YS, Tsakirtzis S, Vakakis AF, Bergman LA, McFarland DM (2009) Physics-based foundation for empirical mode decomposition. AIAA J 47(12):2938–2963.
41 Ricci R, Pennacchi P (2011) Diagnostics of gear faults based on emd and automatic selection of intrinsic mode functions. Mech Syst Signal Process 25(3):821–838.
42 Li C, Wang X, Tao Z, Wang Q, Du S (2011) Extraction of time varying information from noisy signals: An approach based on the empirical mode decomposition. Mech Syst Signal Process 25(3):812–820.
43 Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Sig Process 28(4):357–366.
44 Patterson RD, Holdsworth J, *et al* (1996) A functional model of neural activity patterns and auditory images. Adv Speech Hear Lang Process 3(Part B):547–563.
45 Xu Y, Ren J, Zhang Y, Zhang C, Shen B, Zhang Y (2019) Blockchain empowered arbitrable data auditing scheme for network storage as a service. IEEE Trans Serv Comput 13(2):289–300.
46 Xu Y, Zhang C, Zeng Q, Wang G, Ren J, Zhang Y (2020) Blockchain-enabled accountability mechanism against information leakage in vertical industry services. IEEE Trans Netw Sci Eng 8(2):1202–1213.
47 Xu Y, Zhang C, Wang G, Qin Z, Zeng Q (2020) A blockchain-enabled deduplicatable data auditing mechanism for network storage services. IEEE Trans Emerg Top Comput 9(3):1421–1432.
48 Yamagishi J, Todisco M, Sahidullah M, Delgado H, Wang X, Evans N, Kinnunen T, Lee KA, Vestman V, Nautsch A (2019) Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. https://ieeexplore.ieee.org/document/9358099
49 Das RK, Yang J, Li H (2020) Assessing the scope of generalized countermeasures for anti-spoofing In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6589–6593. IEEE. https://ieeexplore.ieee.org/abstract/document/9053086/?casa_token=t_M6aLgSkwoAAAAA:7m52qVwU913gZOV79c_GPeXg3BjG8DXmK0R-cfYo_1cPpM1zcg6HEop-gcqK8_olpwWsBA0p-Rw
50 Kumar RL, Khan F, Din S, Band SS, Mosavi A, Ibeke E (2021) Recurrent neural network and reinforcement learning model for covid-19 prediction. Front Public Health 9. https://www.frontiersin.org/articles/10.3389/fpubh.2021.744100/full

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com