A Unique Glottal Flow Parameters based Features for Anti-spoofing Countermeasures in Automatic Speaker Verification

Ankita Chadha, Azween Abdullah, Lorita Angeline
School of Computer Science and Engineering
Taylors University
Subang Jaya, Malaysia

***Abstract*—The domain of Automatic Speaker Verification (ASV) is blooming with growing developments in feature engineering and artificial intelligence. Inspite of this, the system is liable to spoofing attacks in the form of synthetic or replayed speech. The difficulty in detecting synthetic speech is due to recent advancements in the Voice conversion and Text-to-speech systems which produce natural, indistinguishable speech. To prevent such attacks, there is a need to develop robust spoof detection systems. In order to achieve this goal, we are proposing estimation of Glottal Flow Parameters (GFP) from speech of genuine speech and synthetic spoof samples. The GFP are further parameterized using time, frequency and Liljencrants–Fant (LF) models. Along with GFP features, the Linear Prediction Cepstrum Co-efficient (LFCC) and statistical parameters are computed. The GFP features are investigated to prove their usefulness in detecting spoofed and genuine speech. The ASV spoof 2019 corpus is used to test the framework and evaluated against the baseline models. The proposed spoof detection framework produces an Equal Error Rate (EER) of 2.39% and tandem Detection Cost Function (t-DCF) of 0.0562 which is found to be better than the state-of-the art technique.**

***Keywords*—Spoof detection; synthetic speech; glottal excitation; speaker verification; voice conversion; text-to-speech**

I. Introduction

The speaker verification system acknowledges the true identity of a known speaker while dismissing the unknown speaker's voice [1]. These systems are bound to be exposed to the infiltrators through spoofing attacks. The intrusion in the form of synthetically generated speech results into spoofing attack on the ASV system. Such an environment is termed as Logical Access (LA) scenario while the one with replay speech is a Physical Access (PA) scenario [2]. These attacks are a result of continuous efforts by researchers in field of Voice Conversion (VC) and Text-to-Speech (TTS) [3]; since their aim is to generate clean, human like speech - with little to no variation in the synthetic speech. Hence, tackling these attacks through means of efficient features and machine learning algorithms are a desideratum. The studies in anti-spoofing or countermeasures have increased tremendously with increasing attacks on main-frame systems such as phone-banking theft, unauthentic access to workplaces or even smart phone devices where speech is used as the identity [3], [4]. So, as authentication is no more limited to finger prints and retina scans, the speech based spoofing attacks are growing and catching attention of many researchers for developing robust spoofing detection schemes. Moreover, the countermeasures developed so far are less than a decade old and still have a scope of improvement in terms of reducing the False Acceptance ratios. Most of the research is based on specific type of attack [5], [6] while few others consider all the types of attack making them universal detectors [7], [8].

II. Related Work

The anti-spoofing measures are solely dependent on two prime techniques: feature representation and spoofed speech classification. The studies on features are significant and need to be based on the nature of input speech which is either genuine or spoofed. Thus, the task is restricted to differentiate between spoofed and genuine speech through appropriate use of features for extracting relevant information from the test speech. The spectral features employed for spoofing detection are Mel-Frequency Cepstral Co-efficient (MFCC) [9], [10], Magnitude and Phase based features [11] such as Log Magnitude Spectrum, Residual Log Magnitude Spectrum, Group Delay (GD), Modified GD (MGD), Instantaneous Frequency (IF), Baseband Phase Difference and Pitch Synchronous Phase (PSP). Additionally, the known fact that the MFCCs represent the human auditory system as it utilizes perceptually similar filter bank analysis, is found to be performing not so well in the anti-spoofing environment [11]. To counter that, the Inverse MFCC (IMFCC) is proposed for spoof detection because it comprises of feature contents which are absent in MFCC [12]. Furthermore, the CFCCIF, CQCC based features were also proposed; out of which CQCCs are considered to outperform in the ASV Spoof 2017 challenge [13], [14].

The features extracted are trained using machine learning algorithms ranging from generative models like i-vectors [15], Gaussian Mixture Models (GMM) [10], [16], Universal Background Models (UBM)[17], [18], [19] and Joint Factor Analysis [15] to discriminative models like Support Vector Machines (SVM) [20], Deep Neural Networks [21], [22], [23] and its variants like Recurrent Neural Networks (RNN) [24], [25], Deep Residual Neural Networks [13] and Convolutional Neural Network (CNN) [26], [27]. The GMM are considered to be efficient in capturing the generality and non-linearities in data [2]. Therefore, we are using the state-of-the art GMM for learning the pattern to differentiate genuine and spoofed speech.

The speech signal generated by lungs act as a source of air that stipulates excitation from glottis resulting into resonating

frequencies traveling through the vocal tract out of the mouth. Hence, the lip radiation is also considered as the part of the production mechanism but is stable. Thus analytically, the contents available from speech may be in the form of meaning of the utterance and individual speaker's identity. For designing the counter-measure to detect an attack, the extraction of speaker related information and artefacts inserted due to synthetic speech is a crucial step. Both identity of speaker and meaning of sample can be interpreted at different areas of the production mechanism like shape of Vocal Tract (VT), nature of Glottal Excitation (GE) or flow and prosody parameters [28]. The work in this research is based on analysing the source of the speech production model, i.e. glottal source estimation technique. The research in [20] used IAIF estimation for glottal flow estimation but focused more on the classifiers (SVM and ELM). Along with this, we consider the VT information which captures the speaker's individuality in the form of LFCC [29] with statistical parameters. Also, the few studies have shown glottal excitation to be independent of VT [30] while some have shown inter-dependency between them [31], [32], [33]. Hence, we found it necessary to explore glottal excitation components of genuine and spoof speech. Furthermore, the scope of the research is also confined to LA attacks as synthetic speech production is becoming more accessible and capturing naturality. This is due to the fact that open source tools and datasets are available for researchers to explore leading to more versatile synthetic speech generators [5], [23], [34], [35].

Thus, the research approach is divided in a three-fold process and is listed as follows:

1) Exploring the Glottal Flow Parameters (GFP) using Quasi-Closed Phase estimation and LF modelling to capture the inaudible artefacts present in the synthetic speech through careful representation of source excitation process.
2) Investigating the performance of these GFP features using objective metrics in the GMM framework.
3) Conducting comparative analysis of the proposed features with the Baseline LFCC features [2].

The article is organized as follows: Section III describes the Glottal excitation estimation based Feature Extraction while Section IV elaborates the Proposed Anti-spoofing based speaker verification system. The Section V presents the experimental results while overall discussion and conclusion are summarized in Sections VI and VII, respectively.

III. Glottal Excitation Estimation Based Feature Extraction

The estimation of source of the speech by filtering out the effects of lip radiation and vocal tract is termed as Glottal inverse filtering (GIF). The first research on glottal source estimation began in 1950s by Miller [36]. Since then, improvements were seen in representing glottal source, but it has been difficult to compute due to lack of ground truth like no EGG information available. Furthermore, studies directed towards utilizing synthetic speech to work on in order to avoid the need for ground truth [37]. In the spoof detection task, this research is analyzing natural as well as synthetic speech (which is indeed spoofed speech). The GIF analysis was initially based on closed phase, iterative and adaptive approaches [38]. The Closed phase estimation is based on the covariance criteria for Linear Prediction (LP) analysis as some samples which are present in closed phase. Another approach that requires prior knowledge of shapes of both vocal tract as well as glottal excitation is the Iterative Adaptive Inverse Filtering (IAIF) [31]. The mixed phased approaches like Complex Cepstrum analysis [39] and zeros of Z-transform (ZZT) [40] are contrasting to the earlier estimation techniques as they consider segregation of glottal and vocal tract information through transformation in another domain (such as frequency or z-domain). Furthermore, the Mean-Square Phase (MSP) is used to approximate the Liljencrants-Fant (LF) model [41]. Most of approaches mentioned so far perform well for low pitched male voices and deteriorate for higher fundamental frequencies (f0) [38]. This research is based on Quasi-Closed Phase (QCP) glottal estimation that uses Weighted Linear Prediction (WLP) in place of covariance criteria as shown in Fig. 1. It is found that this kind of estimation is more robust in the closed phase parts of the speech samples [38]. Also, so far studies have been conducted on VT contents of the speech whereas the glottal excitation is equally important as it bears the source of speech production system.

The speech produced because of convolution in time domain, s_m turns out to be product of individual frequency responses of GE source, $G(z)$ and VT filter $T(z)$. Thus, speech signal $S(z)$ in z-domain is given in Equation 1

$$S(z) = G(z).T(z) \tag{1}$$

So, using the conventional LP approach for portraying the WLP model for m^{th} speech utterance as shown in Equation 2

$$s_m = \sum_{j=1}^{L} s_{m-j} b_j + e_m \tag{2}$$

Where, e_m is excitation signal with j^{th} b_j prediction coefficient of order L. The significant difference between WLP and LP analysis is that the WLP yields the product of weight function W_m with square of the excitation signal given in the form of Total energy residual E (in Equation 3):

$$E = \sum_{m=m_1}^{m_2} (s_m - \sum_{j=1}^{L} s_{m-j} b_j)^2 W_m \tag{3}$$

For auto-correlation criteria, the limits $m_1 = 1$ and $m_2 = M+L$; M is the length of frame. The weight function, W_m is given in Equation 4 using Attenuated Main Excitation (AME) function.

$$W_m = \sum_{i=0}^{N-1} s_{m-1-j}^2 \tag{4}$$

The Glottal Flow waveform obtained from the raw speech samples of genuine speech (Fig. 2a), TTS synthetic speech (Fig. 2b) and the VC speech (Fig. 2c) signify the difference in time, frequency and phase contents of genuine and synthetic speech samples.

The QCP parameters include the time, amplitude and frequency domain traits contributing to 31 Glottal flow descriptors. The time domain parameters considered in this research

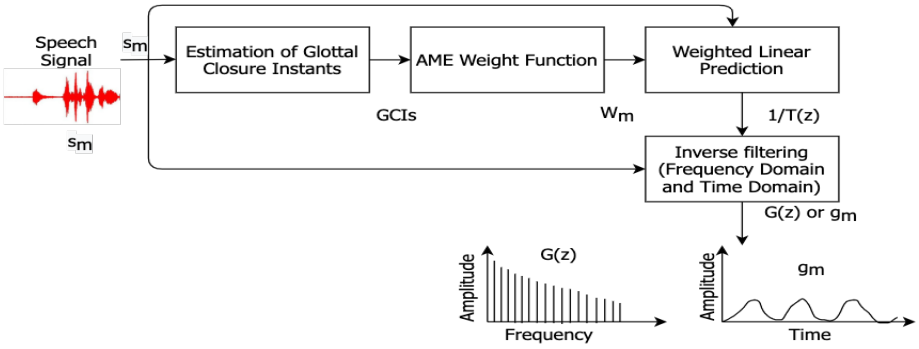


Fig. 1. Block Diagram of GFP based Feature Extraction using QCP Estimation.

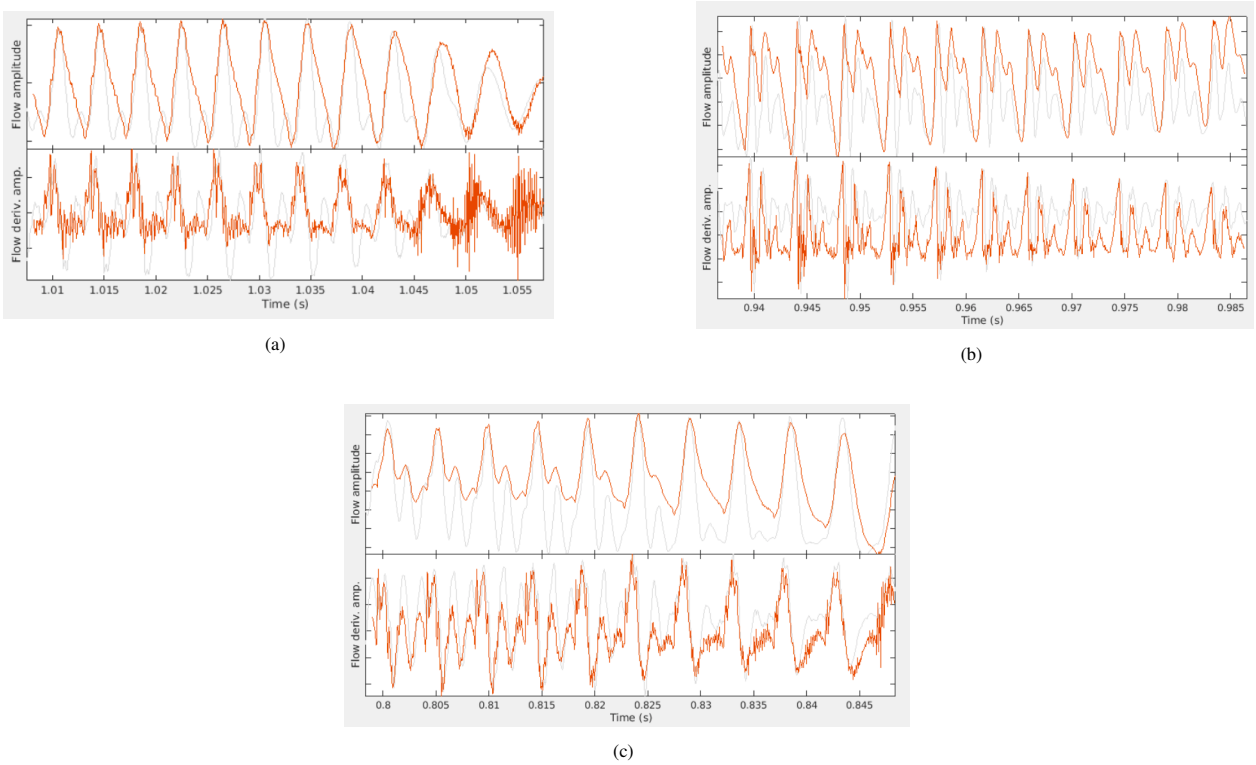


Fig. 2. Glottal Flow Derivative and Amplitude Waveform for (a) Genuine (b) TTS (c) VC Speech Samples.

are based on open quotient (OQ), speed quotient (SQ), and closing quotient (ClQ) while the amplitude parameters are based on Amplitude Quotient (AQ). Lastly, the frequency domain parameters such as Parabolic spectrum parameter (Psp), difference value between amplitude of first and second harmonic (H1-H2) and Harmonic Richness Factor (HRF) which are adapted from [37] are also computed as a part of GFP features.

IV. Proposed Anti-spoofing Speaker Verification Framework

A spoof detection or anti-spoofing algorithm must be designed by carefully choosing the right features which represent the spoof and genuine speech in order to make the differentiation task easier. Hence, the choice of appropriate classifier too, is crucial. To summarize the spoof detection system, there two primary phases, namely the training phase and the testing phase as shown in Fig. 3. The training phase involves extracting the GFP, LFCC and statistical features after pre-processing of the raw speech data. These features are fed to the GMM classifier using associated labels. The individual models for genuine and spoofed samples are used in the testing phase to categorize the unknown test sample. The details steps: parameterization, model training and decision making algorithm are described in further sub-sections.

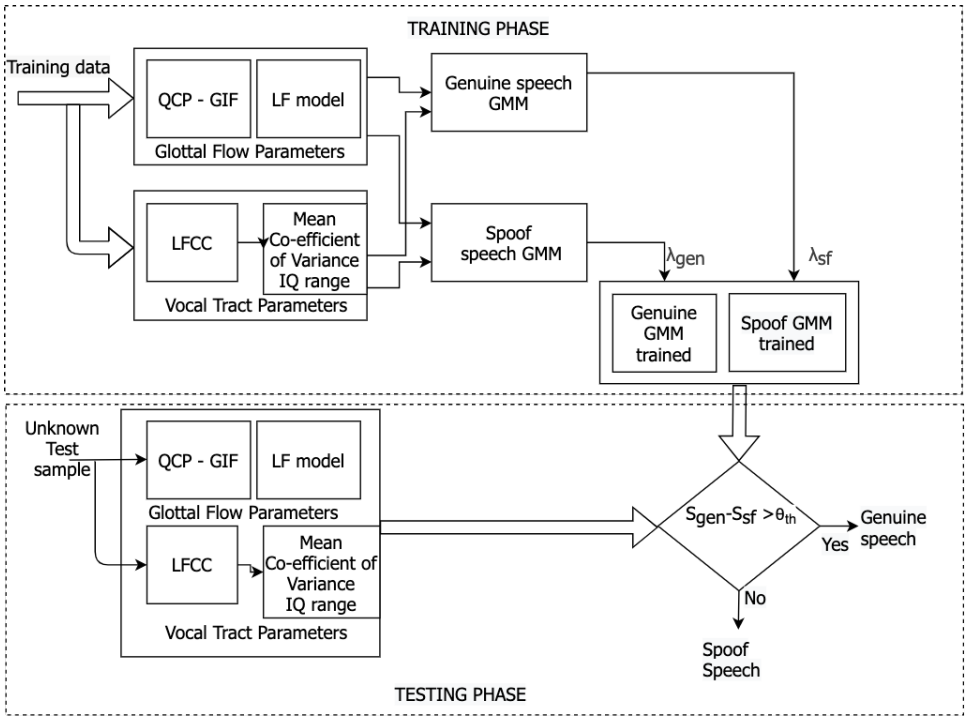


Fig. 3. Block Diagram of Proposed Counter Measure Framework for Automatic Speaker Verification System.

A. Parameterization

During the training stage, the speech samples are low pass filtered along with framing. The silence and pauses at the beginning and end of the sample are removed using Voice activity detection [20]. The VT filter information is represented using LFCC with $20ms$ frame size. Furthermore, the statistical parameters like mean, coefficient of variance (CoV) and Interquartile range (IQR) are combined with the LFCC parameters to form a feature matrix. The LFCC features are found to be more robust than MFCC in terms of noisy speech as it performs well in the higher frequency region (comprising of VT features). The order of LFCC is 19 and its delta and double delta variants are also computed.
The VT filter features alone are not sufficient to represent the speech, especially when the naturally spoken speech needs to be differentiated as against spoofed speech. According to the speech production model, the remaining glottal excitation information is represented using GFP estimation through QCP GIF technique using $30ms$ frame length. The time-based parameters such as OQ, SQ and ClQ are computed in Equation 5:

$$OQ = \frac{(L_{01} + L_c)}{L}, SQ = \frac{L_{01}}{L_c}, ClQ = \frac{L_c}{L} \tag{5}$$

where, L_{01} is the opening phase length (expressed in time), L_c is closing phase length and L is glottal cycle length which in terms of period. The AQ is computed using Equation 6

$$AQ = \frac{A_{max}}{d_{min}} \tag{6}$$

where A_{max} is glottal peak and d_{min} is minimum value of derivative of glottal time waveform. The Normalized AQ (NAQ) is given using Equation 7

$$NAQ = \frac{AQ}{L} \tag{7}$$

Apart from these, the Quasi OQ (QOQ), HRF, Psp and H1H2 are also used as a part of GFP. Additionally, the LF model parameters such as E_e, R_a, R_g and R_k are also considered as they are dependent on the linear source filter model (Table I shows details of parameters). A subjective test is performed (Fig. 4) to discern proficiency of GFP descriptors, box plot analysis is used to display the numerical values of the genuine and spoof speech samples for AQ, QOQ, HRF and H1H2.

From Fig. 4 for AQ and QOQ, it is found that the IQR for genuine and spoof speech are different while for H1-H2 and HRF the IQR values between genuine and spoof speech are slightly similar. Hence, the AQ and QOQ have higher discrimination properties than H1-H2 and HRF.

B. Model Training and Decision Making Algorithm

The GFP parameters, LFCC features, and statistical parameters together form a feature matrix for each sample of the entire data in the spoofed and genuine category individually. In this study, we use the GMM based binary classifier with 512 mixtures for modelling the class labels according to genuine or spoofed speech. The GMM model in case of genuine speech samples λ_{gen} while for the spoofed sample is λ_{sf}. The GMM are considered to capture higher classification accuracy due to their ability to capture generality in case of unknown data samples. For a particular test utterance T, the Log Likelihood

TABLE I. LIST OF DESCRIPTORS BELONGING TO GFP BASED FEATURE EXTRACTION

TIME DOMAIN PARAMETERS	
OQ1 , OQ2	These are Open Quotients computed using Primary and secondary opening of the glottis.
NAQ	Normalized Amplitude Quotient
AQ	Amplitude Quotient
ClQ	Closing Quotient.
OQa	Variant of OQ obtained from LF model
QOQ	Quasi-Open Quotient State
SQ1, SQ2	Speed Quotients These are computed from the primary (OQ1) and openings (OQ2)
TPO	Time corresponding to primary opening
TSO	Time corresponding to secondary opening
TC	Closing time
TMAX	Time corresponding to maximum flow of air pressure
TMIN	Time corresponding to minimum flow of air pressure
TDMIN	Time corresponding to minimum of the derivative
TDMAX	Time corresponding to maximum of the derivative
TQO	Time corresponding to quasi-opening time
TQC	Time corresponding to quasi-closing time
FREQUENCY DOMAIN PARAMETERS	
Psp	Parabolic spectrum parameter corresponds to the second-order polynomial wrt the flow spectrum over a single glottal cycle.
DH12	This is the H1-H2 parameter represented in decibels
HRF	Harmonic richness factor is ratio higher harmonics like f2, f3 etc to f1 (first harmonic)
LF PARAMETERS	
t0	Time corresponding to start of opening phase
tp	Time corresponding to peak of the speech wave
te	Time corresponding to derivative of min peak value
ta	Time corresponding to return phase
Ee	Amplitude corresponding to negative peak of glottal pressure wave in percentage
RA	ta x f0 (where f0 is fundamental frequency)
RG	0.5 f0 x tp
RK	(te-tp)/ tp
OQ	(te + ta) f0
QO	te x f0

Ratio (LLR) is computed from likelihood values of genuine and spoofed speech models. The decision (R) of the test utterance being genuine or spoofed is relying on the LLR as shown in Equation 8

$$R = log(p(T|\lambda_{gen})) - log(p(T|\lambda_{sf})) \tag{8}$$

Where, the likelihood scores obtained from GMM for genuine and spoofed speech samples are $s_{gen} = log(p(T|\lambda_{gen}))$ and $s_{sf} = log(p(T|\lambda_{sf}))$ respectively.

V. EXPERIMENTAL RESULTS

The research is based on ASV spoof 2019 dataset [42] which was the part of ASV spoof challenge held in 2019. The corpus consists of 20 speakers and more than fifty thousand samples in LA attack samples. For training we used 2580 genuine and 22800 spoof samples while 23400 samples are used for development purpose as shown in Table II.

TABLE II. NUMBER OF SAMPLES IN ASV SPOOF 2019 CORPUS FOR TRAINING AND DEVELOPMENT

	Subset	
Logical Access	**Training Data**	**Development Data**
Genuine	2580	2548
Spoof	22800	22296
Total	25380	24844

So far, this is the only dataset with such a wide variety of samples and attack types. The state-of-the-art LFCC-GMM technique is considered as the baseline approach [2]. Furthermore, the process of binary classification leads to two error types: False Acceptance Ratios (FRR) and the False Rejective Ratios (FRR). A standalone spoof detection scheme may falsely reject a genuine sample assuming it to be spoofed or falsely accept an imposter sample assuming it to be genuine. Based on these errors, the DET is used to measure performance of the features used. The operating point obtained from the DET curve is the EER which is another metric for evaluating the spoof detection performance [2]. Lastly, the normalized tandem-Detection Cost Function (t-DCF) [2] is also used to measure performance as it does not require pre-setting of decision threshold and is given in Equation 9

$$norm\ \ t - DCF = p_{FR} + a\ p_{FA} \tag{9}$$

Where p_{FR} probability for scores which are less than set threshold considered as rejected while p_{FA} is the probability for scores which are greater than the set threshold (a) considered as accepted test sample. The ASV and CM scores performance, DET Curve and CM results using EER and t-DCF plots are depicted from Fig. 5 to Fig. 7 and Table III.

The Fig. 5 depicts probability density function (pdf) of ASV and CM scores. The CM scores are for Baseline (red) and the Proposed model (blue). Both models are bimodal except in case of the Proposed model the density has smaller peak in comparison to a more definitive peaks for Baseline model signifying lower pdf for the baseline with two opposite distributions. Fig. 6 and Fig. 7 show the t-DCF and DET curves

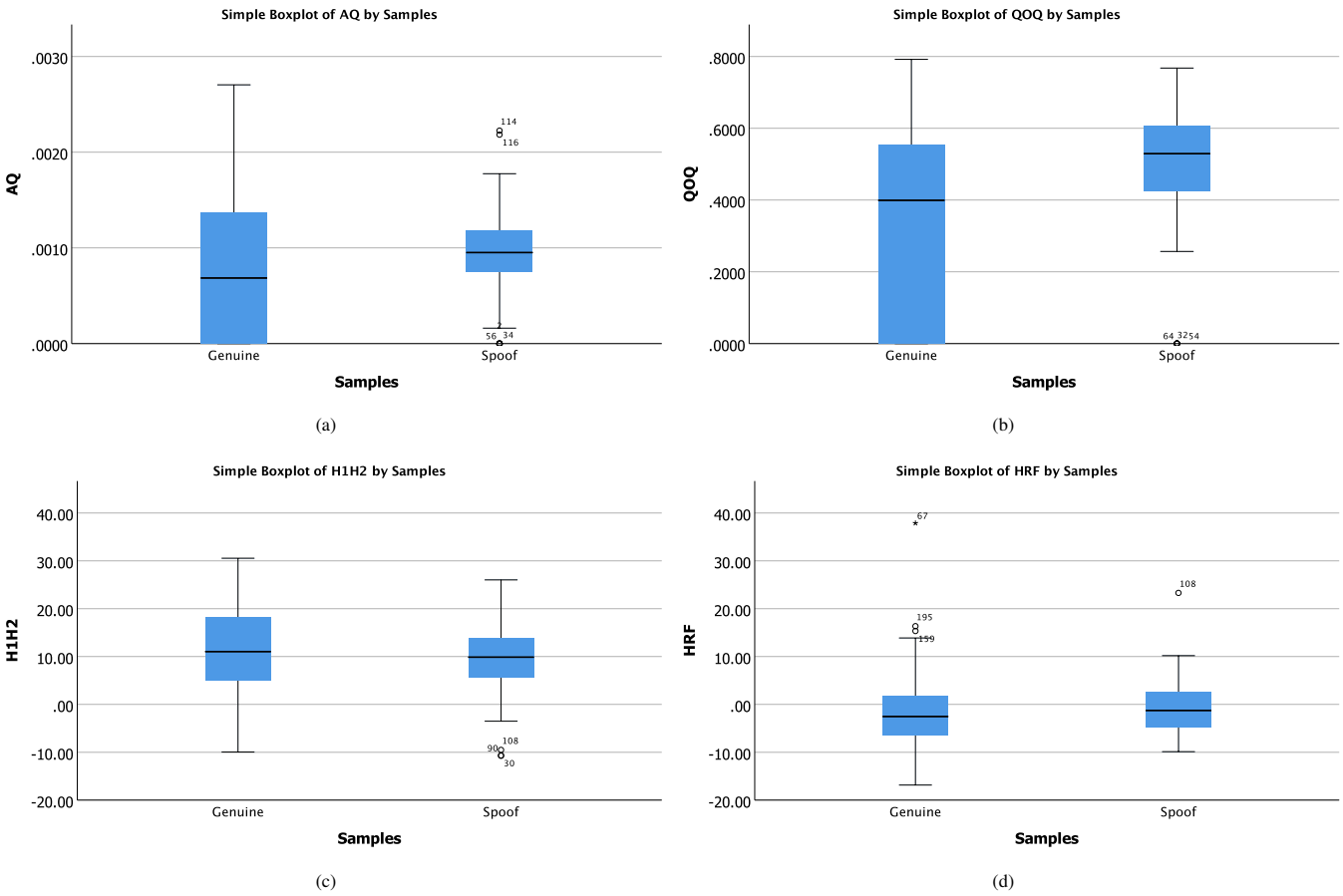


Fig. 4. Subjective Analysis of GFP based Features using ASV Spoof 2019 Dataset for (a) AQ (b) QOQ (c) H1H2 (d) HRF.

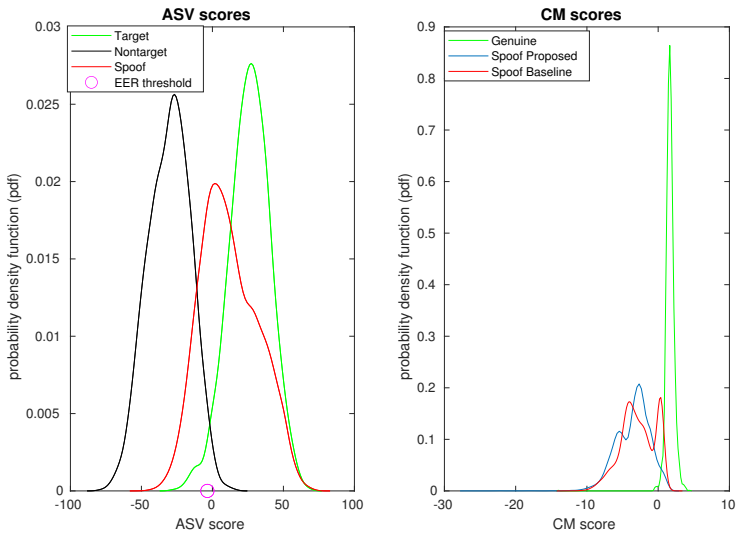


Fig. 5. Probability Density Function (PDF) for Scores of ASV and CM.

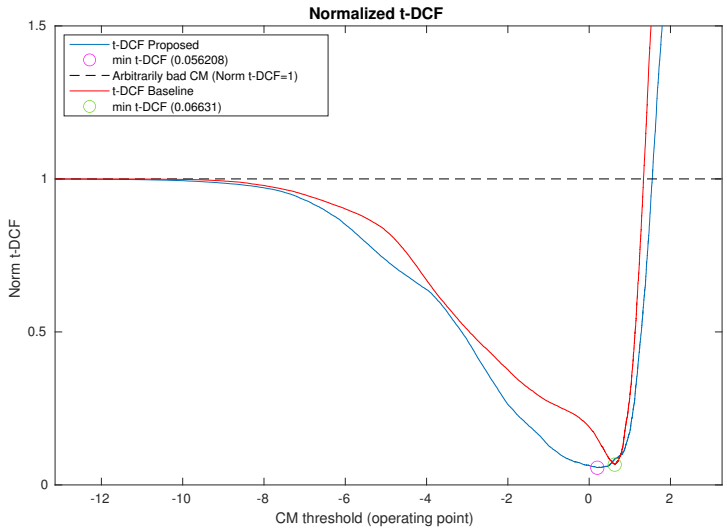
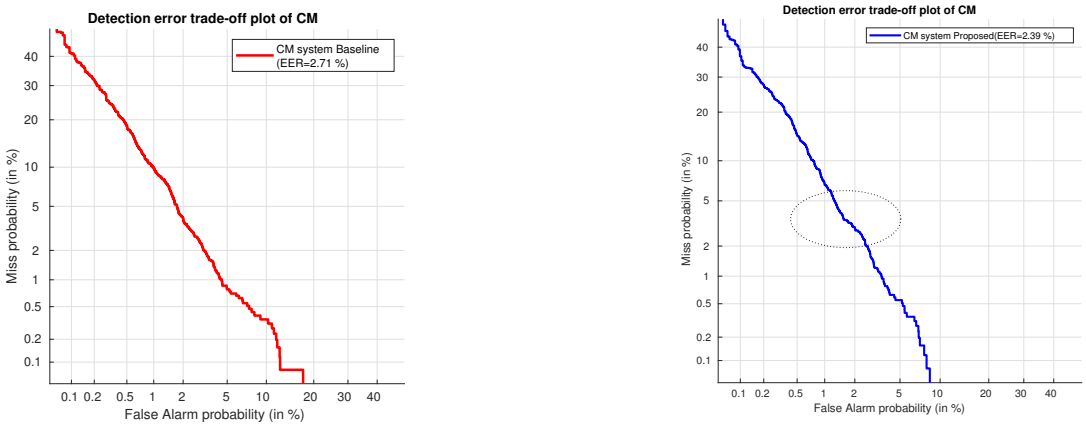


Fig. 6. Normalized t-DCF Plot for Baseline and Proposed CM Algorithm.



(a) (b)

Fig. 7. DET Curve for (a) Baseline CM (b) Proposed GFP CM.

respectively. The t-DCF is lower for proposed technique in comparison to the baseline model. Also, the DET curve shows slightly lower EER for proposed technique in contrast to the baseline method (shown in Table III).

TABLE III. EER AND NORMALIZED T-DCF SCORE FOR BASELINE AND PROPOSED TECHNIQUE

CM type	EER %	t-DCF
Baseline LFCC -GMM	2.708	0.066
Proposed GFP+ LFCC+ stats -GMM	2.390	0.056

VI. DISCUSSION

The GFP-based features are unique and not much explored in the spoof detection domain. The Glottal-flow plots for synthetic speech highlight the significant difference in the amplitude, time, and frequency information from the genuine speech. This ascertains the importance of proposed GFPs in addition to VT parameters in developing countermeasures. Also, the selection of the right GFPs is crucial. Thus, we plotted the box plot to investigate which parameters are more reliable than the others. For instance, the AQ captures the glottal peaks accurately and due to the synthetic nature of spoofed speech, the amplitude information is found to be deviating from genuine speech. While on the other hand, the HRF represents the quality of speech which may perceptually similar. Hence, detecting spoofed speech from genuine is slightly difficult with HRF and similar parameters. In contrast, the GFPs on the whole when used in the conjunction with VT parameters show improvement in the EER and t-DCF when compared to the baseline technique. This might be due to the fact that missing glottal flow information is now fulfilled by the 31 QCP Glottal features that represent amplitude along with with time-frequency contents; and also due to the fact that the high pitched voices are now easily detected with these proposed GF features leading to better results.

VII. Conclusion

The main role of a counter measure is to prevent any unauthentic access. For doing so, the kind of attack and spoofed speech must be analysed. Hence, in this research, we focused on the synthetic speech attack using unique QCP estimation for extracting GFP from both genuine as well as spoof speech. Since, the GFP represents the source of attack samples, the minute differentiation between genuine and spoof speech was magnified with GFP. As a result, GFP certainly added the information contents to the features set by further reducing the EER from 2.70% for Baseline LFCC to 2.39%. So, the FAR and FRR can be reduced by extracting relevant information from spoofed speech. Additionally, the GMM classifier captured the non-linearities quite well as the conjugative contribution of GFP and LFCC provided sufficient data for better classification accuracy. Also, this research can further be extended for replay speech where noise based artifacts may be present and GFPs are found to perform significantly well in noisy speech. In addition to the improvements obtained by employing the QCP based GF parameters, there are two prime limitations of these features: first, the QCP based GIF requires precise estimation of GCI. This can be explored in the future by investigating more appropriate GCI estimation techniques. Secondly, the unstable filter parameters contribute to computational complexity while extracting these features. From future prospects, the prosodic features may be explored in conjunction with source filter parameters for further reducing the EER and improving the countermeasure performance.

Acknowledgment

The authors would like to thank the Taylors University, Malaysia for sponsoring this research. Also, a genuine appreciation for School of Computer Science and Engineering for their support.

References

[1] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagish, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 4, 2016.

[2] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. Aik Lee, V. Vestman, and A. Nautsch, "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan *." [Online]. Available: http://dx.doi.org/10.7488/ds/1994

[3] A. Y. Kuznetsov, R. A. Murtazin, I. M. Garipov, E. A. Fedorov, A. V. Kholodenina, and A. A. Vorobeva, "Methods of countering speech synthesis attacks on voice biometric systems in banking (Review article)," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 21, no. 1, 2021.

[4] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of Audio Deepfake Detection," 2020.

[5] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer Learning from Speech Synthesis to Voice Conversion with Non-Parallel Training Data," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, 2021.

[6] K. Phapatanaburi, L. Wang, S. Nakagawa, and M. Iwahashi, "Replay Attack Detection Using Linear Prediction Analysis-Based Relative Phase Features," *IEEE Access*, vol. 7, 2019.

[7] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, 2017.

[8] M. G. Kumar, S. R. Kumar, M. S. Saranya, B. Bharathi, and H. A. Murthy, "Spoof Detection Using Time-Delay Shallow Neural Network and Feature Switching," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings*, 2019, pp. 1011–1017.

[9] G. Nijhawan, "Robust automatic speaker recognition system," Ph.D. dissertation, Manav Rachna International University, 2015. [Online]. Available: http://hdl.handle.net/10603/104402 http://shodhganga.inflibnet.ac.in:8080/jspui/handle/10603/104402

[10] A. Paul, R. K. Das, R. Sinha, and S. R. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *2016 International Conference on Signal Processing and Communications, SPCOM 2016*, 2016.

[11] L. Liu and J. Yang, "Study on Feature Complementarity of Statistics, Energy, and Principal Information for Spoofing Detection," vol. 8, 2020.

[12] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-Janua, 2015.

[13] Z. Li, J. Wei, Q. S. . A.-P. C. On, and U. 2021, "Time-frequency Resolution Optimization Features on Spoof Detection," in *Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9407631/

[14] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," 2018.

[15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, 2011.

[16] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVSpoof 2019: Future horizons in spoofed and fake audio detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, 2019, pp. 1008–1012.

[17] C. Hanilçi, "Data selection for i-vector based automatic speaker verification anti-spoofing," *Digital Signal Processing*, vol. 72, pp. 171–180, jan 2018.

[18] H. Zeinali, H. Sameti, and L. Burget, "HMM-based phrase-independent i-vector extractor for text-dependent speaker verification," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 7, 2017.

[19] M. Singh and D. Pati, "Usefulness of linear prediction residual for replay attack detection," *AEU - International Journal of Electronics and Communications*, vol. 110, p. 152837, oct 2019.

[20] R. Rahmeni, A. B. Aicha, and Y. B. Ayed, "Speech spoofing countermeasures based on source voice analysis and machine learning techniques," in *Procedia Computer Science*, vol. 159, 2019.

[21] H. Yu, Z. H. Tan, Z. Ma, R. Martin, and J. 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.

[22] M. Y. Faisal and S. Suyanto, "SpecAugment Impact on Automatic Speaker Verification System," in *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*. Institute of Electrical and Electronics Engineers Inc., dec 2019, pp. 305–308.

[23] R. Liu, B. Sisman, and H. Li, "Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability," in *INTERSPEECH*, 2021. [Online]. Available: https://ttslr.github.io/i-ETTS.

[24] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, 2016.

[25] N. Do, "Neural networks for automatic speaker, language, and sex identification," feb 2016. [Online]. Available: https://dspace.cuni.cz/handle/20.500.11956/77265

[26] S. Saranya, S. Rupesh Kumar, and B. Bharathi, "Deep Learning Approach: Detection of Replay Attack in ASV Systems," in *Advances in Intelligent Systems and Computing*, vol. 1118. Springer, jun 2020, pp. 291–298.

[27] J. Yang, H. Wang, R. K. Das, and Y. Qian, "Modified Magnitude-Phase Spectrum Information for Spoofing Detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, 2021.

[28] D. Paul, M. Pal, and G. Saha, "Spectral features for synthetic speech detection," *IEEE Journal on Selected Topics in Signal Processing*, vol. 11, no. 4, 2017.

[29] G. P. Prajapati, M. R. Kamble, and H. A. Patil, "Energy separation based features for replay spoof detection for voice assistant," in *European Signal Processing Conference*, vol. 2021-Janua, 2021.

[30] A. Barney, A. De Stefano, and N. Henrich, "The effect of glottal opening on the acoustic response of the vocal tract," in *Forum Acusticum Budapest 2005: 4th European Congress on Acustics*, 2005.

[31] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2-3, 1992.

[32] T. Drugman, T. Dutoit, and B. Bozkurt, "Excitation-based Voice Quality Analysis and Modification," jan 2020.

[33] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech and Language*, vol. 28, no. 5, 2014.

[34] C. Y. Huang, Y. Y. Lin, H. Y. Lee, and L. S. Lee, "Defending Your Voice: Adversarial Attack on Voice Conversion," in *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, 2021.

[35] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-Tacotron: Spectrogram-Free End-to-End Text-to-Speech Synthesis," 2021.

[36] R. L. Miller Bell, "Nature of the vocal cord wave," *Journal of the Acoustical Society of America*, vol. 31, no. 6, 1959.

[37] P. Alku, "Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, 2011.

[38] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, 2014.

[39] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[40] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of Z-transform representation with application to source-filter separation in speech," *IEEE Signal Processing Letters*, vol. 12, no. 4, 2005.

[41] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, 2013.

[42] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y. H. Peng, H. T. Hwang, Y. Tsao, H. M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L. J. Liu, Y. C. Wu, W. C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J. F. Bonastre, A. Govender, S. Ronanki, J. X. Zhang, and Z. H. Ling, "ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech," *Computer Speech and Language*, vol. 64, p. 101114, nov 2020.