iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI: http://doi.org/10.22214/ijraset.2020.4055

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

An Overview on Automatic Speaker Verification System Techniques

Ajila A[1], Smitha K S[2]
[1]*PG Scholar, Dept. of ECE, LBS Institute of Technology for Women, Kerala, India*
[2]*Assistant Professor, Dept. of ECE, LBS Institute of Technology for Women, Kerala, India*

Abstract: ***Automatic Speaker Verification (ASV) is a widely used voice-based biometric system. Like every other biometric system, ASV is also vulnerable to malicious attacks. Replay attacks in ASV systems are a greater threat compared to other attacks. Many types of research have been conducted to find the countermeasures for the replay attacks. Feature extraction and classification are the two fundamental processes of speaker verification. A wide variety of features and classifiers were implemented in the past studies to check their performances. The feature extraction methods discussed in this paper are Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Discrete Wavelet Transform (DWT) and Perceptual Linear Prediction (PLP). Some of the classifiers that are discussed are Gaussian Mixture Model (GMM), Gaussian Mixture Model-Universal Background Model (GMM-UBM), Hidden Markov Model (HMM), Support Vector Machine (SVM), Dynamic Time Wrapping (DTW) and deep learning networks like Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). This paper intends to focus on the various attacks in the ASV system along with these two fundamentals of the system.***

Keywords: Speech Processing, Automatic Speaker Verification (ASV), Spoof Detection, Feature Extraction, Classifiers.

I. INTRODUCTION

Biometric identification is a form of identity-based on an individual's behavioral or biological characteristics. These characteristics are classified into two: anatomical and behavioral. Anatomical characteristics include identification based on face, palm, fingerprint, signature, etc. Behavioral characteristics include keystroke dynamics, gait analysis, etc. [1]. Voice biometrics can be included either as an anatomical characteristic or as a behavioral characteristic. Some of the biometric systems fail to secure the data and also fails in operation when it comes to practical applications. Over the past decades, researches being conducted to rectify the problems that are faced by almost all biometrics.

Voice biometrics is a high demanded research area for over the last decade. Speaker recognition and speaker verification are those popular areas. Speaker recognition is basically the identification of a person from the characteristics of his/her voices. On the other hand, speaker verification is verifying the claimed person's identity based on the input speech samples. The speaker recognition systems are further classified into Speaker Identification (SID) system and Automatic Speaker Verification (ASV) system. The SID system tries to identify the unknown speaker's identity by using the known identities and the ASV system verifies the claimed person's identity based on the pre-recorded speech samples from the speaker.

In practical case scenarios, the ASV system is prone to get attacked in 9 positions as shown in Figure 1 [2]. Point 1 is the microphone point and point 2 is the transmission point. These two points come under direct attack regions. Other points come under the indirect attack regions and for these types of attack regions the attacker should need access to the ASV system. Point 3 is the override feature extraction point, point 4 is modify probe to feature, point 5 denotes the modify speaker database, point 6 is modify biometric reference, point 7 is override classifier, point 8 is modify score and finally point 9 is override decision.

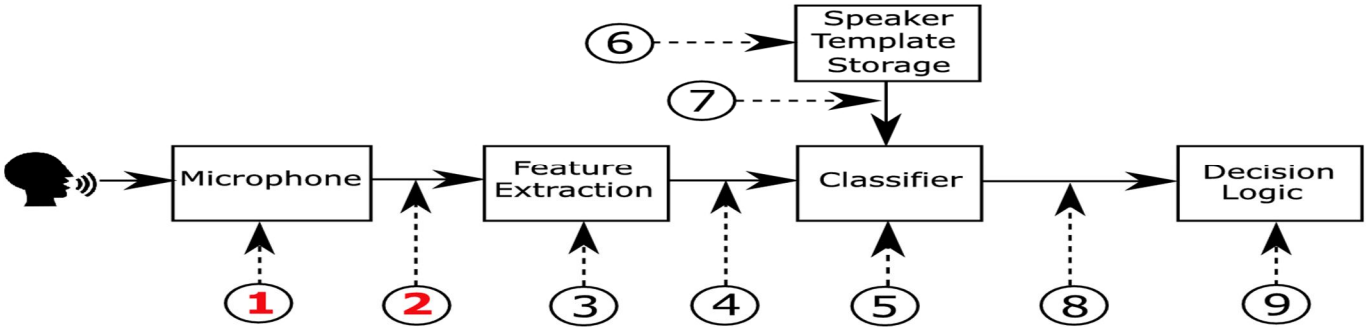


Figure 1. Possible attacks in an ASV system

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

II. ASV SYSTEM: SPOOFING ATTACKS

There are mainly five types of spoofing attacks in an ASV system. They are Impersonation [3], Voice Conversion (VC) [4], Speech Synthesis (SS) [5], Twins [6] and Replay [7] as shown in Figure 2.

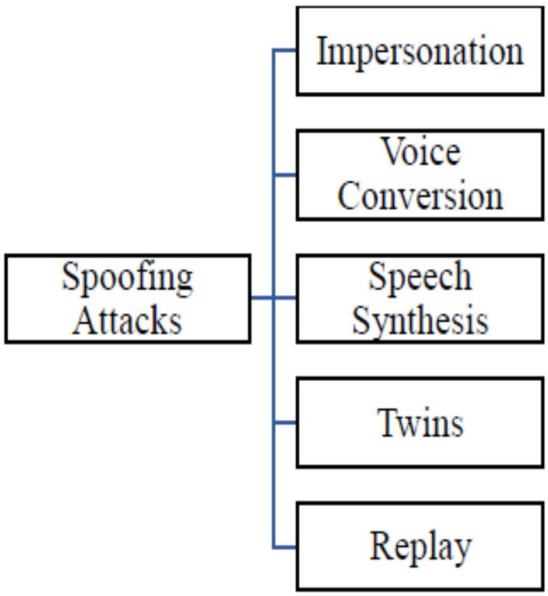


Figure 2. Various spoofing attacks

A. Impersonation

Impersonation is one of the most obvious spoofing techniques where it is known for human-altered voices, also known as mimicking. Here, the attacker tries to mimic the speaking of the target speaker. For these types of attacks, the attacker does not need any technical equipment or any computer knowledge for mimicking the target speaker. For better mimicking purposes, the attacker tries to mimics the prosodic features of the target speaker's voice. Professional mimicking tries to imitate the prosody features along with the other features like accent, pronunciation and other high-level speaker traits.

B. Voice Conversion (VC)

Voice conversion is a process of converting the source speaker's voice similar to the target speaker. The VC needs some kind of equipment for the conversion process. VC mainly deals with the segmental and suprasegmental features of the source speaker also try to keep the language content similar. The main approaches used in VC are spectral mapping and prosody conversion. Spectral mapping normally deals with frequency wrapping, statistical parametric and unit selection. While the prosody conversion plays an important role in characterizing the speaker's individuality.

C. Speech Synthesis (SS)

Speech Synthesis is also known as Text-to-Speech (TTS) where the input text is converted to the speech signal. These speech signals are fed into the ASV system for the spoofing attack. SS is normally giving machine-generated signals. SS has commonly used spoofing because it uses computerized techniques and the output from the SS is normally high-quality voice signals. SS uses the properties of the claimed person's voice characteristics and spectral cues of natural speech. Since the voice generated from the SS is high-quality, it is obvious that these signals have more energy compared to the natural speech signals.

D. Twins

These types of attacks are not very common since these attacks are only applicable when it comes to twins. The speech data is sufficient to identify a human being, however, when it comes to twins, it can be relatively difficult to identify the same. When it comes to identical twins, the system is likely to fail. Identical twins have the same spectrographic patterns, but they can be identified with correct ASV systems. The pattern of the speech signal, pitch, contours, and spectrograms are similar when it comes to twins, if not identical. So, the target speech signal and source speech signal can be identified when it comes to foolproof in the ASV system.

E. Replay

Replay attacks are the most commonly used attacks since it does not need any equipment or computer knowledge. A replay is a form of spoof attack in the ASV system using a pre-recorded speech sample collected from a genuine target speaker. Replay attacks are the widely used attack since it does not need any special knowledge in speech processing. The replay spoofed sound signal can have some characteristic change because of the changes in recording devices and environmental changes. So, a replay attack is the most difficult attack to distinguish and identify and has the highest possibility to spoof the system.

The replay attack scenario in the ASV system is shown in Figure 3 below [8]. In case of a genuine attempt, the target speaker directly gives his/her voice samples to the system and gets accessed. This case is known as the actual speech. While in case of a spoofed attack, the target speaker's voice is recorded first by the intruder by his/her playback devices before giving in to the ASV system. This process is called the replay spoof speech. The difference between the actual speech and replay spoof speech is the use of recording and playback devices and also there will be a characteristic change in the voice samples since the use of external devices.

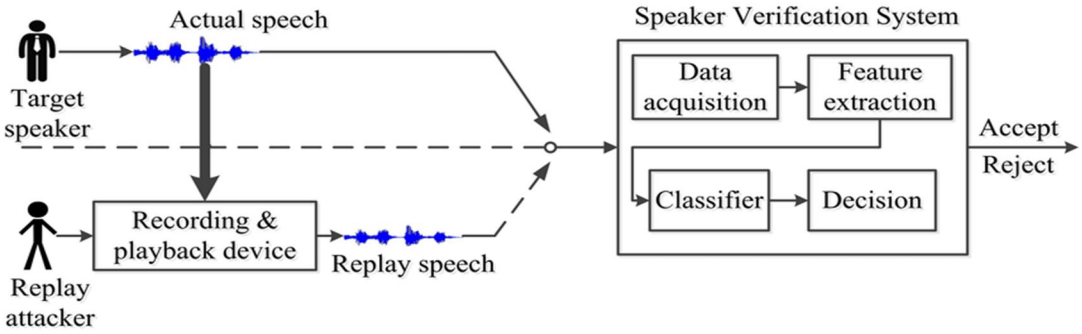


Figure 3. Replay attack scenario in ASV system

III. REPLAY SPOOF DETECTION ARCHITECTURE

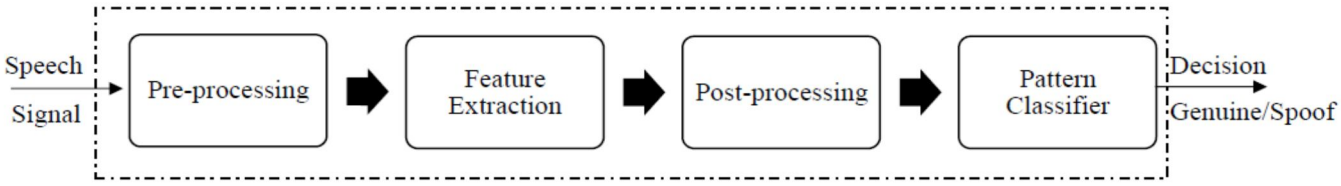


Figure 4: System Architecture

A replay spoof detection consists of four major steps [9]. These are Pre-processing, Feature extraction, Post-processing and classifier. These steps are explained below.

A. Pre-processing

Pre-processing is considered as the first phase in any ASV system. Pre-processing mainly modifies the input speech signal so that the feature extraction and analysis of the speech will be easy in further steps. Pre-processing is done in input signal mainly to avoid any noise or background disturbances. The main pre-processing steps are pre-emphasis, silence removal, etc.

B. Feature Extraction

Feature extraction is the process of transforming raw signals into some type of parametric representation. The input signal is normally a random process and cannot be extracted as such for classification. So, the role of feature extraction is to map the input raw signal into a new vector space which is a more understandable representation of the input signal. The most commonly used feature extraction methods are Constant Q-Cepstral Coefficient (CQCC), Mel-Frequency Cepstral Coefficients (MFCC), etc.

C. Post-processing

Post-processing is a necessary step used for correcting systematic inaccuracies if any in the system. This technique is mainly used for extracting the features of a new vector space. Such that in this space the data will be more distinguishable. Post-processing composes of feature normalization, dimensionality reduction, etc.

D. *Pattern Classifier*

The pattern classifier as the name suggests is verified by using various patterns. After effective feature extraction, the main task is to classify whether the input signal is genuine or spoof. By using patterns of the speech signal, one can identify the signal is spoofed or genuine. Various classifiers used are Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Convolutional Neural Network (CNN), etc.

IV. FEATURE EXTRACTION TECHNIQUES

This section describes some of the commonly used feature extraction techniques in speech spoofing.

A. *Mel-Frequency Cepstral Coefficient (MFCC)*

This is one of the most commonly used feature extraction methods. MFCC computation is used as a replication of the human hearing system. It is intended to artificially implement the human ear's working principle [10]. MFCC features are based on the recognized discrepancy of the human ear's critical bandwidth with frequency filters spaced linearly at low frequencies. MFCC is based on the signal decomposition with the help of filter banks. The fundamental steps of an MFCC algorithm are shown in the below figure, Figure 5.

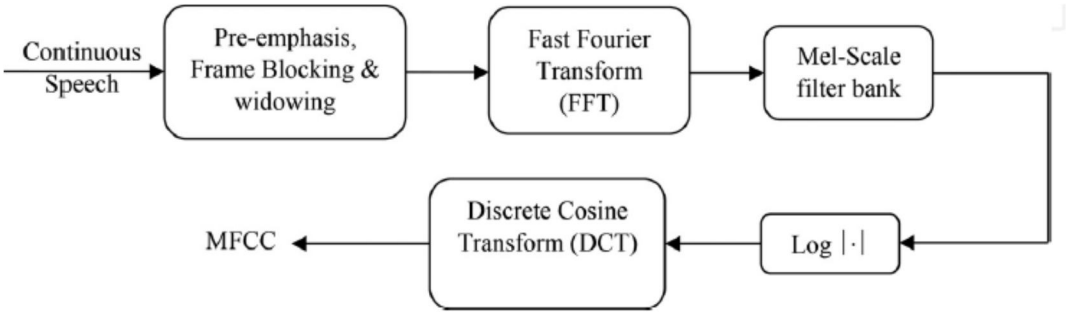


Figure 5: Block diagram of MFCC

Cepstral coefficients are usually accurate when it comes to voice-related applications. It is widely used for both speaker verification and speaker recognition. When background noise is present MFCC will not be able to give accurate output and may not be ideal for normalization [11].

B. *Linear Prediction Coefficients (LPC)*

LPC basically imitates the human vocal tract. It provides accurate speech features. The evaluation of LPC is done by approximating the formants, estimating the concentration and frequency of the residue left behind. It is a powerful speech analysis method and is known for its formant estimation method [12]. It is in encoding high-quality speech at a low bit rate. LPC is also used for speech reconstruction. The block diagram of LPC is shown in Figure 6.

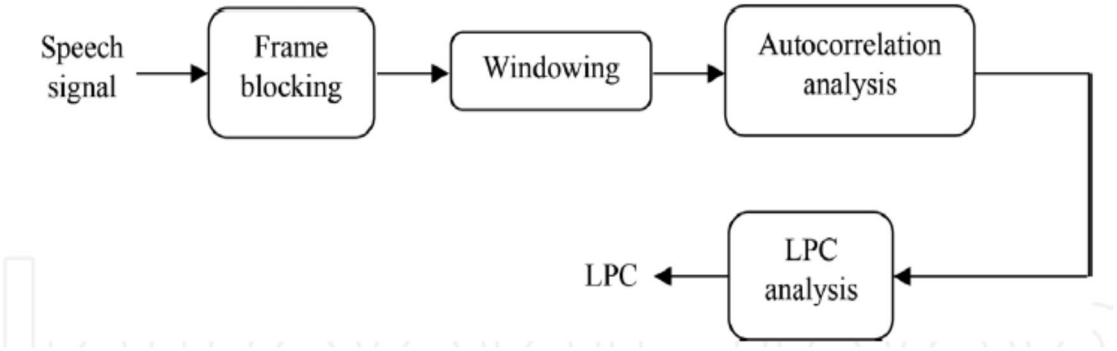


Figure 6: Block diagram of LPC

LPC selects the vocal tract information from the given input speech. LPC represents consistent source behaviors. The speech parameters are more accurately calculated using LPC when compared to other feature extraction methods. The main lack of LPC is the aliased autocorrelation coefficients.

C. Linear Prediction Cepstral Coefficients (LPCC)

LPCC is the spectral coefficient derived from the LPC spectral envelope. LPCC is the Fourier transform illustration of the logarithmic magnitude spectrum of LPC [13]. The cepstral analysis is commonly used in speech processing because it has the ability to symbolizing speech waveforms and characteristics with a limited size of features. LPCC is more often used because the use of limited correlated features will give more accurate output.

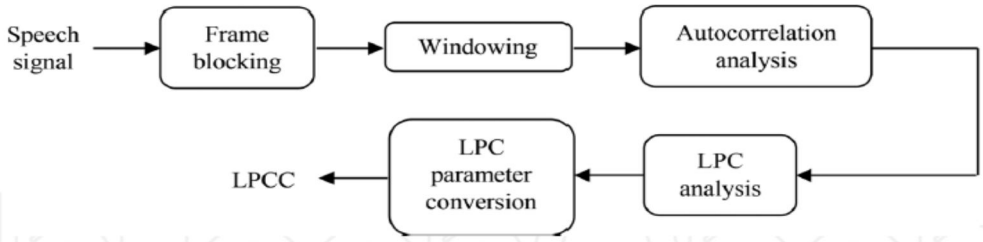


Figure 7: Block diagram of LPCC

LPCC has a lower error rate when compared to LPC. Cepstral analysis on the high pitch speech signal gives small source-filter separability in the frequency domain. These cepstral features are highly sensitive to noise.

D. Discrete Wavelet Transform (DWT)

DWT is an extension of the Wavelet Theory (WT) that enhances the flexibility of the decomposition process [14]. It is a highly flexible and efficient method for sub-band breakdowns of speech signals. It is specially designed for evaluating a finite set of samples over the set of scales. The DWT is shown in Figure 8.

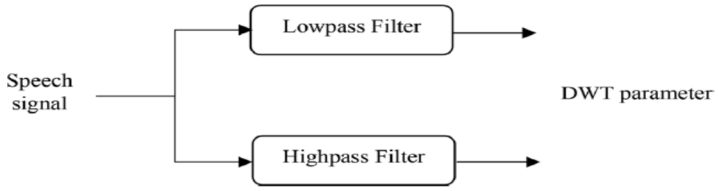


Figure 8: Block diagram of DWT

DWT has two related functions called scaling function and wavelet function [15]. These functions help in computing the DWT. It provides information regarding different frequency scales. These enhance the information regarding the speech signal since it provides accurate information about the frequencies it belongs to. It provides enough number of frequency bands for speech analysis.

E. Perceptual Linear Prediction (PLP)

PLP combines various techniques like intensity-to-loudness compression, critical bands, and equal loudness pre-emphasis to extract the relevant information from the speech signal. PLP acts similarly to MFCC because it too does imitate the process of the human hearing method. It provides minimum resolution in high frequency. It is a combination of both spectral analysis and linear prediction analysis. The speech spectrum resembles an auto-regressive all-pole model [16]. These auto-regressive coefficients can be converted to cepstral variables. It is noise resistant to an extent and is sensible to additional noise in the system. The filter shape used in PLP is a trapezoidal filter.

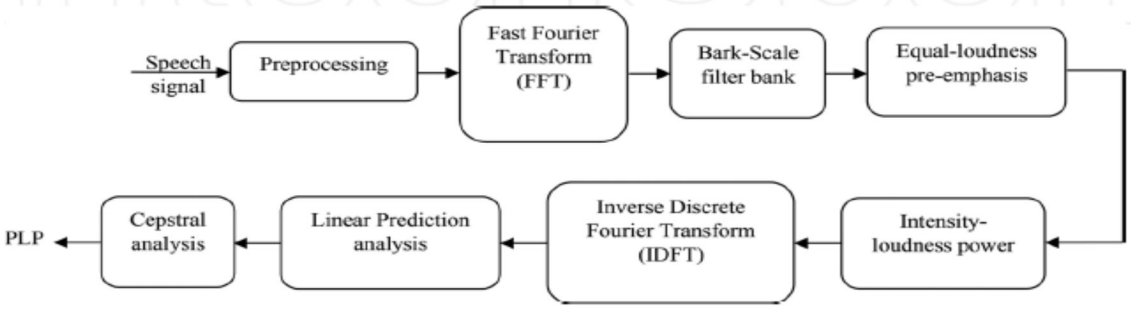


Figure 9: Block diagram of PLP

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

V. CLASSIFIERS

After the effective extraction of the features, the speech vectors should be used in a way to obtain its maximum results. The extracted features are used to train a classifier which helps to classify according to the words present in the speech input. This section describes some of the commonly used classifiers in speech spoofing.

A. Gaussian Mixture Model (GMM)

GMM is the widely used classifier in speech spoof detection and it also yields accurate output. It is a parametric probability density function that is normally represented as a weighted sum of Gaussian component densities [17]. Each class is represented as a weighted sum of M multivariate Gaussians,

$$p(x/\lambda) = \sum_{i=1}^{M} w_i p_i(x)$$

where w_i is the i^{th} mixture weight and $p_i(x)$ is the D-variate Gaussian density function. It has a mean μ_i and a covariance matrix Σ_i. Also, the model parameters are denoted by,

$$\lambda = \{\omega_i, P_i, \Sigma_i\}_{i=1}^{M}$$

B. Gaussian Mixture Model-Universal Background Model (GMM-UBM)

GMM-UBM is based on modeling a certain amount of data from all the classes pooled to develop a UBM and this model is adapted to each class [17]. The main disadvantage that GMM-UBM faced was its session variability in testing and training conditions. GMM-UBM can be used to calculate high-dimensional supervectors.

C. Hidden Markov Mixture (HMM)

HMM comes under a more speculative approach. It is a computationally practical and easy approach when it comes to speech spoofing. It is characterized as a finite state Markov model and sets as of output distributions [18]. The main limitation of HMM is it is quite complex to examine the errors found in the output. So, HMM is not frequently used for spoofing techniques when it comes to enhancing the performance of the system.

D. Support Vector Machine (SVM)

SVM is an effective classifier when it comes to speech spoofing. It is a binary classifier and is basically constructed from Kernel functions. It constructs a linear model based on support vectors in order to estimate decision function [17]. SVM is used to classify data sets. SVM is basically the transformed training patterns and is equally close to the hyperplane of separation.

E. Dynamic Time Wrapping (DTW)

DTW is basically an algorithm that is used for calculating the similarity between two series that may differ in time. Any data say speech, video, etc. can be transformed into linear representation and analysis with DTW. The disadvantage of DTW is it lacks continuity when compared to other classifier approaches.

F. Convolutional Neural Network (CNN)

A typical CNN consists of a convolutional network and a fully connected layer. The convolution module consists of two layers a convolutional layer and a pooling layer [19]. The convolutional layer performs the convolution operation to generate the output from the local regions of the input features of the previous layer. The pooling layer performs the down-sampling of the feature map of the previous layer and generates new feature maps with a reduced resolution. In almost all cases, max-pooling layers are used on CNN.

G. Recurrent Neural Network (RNN)

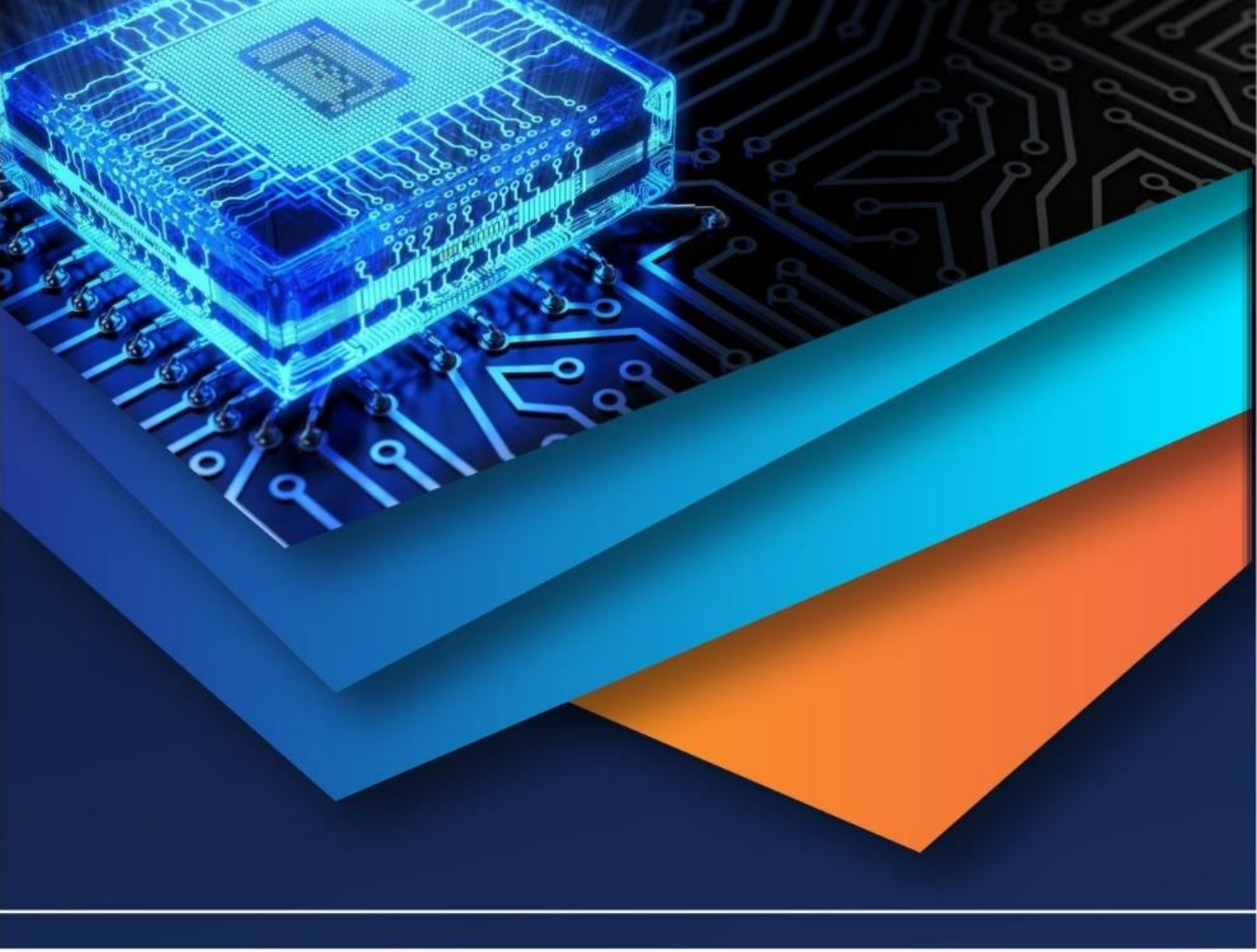
RNN is a feed-forward deep model like CNN. RNN basically does not consider the current input but also takes into consideration what they perceived during the previous time [19]. That is, RNN has two sources of inputs the present input and the recent past inputs. RNNs are basically distinguished from feedforward networks by a feedback loop connected from the past decisions made. RNN can be used when there is an arbitrarily long sequence of information.

VI. CONCLUSION

This work aims for an overview of the automatic speaker verification system. The spoofing in the voice biometric system is a greater threat to the identification system technology. Many types of research have been done to find the countermeasures for the spoofing attacks. This paper provides a review of commonly used methods in some researches. Feature extraction and classification are the key elements in identifying the spoofing. Some of the feature extraction and classifier methods are discussed in this work. Also, the researches show that introducing effective feature extraction methods and classifiers can improve spoof detection in speech processing.

REFERENCES

[1] Togneri, Roberto and Daniel Pullella, "An overview of speaker identification: Accuracy and robustness issues," in IEEE circuits and systems magazine 11, no.2, 2011, pp.23-61.
[2] Sahidullah M., Delgado H., Todisco M., Kinnunen T., Evans N., Yamagishi J. and Lee K. A., "Introduction to voice presentation attack detection and recent advances," in Handbook of Biometric Anti-Spoofing, Springer, Cham, 2019, pp. 321-361.
[3] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, China, 2004, pp. 145–148.
[4] Y. Stylianou, "Voice transformation: A survey," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 2009, pp. 3585–3588.
[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.
[6] H. A. Patil and K. K. Parhi, "Variable length Teager energy-based Mel cepstral features for identification of twins," in International Conference on Pattern Recognition and Machine Intelligence, New Delhi, India, 2009, pp. 525–530.
[7] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in IEEE Annual Summit and Conference in Asia-Pacific Signal and Information Processing Association (APSIPA-ASC), Chiang Mai, Thailand, 2014, pp. 1–5.
[8] Singh Madhusudan and Debadatta Pati. "Usefulness of linear prediction residual for replay attack detection", in AEU-International Journal of Electronics and Communications, 2019, Volume 110: pp.152837.
[9] Wu Z., Evans N., Kinnunen T., Yamagishi J., Algre F. and Li H., "Spoofing and countermeasures for speaker verification," in Speech Communication, 66, 2015, pp. 130-153.
[10] Chakroborty S., Roy A., Saha G., "Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification," in IEEE International Conference on Industrial Technology, 2006. ICIT 2006. pp. 387-390.
[11] Chu S., Narayanan S., Kuo C. C., "Environmental sound recognition using MP-based features," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE; 2008. pp. 1-4.
[12] Gill A. S., "A review on feature extraction techniques for speech processing," in International Journal of Engineering and Computer Science. 2016, 5(10): pp.18551-18556.
[13] Wu Q. Z., Jou IC, Lee S. Y., "On-line signature verification using LPC cepstrum and neural networks," in IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 27(1): 1997, pp.148-153.
[14] Turner C, Joseph A., "A wavelet packet and Mel-frequency cepstral coefficients-based feature extraction method for speaker identification," in Procedia Computer Science, 2015, pp. 416-421.
[15] Nehe N. S., Holambe R. S., "DWT and LPC based feature extraction methods for isolated word recognition," in EURASIP Journal on Audio, Speech, and Music Processing. 2012, 2012(1):7.
[16] Hermansky H., "Perceptual linear predictive (PLP) analysis of speech," in The Journal of the Acoustical Society of America, 87(4), 1990, pp.1738-1752.
[17] Hanilçi, Cemal, Tomi Kinnunen, Md Sahidullah, and Aleksandr Sizov. "Classifiers for synthetic speech detection: A comparison," 2015.
[18] Singh A., Panchal T., and Saharan M., "Review on Automatic Speaker Recognition System," in International Journal of Advanced Research in Computer Science and Software Engineering, 3(2), 2013, pp.350-354.
[19] Qian, Yanmin, Nanxin Chen, Heinrich Dinkel, and Zhizheng Wu. "Deep feature engineering for noise robust spoofing detection." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25, no. 10, 2017, pp. 1942-1955.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY