Spoofing Detection and Countermeasure in Automatic Speaker Verification System using Dynamic Features

Medikonda Neelima, I. Santi Prabha

Abstract*: This present paper aims to extract robust dynamic features used to spoofing detection and countermeasure in ASV system. ASV is a biometric person authentication system. Researchers are aiming to develop spoofing detection and countermeasure techniques to protect this system against different spoofing attacks. For this, replayed attack is considered, because of very common accessibility of recording devices. In replay spoofing, the speech utterances of target (genuine) speakers are recorded and played against ASV system for gaining access unauthorizedly. For this purpose, as a first step, different dynamic features will be extracted for each speech sample. For feature extraction MFCC, LFCC, and MGDCC feature extraction techniques are used. As a second step, a classifier is used to classify whether the given speech sample is genuine or not. As a classifier, GMM and universal background model is used. In this present work, GMM based ASV system and Countermeasure systems using different feature extraction techniques are developed, and the performance of the methods is evaluated using EER and t- DCF. Basing on the performance values, the best feature extraction technique is selected.

Keywords* : *Automatic Speaker Verification (ASV) system, Equal Error Rate (EER), False Acceptance Rate (FAR), False Rejection Rate (FRR).

I. INTRODUCTION

Speech is a common mode of communication. ASV system has become a cost-effective and reliable approach over recent years for person recognition. ASV system is used to verify the person's identity using speech utterance. There are many applications of ASV system, such as access control in smart phones, phone banking, credit card activation, trading, password resetting, etc. Recent advancements in biometric technology have significantly improved ASV performance in different applications.

There are four attacks to spoof a biometric authentication system. They are replay attack, voice conversion attack, speech synthesis attack, and impersonation attacks. A replay attack is made by pre-recording samples taken from a genuine speaker. Voice conversion technique is to change given speaker's sample into the required speaker's sample so that converted sample resembles genuine speaker's sample. Speech synthesis is also treated as a spoofing attack where a natural-sounding artificial speech is generated. Impersonation is a very common spoofing attack where attacker tries to mimic the speaker's voice without any knowledge about the system. These four spoofing attacks confuse the ASV system for deciding whether to accept or reject a particular speaker.

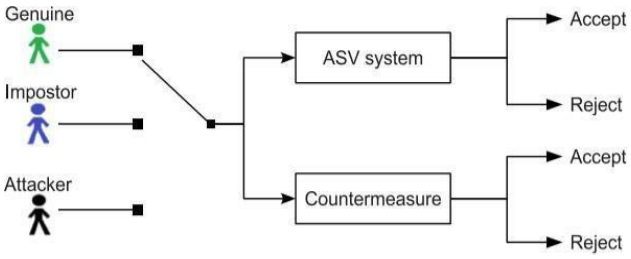

Fig. 1.Illustration of the decision made by an ASV system and countermeasure

To prevent spoofing attacks, a countermeasure is developed such that a decision is made whether the access attempt is from a genuine speaker or a spoofing attack. Countermeasure has to be developed, which decreases FAR while not increasing FRR. From Fig. 1, it can be shown that an ASV system can make decisions depending on whether the identity claim is genuine or impostor/attacker [1]. For a genuine sample, the system accepts it, and for impostor/attacker sample, the system rejects it. In a similar way, the countermeasure can make decisions depending on whether the identity claim is a human sample or a spoofed sample. For a human sample, the system accepts it, and for a spoofed sample, the system rejects it.

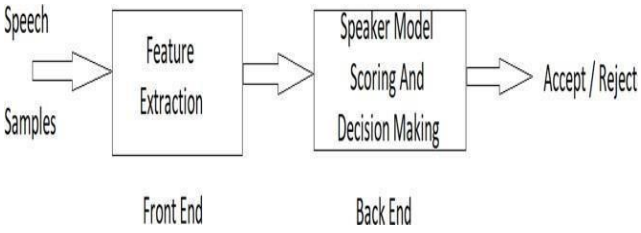

Fig. 2.Proposed ASV and countermeasure system

The proposed ASV and countermeasure system is shown in fig. 2.

Front End: First block is a feature extraction block. This block is a front end which is used to extract different features used for spoofing detection and countermeasure [2].

Revised Manuscript Received on January 15, 2020.

* Correspondence Author

Medikonda Neelima*, Ph.D. Scholar, E.C.E. Department, JNTUK, Kakinada, Andhra Pradesh, India. Email:mneelima@gvpce.ac.in.

I. Santi Prabha, Professor, E.C.E. Department, JNTUK, Kakinada, Andhra Pradesh, India. Email: santiprabha@yahoo.com

In this, the dynamic features are extracted using MFCC, MGDCC, and LFCC techniques.

Back End: Second block is a speaker model used for scoring the system and for decision making. As a speaker model in this present paper, Gaussian Mixture Model (GMM) is used. The GMM-UBM back end is shown in fig 6. The countermeasure is also developed using the same system.

II. FEATURE EXTRACTION FRONT END

Dynamic features are extracted from sample speech utterances. Previous studies reported that MFCC, LFCC, and MGDCC work better than other features [3].

A. Mel-Frequency Cepstral Coefficients (MFCC)

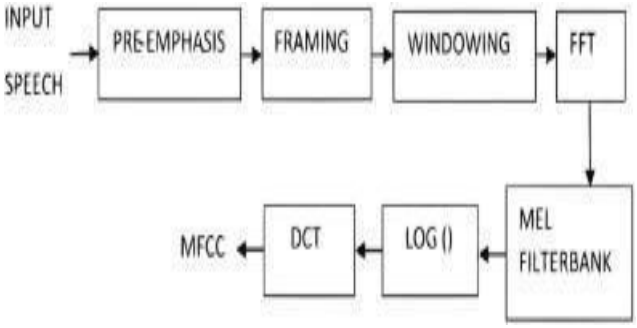

Fig. 3.Block representation of MFCC extraction

MFCC is a prominent well-known popular feature for speaker verification. MFCC is the commonly used feature in any ASV system, which is used as a primary feature extraction technique [1]. It is similar to the human audio perception system. The steps involved in extracting the features are shown in fig 3. The first step is pre-processing. In this, a speech sample is applied to a filter for boosting the energy content in a higher frequency component. Speech sample is non-stationary in nature; hence it is split into 20 to 30 msec shorter frames to maintain the signal almost stationary. While framing the speech samples, to protect from not losing the spectral content, the frames are passed through a Hamming window function. To convert speech sample to spectral domain, a frequency transform such as Discrete Fourier Transform (DFT) is used. Resulting spectrum is passed through a series of 26 triangular shaped filters, usually known as a mel filter bank. These mel filters are commonly spaced equally and model mel scale as given in following expression

$$mel(f) = 2592 * log\left(1 + \frac{f}{700}\right) \quad (1)$$

This is an equation used to change frequency (f) from Hertz to mel. Cepstrum of time sequence s(n) and is calculated from the inverse transform of the logarithm of the spectrum. The Mel- cepstrum is standard parameterization used in speaker verification and is referred to as MFCC.

B. Linear Frequency Cepstral Coefficients (LFCC)

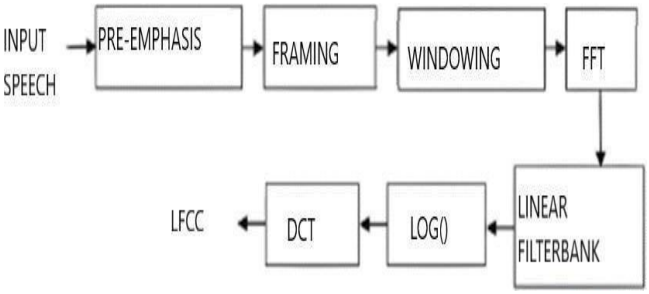

Fig. 4.Block representation of LFCC extraction

Processing steps for extracting LFCC coefficients is shown in fig 4. The difference between MFCC and LFCC extraction is, in LFCC linear filter bank is used instead of mel filter bank.

C. Modified Group Delay Cepstral Coefficients (MGDCC)

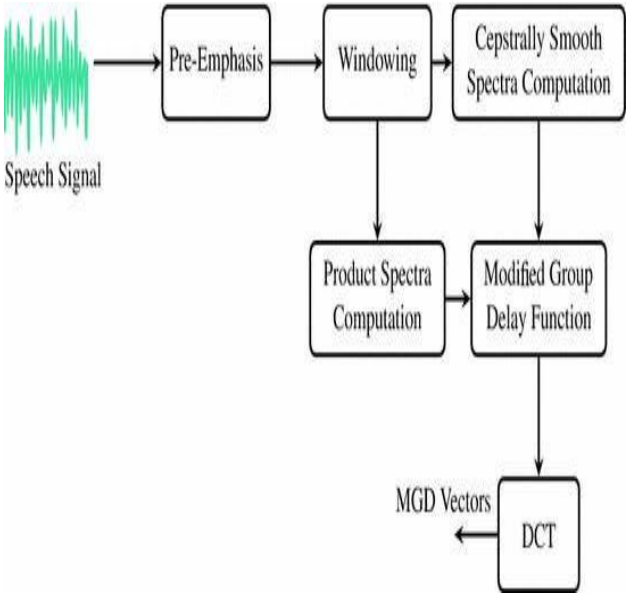

Fig. 5.Block representation of MGDCC extraction

The steps involved in extracting MGDCC features are shown in fig 5 [5]. Group delay from the Fourier transform is obtained by calculating the derivative of phase and negating the resultant. Group delay improves the most necessary features from the speech spectrum.

Previous work on spoofing detection using group delay states that group delay with some modification is better than initially obtained group delay [5]. The extraction of cepstral coefficients from the modified group delay function, DCT is applied as follows

$$c_i = \sum_{m=0}^{M} g(m) cos\left(\frac{\pi i}{M}(2m+1)\right) \quad (2)$$

Where M is the order of DCT and g(m) is the group delay function.

III. SPEAKER MODEL BACKEND

Back-end processing systems have been successfully used for spoofing detection. Such an approach is a GMM [7]. GMM is a popular classifier, and it is mostly used by speech and speaker recognition. After features are extracted, the GMM is built separately for genuine and spoof. The mixture components number used in GMM is 512. The log-likelihood based score is computed using the following formula

$$score(S) = llk(S \mid \lambda_{genuine}) - llk(S \mid \lambda_{spoof}) \quad (3)$$

Where $S=\{s_1,s_2,\ldots\ldots s_T\}$ is the feature vector of test utterance, T is the sampling period and models represented as $\lambda_{genuine}$ and λ_{spoof} respectively.

The block diagram is shown in fig. 6 is a GMM-LLR (GMM-Log-Likelihood Ratio) based spoofing detection score measurement system

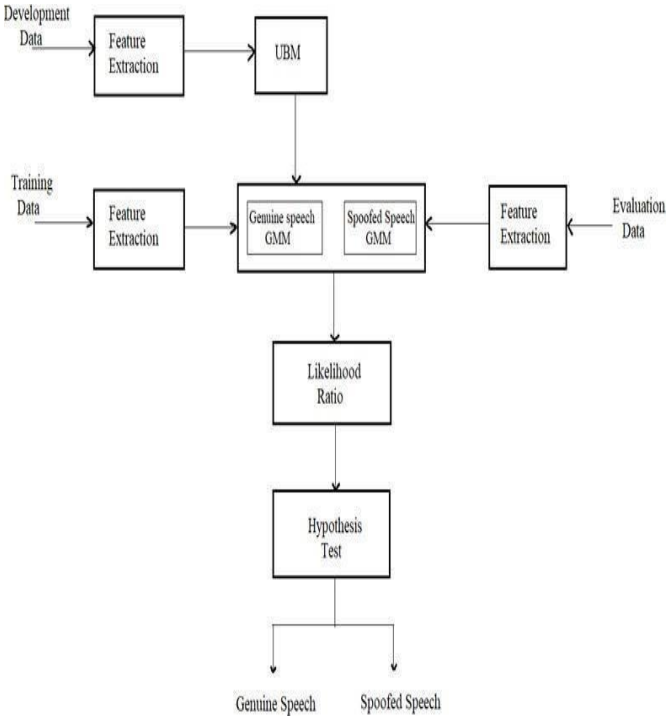

Fig. 6.GMM-UBM backend system

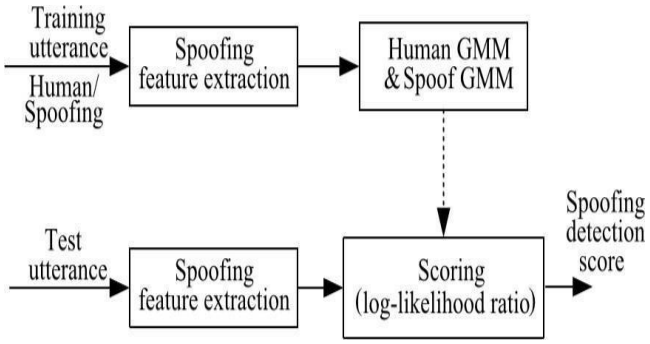

Fig. 7.Block representation of a GMM-LLR scoring system

Score generated by system depicted in Fig. 7 is used to estimate the FRR and FAR values finally from which the EER is estimated.

IV. EXPERIMENTAL SETUP

A. Database

The ASVspoof 2017 database is developed by the RedDots database [6]. The genuine utterances for this work are taken from database subset. Replayed data is recorded with the help of volunteers with replay and playback devices. The entire database is divided into training, development, and evaluation sets. The detailed number is shown in table I.

Table I: Description of ASV Spoof 2017 database

Subset	#Speakers	#Non replay	#Replay
Training	10	1508	1508
Development	8	760	950
Evaluation	24	1298	1200 8

B. Evaluation metric

Evaluation of the proposed ASV and countermeasure systems is carried out with different tests by providing genuine and spoofed samples separately. ASV system decides to accept or reject claimed identity and results in any one of four outcomes.

Speech utterance	Decision	
	Accept	Reject
Genuine	Correct acceptance	False rejection
Spoofed	False acceptance	Correct rejection

In these results, two are correct outputs, and two are incorrect outputs. Incorrect outputs are false acceptance and false rejection. To evaluate the system performance, FAR and FRR are to be estimated. The FAR and FRR are having an inverse relationship that one value may be reduced with increasing the other. The system parameters and coefficients are selected such that which brings out the compromise between FAR and FRR. The compromising solution is measured in terms of EER. In addition to these measures, Detection Cost Function (DCF) is also a considerable metric [1]. It supports evaluation with a combination of ASV and spoofing countermeasures.

tandem Detection Cost Function (t-DCF) proposed in [4] is a performance measure which is used in evaluating system in taking decisions from combined ASV and Countermeasure system. This measure is basing on comparing detection error values of different systems, i.e., ASV and CM systems, and actions taken on a number of trials.

V. RESULTS AND DISCUSSION

ASV system and CM system with different features and the corresponding percentage EER is given in table II and table III, respectively. Three different types of features are observed. Those are statistical (stat), delta features (delta), and double delta features (double delta). From table II, it is observed that EER(%) on development and evaluation data is very high on ASV system with both MFCC and MGDCC features and GMM classifier.

From table III, it is observed that EER(%) on development and evaluation data is comparatively low on CM system with LFCC features and GMM classifier for all statistical, stat+delta, and stat+delta+double delta features. From table III, it is also observed that EER(%) on development data is 6.19 with MFCC+GMM as ASV system and LFCC+GMM as CM system with statistical + delta features.

Table II: EER(%) of spoofing detection with ASV system

ASV Feature set + Model	Dev EER (%)	Eval EER (%)
MFCC+GMM	39.07	40.30
MGDCC+GMM	30.93	40.76

Table III: EER(%) of spoofing detection with CM system

Features	CM Feature set + Model	Dev EER (%)	Eval EER (%)
Stat	LFCC+GMM	6.32	37.60
Stat + delta	**LFCC+GMM**	**6.19**	35.59
Stat + delta + double delta	LFCC+GMM	7.89	**33.58**

Table IV: t-DCF values of ASV and CM systems on development and evaluation data

Features	ASV Feature set + Model	CM Feature set + Model	Dev t-DCF	Eval t-DCF
Stat	**MFCC+GMM**	**LFCC+GMM**	**0.137**	0.743
	MGDCC+GMM	LFCC+GMM	0.165	0.739
Stat + delta	**MFCC+GMM**	**LFCC+GMM**	**0.138**	0.712
	MGDCC+GMM	LFCC+GMM	0.172	**0.709**
Stat + delta + double delta	MFCC+GMM	LFCC+GMM	0.171	0.716
	MGDCC+GMM	LFCC+GMM	0.197	0.710

t- DCF values on ASV and CM for different features are noted in table IV. It is observed that better t- DCF is obtained with statistical features for MFCC+GMM as ASV system and LFCC+GMM as CM system.

VI. CONCLUSION

In this paper, the robust features are extracted and are used for spoofing detection and countermeasure in automatic speaker verification system. For feature extraction, MFCC, LFCC, and MGDCC techniques are adopted. Statistical, delta, and double delta features are extracted with LFCC technique and applied to the classifier. As a classifier, GMM is used. The proposed method was evaluated on ASVspoof 2017 database. EER and t-DCF are taken as evaluation metrics. Better EER values are obtained with LFCC+GMM system on both development and evaluation data. t-DCF values for different systems is observed. For the development data set, the best t-DCF value (0.1379) is obtained with MFCC+GMM as ASV system and LFCC+GMM as CM system with statistical features. From the results it can be shown that, delta features along with statistical features are also giving acceptable t-DCF. With this discussion it can be concluded that ASV system with MFCC features and GMM classifier, CM system with LFCC features and GMM classifier performs better with both statistical and statistical + delta features.

FUTURE SCOPE

Further improvements to this system can be made. Analysis can be carried out with other classifiers. More and better feature extraction techniques can be adapted. Further study can be carried out on a real replay attack scenario. Performance can be analyzed using other spoofing attacks.

REFERENCES

1. Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communications, vol. 66, pp. 130-153, 2015.
2. S. Kaavya, V. Sethi, E. Ambikairajah, and H. Li, "Front- End for Antispoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition," IEEE Journal of selected topics in signal processing, vol. 11, no. 4, June 2017.
3. Hong Yu, Zheng-Hua Tan, Zhanyu Ma, Rainer Martin, and Jun Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," IEEE Transactions on neural networks and learning systems, vol. 29, no.10, pp. 4633-4644, Oct. 2018.
4. Tomi Kinnunen, Kong Aik Lee, Hector Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker verification," Odyssey 2018 The Speaker and Language Recognition Workshop 26-29 June 2018, Les Sables d'Olonne, France.
5. Longbiao Wang, Seiichi Nakagawa, Zhaofeng Zhang, Yohei Yoshida, and Yuta kawakami, "Spoofing speech detection using Modified Relative phase information," IEEE Journal of selected topics in signal processing, vol. 11, no. 4, June 2017.
6. Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee, Md Sahidullah, Massimiliano Todisco, and Hector Delgado, "ASVspoof2017: automatic speaker verification spoofing and countermeasure challenge evaluation plan," http://www.spoofingchallenge.org/.
7. D. A. Reynolds, "Speaker identification and verification using Gaussian Mixture speaker models," Speech Commun., vol. 17, no. 1-2, pp. 91-108, 1995.

AUTHORS PROFILE

Medikonda Neelima, received her B.Tech. degree in Electronics & Communication Engineering from M.L.E.C. Engineering college, Singarayakonda, M.E. degree in Electronic Instrumentation from Andhra University, Visakhapatnam, and pursuing Ph.D in the area of Speech Signal Processing at JNTUK, Kakinada. At present, she is working as an Assistant Professor in GVP college of Engineering(A), Visakhapatnam. She has more than 13 years of teaching experience. Her research areas of interests are Speech Signal Processing, Digital Signal Processing and Image Processing. She guided more than 20 B.Tech. and M.Tech. projects. She has published 3 Research papers in International Conferences. She is a life member of Professional organization IAENG.

Dr I. Santi Prabha, is Professor in ECE department of JNTUK. She did her B.Tech & M.Tech with specialization In Instrumentation and Control Systems from JNTU College of engineering, Kakinada. She was awarded with Ph.D. in Speech signal processing by Jawaharlal Nehru Technological University in 2005. She has 34 years of experience in teaching. She has guided 5 Ph.D. scholars and is presently guiding 9 Ph.D scholars. She has guided more than 100 M.Tech. Projects. She worked as Head of ECE Department and subsequently served as Director of EOW&G, JNTUK, Kakinada during 2009 – 2015. She served as Rector, JNTUK, Kakinada during 2018-2019. She is a member of various professional organizations like Fellow of Institution of Engineers (India), Fellow member of The Institution of Electronics and Telecommunication Engineers and Life member of Indian Society for Technical Education. She has published more than 100 technical papers in National and International journals/ conferences.