Exploring Source Features with Deep Residual Neural Networks for Replay Attack Detection

Suresh Veesa*, Badugu Vamsi Krishna† and Madhusudan Singh‡
* Department of Electronics and Communication Engineering
National Institute of Technology Nagaland
Chumukedima-797103, Nagaland, India
E-mail: veesa1034@gmail.com, vamsi@nitnagaland.ac.in and madhu@nitnagaland.ac.in

***Abstract*—Replay spoofing raises a serious challenge to the reliability of automatic speaker verification (ASV) systems, particularly in real-world applications. While most countermeasures have concentrated on spectral features, excitation source information based features remains underexplored. This study addresses this gap by leveraging Linear Prediction Residual (LPR) features, which capture critical excitation source characteristics relevant for replay detection. Specifically, we investigate the effectiveness of Residual Constant Q Cepstral Coefficient (RCQCC), Residual Mel-Frequency Cepstral Coefficient (RMFCC), and Residual Phase Constant Q Cepstral Coefficient (RPCQCC) features, in conjunction with log-spectrogram representations. A deep residual neural network (DRNN) classifier is developed to fully exploit these LPR-based features. Evaluation on the ASVspoof 2017v2.0 (17PA) and ASVspoof 2019 PA (19PA) datasets demonstrates that fusing spectral and source-based features significantly improves detection performance. The best fusion model reports an equal error rate (EER) value of 8.40% on 17PA and records a tandem detection cost function value of (t-DCF) 0.1447 on 19PA. These findings highlight the value of integrating excitation source information with spectral features and advanced deep learning models to strengthen replay spoofing countermeasures in ASV systems.**

I. Introduction

Automatic Speaker Verification (ASV) is a biometric technology that authenticates a person's identity using their unique voice characteristics [1], [2]. It can be applied in a variety of domains, particularly for security-related applications where reliable user authentication is essential [3]. Despite its advantages, ASV is susceptible to spoofing attacks, wherein attackers use fake or manipulated speech to deceive the system [4]. Thus, four major spoofing attacks popular in the literature are impersonation [5], speech synthesis [6], deepfake [7], and replay [8]. Impersonation attacks are rare because mimicking voices is difficult. Speech synthesis, voice conversion, and deepfakes require expertise in speech signals analysis. A replay attack uses recorded voice to deceive an ASV system. Among various types of spoofing, replay attacks are especially concerning because they are easy to perform and increasingly common. As a result, replay attacks have become a central concern in designing robust countermeasures for ASV systems [8]. Early approaches predominantly relied on spectral features and traditional classifiers. For example, study [9], [10] proposed the use of sub-band energy-based features, including Teager energy cepstral coefficients (TECC), combined with Gaussian mixture model (GMM) classifiers, achieving an equal error rate (EER) of 11.41% on the ASVspoof 2017 dataset. Similarly, in study [11], cochlear filter cepstral coefficients (CFCC) and instantaneous frequency-based features were proposed and further enhanced using energy separation algorithms. In [12], two models— Constatn Q cepstral coefficient (CQCC) - GMM and Linear frequency cepstral coefficient (LFCC) - GMM were proposed, achieving an EERs of 9.87% and 11.96%, respectively, on the ASVspoof 2019 PA database.

In another study [13] explored the excitation-based features, proposing the all pole group delay function (APGDF) and reporting an EER of 12.44% when fused with spectral features. In addition, several studies have investigated multi-feature fusion approaches to improve the robustness of spoofing detection. Studies [14], [15] have proposed using spectral-based CQCC features and energy information, combined with a ResNet classifier. By employing fusion techniques, these approaches achieved improved results on the ASVspoof2019 PA database. Deep learning innovations have played a key role in shaping modern anti-spoofing systems. Deep neural networks, particularly CNN and their variants, have demonstrated effectiveness in modeling complex speech representations for spoofing detection [16], [17]. Despite these advances, integrating source features with deep learning frameworks remains a promising direction for developing more effective anti-spoofing techniques [18]. Accordingly, this research utilizes Deep Residual Neural Networks (DRNN) to analyze linear prediction residual (LPR) features for effective replay attack detection. The rest of this paper is arranged in the following manner: Section II presents the LPR source representations. Section III describes the model design and implementation. Section IV provides the experimental results along with a detailed discussion. Lastly, Section V presents the conclusion and highlights potential directions for future work.

II. LP Residual Source Information

Speech is produced by the interaction of two main components: the vocal tract and the excitation source. The vocal tract acts as a resonator, shaping the sound, while the excitation source comes from rapid airflow impulses originating from the vibrations of the vocal folds [20], [21]. Speaker-specific traits are mainly reflected in the resonances of the vocal tract,

979-8-3315-7206-8/25/$31.00

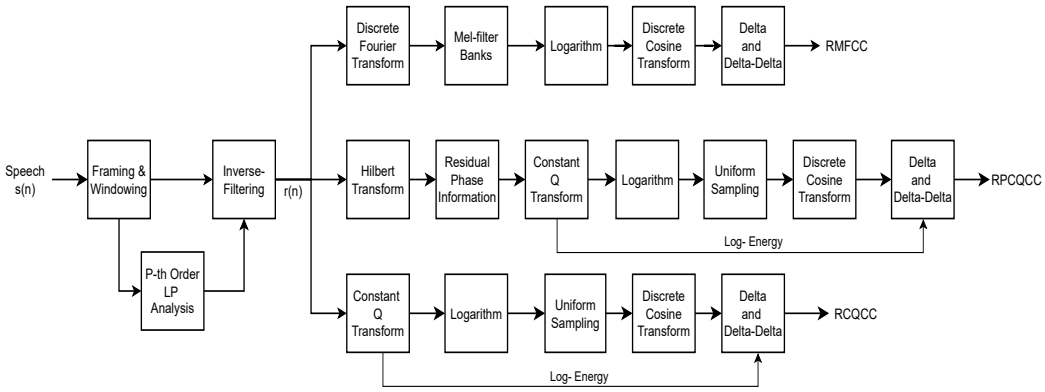

Fig. 1. The figure illustrates the feature extraction stages for RCQCC, RMFCC, and RPCQCC.

TABLE I
OVERVIEW OF ASVSPOOF2017 VERSION 2.0(17PA) AND ASVSPOOF2019 PA(19PA) DATASETS.

Database	17PA		19PA	
	Natural	Replay	Natural	Replay
Training-set	1508	1508	5400	48600
Development-set	760	950	5400	24300
Evaluation-set	1298	12008	18090	116640

TABLE II
RECENT STUDIES BENCHMARK PROPOSED MODELS USING THE ASVSPOOF 2017 V2.0 DATASET.

Systems	EER
APGDF+GCQCC-GMM[13]	12.45
TECC-GMM[9]	10.87
CQCC+LFCC+MFCC+TECC-GMM[9]	10.45
CFCCIF-QESA- ResNet[19]	16.71

while the excitation source appears as harmonics in the speech signal. To analyze these components, linear prediction (LP) is commonly used in speech processing. The LP spectrum captures the properties of the vocal tract, while the LPR highlights the details of the excitation source. According to LP theory, the LPR signal is the difference calculated between the true speech signal ($x(n)$) and its predicted value ($\hat{x}(n)$), as shown below [22], [23]:

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^{p} \alpha_k x(n-k) \qquad (1)$$

Here, a_k are the linear prediction-coefficients and p is the prediction order. When speech is replayed through different devices, it often gets distorted by noise, which can alter the excitation source information [24]. These distortions can actually help us identify replay attacks, as they provide clues that are not present in genuine speech. In this paper, we focus on extracting features from the LPR signal to better capture these excitation-specific clues. We design three types of features: Residual Constant Q Cepstral Coefficient (RCQCC), Residual Mel-Frequency Cepstral Coefficient (RMFCC), and Residual Phase Constant Q Cepstral Coefficient (RPCQCC), which also uses phase information. The best LP order for extracting these features was chosen based on experimental results. The extraction process is shown in Figure 1 and explained in the next sections. By focusing on these source features, our approach aims to improve replay attack detection and complement existing methods that rely mainly on spectral features.

TABLE III
BASELINE AND CUTTING-EDGE MODELS FOR COMPARING PROPOSED MODELS USING THE ASVSPOOF 2019PA DATASET.

Baseline	Developmen-set		Evaluation-set	
	t-DCF	EER	t-DCF	EER
LFCC-GMM[12]	0.2554	11.96	0.3017	13.54
CQCC-GMM[12]	0.1953	9.87	0.2454	11.04
CQCC-ResNet[14]	0.1026	4.30	0.1070	4.43
Energy+FM-ResNet[15]	0.0934	3.93	0.1480	6.20

A. RCQCC

Recent studies have demonstrated the effectiveness of LPR source features in speaker identification tasks [25], [26]. Building on this foundation, we introduce the RCQCC as an LPR-based feature for detecting the replay spoofing attacks in this study. The extraction process for RCQCC begins by segmenting the speech signal into frames using a Hamming window, followed by linear prediction (LP) analysis to obtain the residual signal. The Constant-Q Transform (CQT) is then applied to the residual signal to capture its frequency characteristics. Subsequently, the resulting coefficients undergo logarithmic scaling, uniform sampling, and a discrete cosine transform (DCT). In the final feature vector, the zeroth DCT coefficient is replaced with the log-energy of the frame. To capture dynamic information, delta and delta-delta coefficients are computed and appended. The complete RCQCC feature set consists of 19 static coefficients and log-energy, along with their Δ and $\Delta\Delta$ coefficients, yields a 60-dimensional feature vector. The linear prediction order is set to 18 for the ASVspoof 2017 (17PA) dataset and 10 for the ASVspoof 2019 (19PA) dataset. The overall RCQCC extraction process is illustrated in Figure 1.

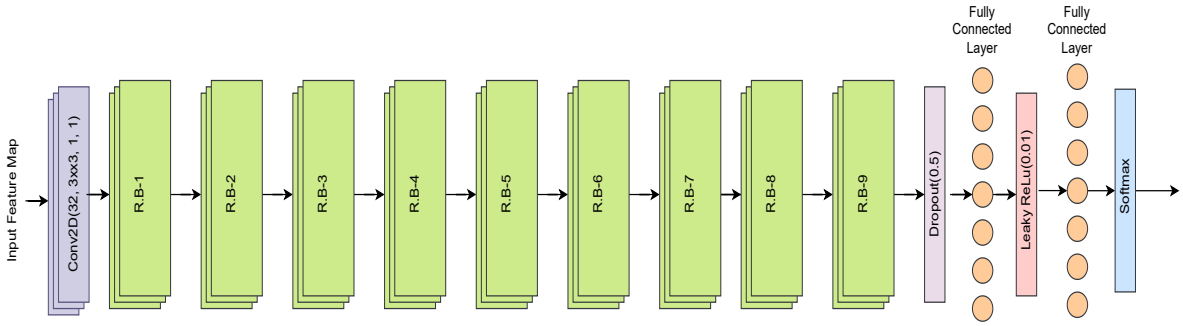

Fig. 2. DRNN model architecture with Residual Blocks(R.B-1 to 9) for the LP residual source features. Detail residual block structure shown in Fig. 3

B. RMFCC

Another LPR source feature explored in this study is the RMFCC, which is designed to enhance replay attack detection. The extraction procedure for RMFCC features is illustrated in Figure 1. Initially, the speech signal is segmented into frames using a Hamming window of 20 ms duration with a 10 ms overlap. LP analysis is then performed using an LP order of 22 for the 17PA dataset and 12 for the 19PA dataset. Inverse filtering is applied to each speech frame to obtain the LPR signal. Subsequently, A Discrete Fourier Transform is utilized to convert the LPR signal into its corresponding frequency components, which is then processed through 24 irregular mel-frequency triangular bandpass filters. The log energies of these filter outputs are calculated and subjected to a DCT to generate static cepstral coefficients. To capture temporal dynamics, the static coefficients are augmented with their delta (Δ) and delta-delta ($\Delta\Delta$) coefficients, resulting in the final RMFCC feature set. Specifically, 19 static coefficients are used, yielding a 57-dimensional feature vector.

C. RPCQCC

Additionally, this study proposes the RPCQCC as a phase information-based LPR source feature. The RPCQCC feature is extracted by applying constant-Q cepstral analysis to the residual phase signal, specifically the cosine function of the phase spectrum ($\cos\theta(n)$). The extraction process for RPCQCC is illustrated in Figure 1. The procedure for computing RPCQCC closely follows that of RCQCC, with the primary distinction being the use of the residual phase signal instead of the residual magnitude. Nineteen static coefficients, along with their delta (Δ) and delta-delta ($\Delta\Delta$) coefficients and log-energy, are computed, yielding in a 60-dimensional feature-set. An LP order of 18 is used for RPCQCC feature in both the 17PA and 19PA databases in this study. This approach effectively captures excitation-specific phase information present in the LPR signal [27], [28]. The RPCQCC feature is thus designed to detect such phase disturbances in LPR samples, making it a valuable tool for distinguishing between genuine and replayed speech signals.

D. Log-spectrogram

Spectrograms have proven to be highly effective for classification tasks across image, text, and audio domains, and are particularly well aligned for deep learning-based methods [29]. In this study, we utilize log-spectrogram features to capture the time-frequency properties of speech signals. The extraction process begins by segmenting the raw speech signal into overlapping frames using a windowing technique. Subsequently, the short-time Fourier-transform(STFT) is computed on each frame using a Hamming window with a window size of 2048 samples and a 25% overlap. The magnitude spectrum of each frame is then calculated and transformed to a logarithmic scale. This log transformation accentuates prominent features of the signal while reducing the influence of less significant components. The resulting log-spectrogram provides a detailed representation of the signal's time-frequency structure, making it particularly useful for identifying artifacts and distortions introduced by replay attacks.

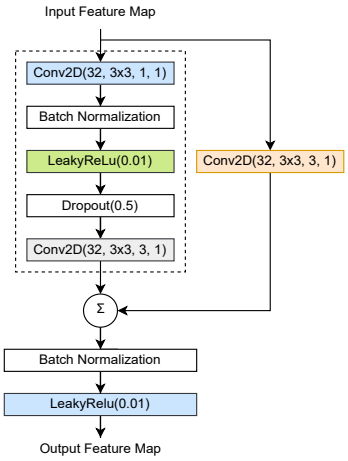

Fig. 3. Proposed Architecture of the residual Network Block of DRNN

III. Model Design

This study employs an efficient DRNN classifier for processing LPR source features. The DRNN framework is consistently applied across all input feature types, with minor adjustments to accommodate the dimensionality of each feature set. Figure 2 illustrates the overall DRNN architecture. Input features

TABLE IV
RESULTS OF PROPOSED AND THEIR FUSION MODELS USING 17PA AND 19PA DATASETS.

	17PA	19PA (Dev)		19PA (Eval)	
Models	EER	t-DCF	EER	t-DCF	EER
Log-Spectrogram-DRNN(s1)	18.10	0.1193	**4.28**	0.1588	**5.47**
RCQCC-DRNN(s2)	**14.71**	0.2580	11.35	0.4399	16.68
RMFCC-DRNN (s3)	21.03	0.4769	18.33	0.5619	20.67
RPCQCC-DRNN (s4)	14.84	0.3355	13.50	0.5352	19.34
s1+s2	11.70	**0.0906**	**3.72**	**0.1447**	**5.36**
s1+s3	14.33	0.1363	5.13	0.1597	5.81
s1+s4	11.09	0.1013	3.86	0.1506	5.42
s1+s2+s3	**10.32**	0.1156	4.66	0.1489	5.74
s1+s2+s3+s4	**8.40**	0.1115	**4.59**	0.1586	**6.06**

are treated as grayscale images and passed through a two-dimensional convolutional layer with 32 filters of size 3×3, using a stride and padding of 1. The output from this initial layer, comprising 32 channels, is then processed through nine residual blocks. Each residual unit comprises a 2D convolutional layer with 32 filters of size 3×3, employing a stride and padding of 1. This is followed by batch normalization, a Leaky ReLU activation function, and a dropout layer set at a 0.5 probability. A second Conv2D layer with 32 filters and a stride of 3 is then applied, along with batch normalization and dropout. A skip connection is incorporated to facilitate gradient flow and align the input and output dimensions via an additional Conv2D layer on the bypass path, as shown in Figure 3. This residual design helps address the gradient vanishing issue and enables the training deeper models, which has been shown to improve classification performance in various audio and speech tasks. After the residual blocks, a dropout layer with a probability of 0.5 is utilized, succeeded by a fully connected layer employing Leaky ReLU activation with a negative slope of $\alpha = 0.01$. The output is then passed to another fully connected layer with two nodes to produce raw classification scores. Finally, a softmax layer converts these scores into a probability distribution over the target classes. The DRNN is trained using the cross-entropy loss function to balance the weights assigned to both bonafide and spoofed speech segments. The model is trained using the Adam optimizer at a learning rate of 5×10^{-5}, for 100 epochs with a batch size of 32, aiming to minimize the loss function. During evaluation, the model parameters corresponding to the epoch with the highest validation performance are selected. The final countermeasure values are derived from the log-likelihood ratio computed using the softmax outputs, which are subsequently used to calculate the evaluation metrics.

A. Validation Metrics

In this research, system performance is evaluated using two standard metrics: the Equal Error Rate (EER) and the tandem Detection Cost Function (t-DCF). The EER corresponds to the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR), providing a single, balanced measure of accuracy. In contrast, the t-DCF serves as a cost-oriented evaluation framework that simultaneously considers the effectiveness of the ASV system and its integrated spoofing countermeasure [30], [31].

B. Database

The presented models are assessed using the ASVspoof 2017(17PA) and 2019 (19PA) datasets, each split into three subsets: training, development, and evaluation. The 17PA dataset focuses on realistic replay attack scenarios, while 19PA introduces additional complexity with varied devices and environments. Details of the speech samples in both datasets are provided in Table I.

IV. RESULTS AND DISCUSSION

This study presents models that integrate both spectral and LPR source-based features to enhance replay attack detection. The proposed approaches are evaluated on the 17PA and 19PA datasets, and their corresponding result analysis is presented below subsections.

A. Results on 17PA

The proposed models were evaluated on the 17PA dataset, which comprises both standard and realistic speech samples. The model was trained using a combination of the training and development datasets, and its performance was evaluated on the separate evaluation set. A summary of the results is provided in Table IV, while Table II offers a comparative analysis with recent approaches. In comparative evaluation, the log-spectrogram-DRNN (s1), RCQCC-DRNN (s2), RMFCC-DRNN (s3), and RPCQCC-DRNN (s4) models attained EERs of 18.10%, 14.71%, 21.03%, and 14.84%, respectively.. Notably, the s2 model, which is based on magnitude information, yielded the best individual performance. To further investigate the effectiveness of feature fusion, several combinations of spectral and source-based models (s1+s2, s1+s3, s1+s4, and s1+s2+s3) were evaluated using weighted sum score-level fusion. The s1+s2+s3 combination produced the best result among the fused models, while the full combination (s1+s2+s3+s4) achieved an EER of 8.40%, representing the most effective configuration for replay attack detection. Figure 4 represents the performances of proposed models in terms of EER. Furthermore, the proposed fusion model outperformed state-of-the-art approaches listed in Table II, highlighting the

significant contribution of source-based features to the detection of replay spoofing attacks.

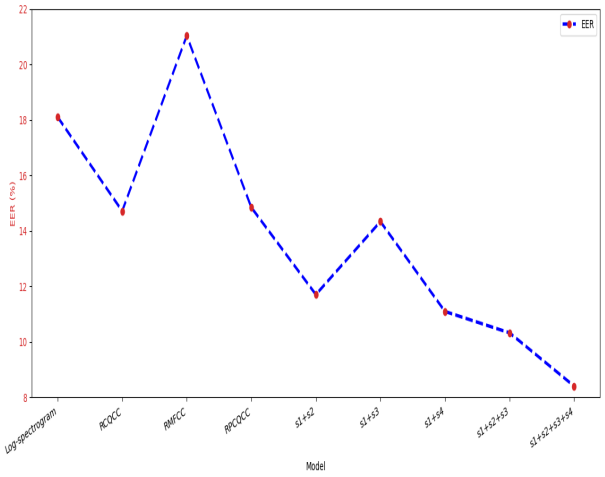

Fig. 4. Comparison of proposed models perfromances (Using 17PA)

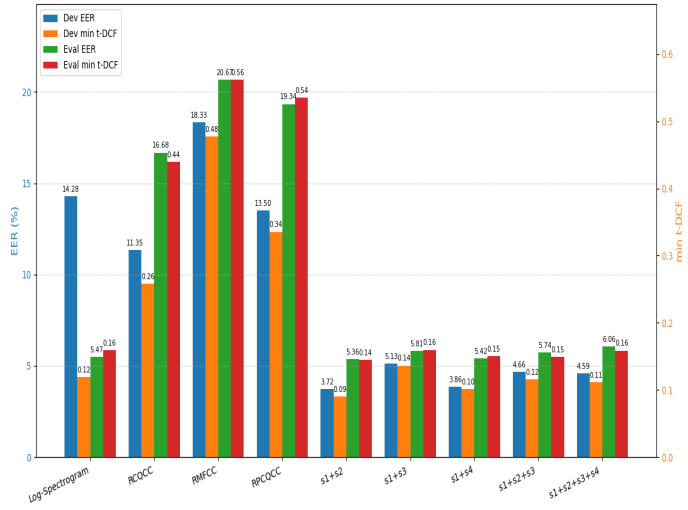

Fig. 5. Comparisons of proposed models performances (Using 19PA Development & Evaluation sets)

B. Results on 19PA

To further validate the effectiveness of the proposed models, evaluations were conducted using the 19PA dataset. The models were optimized using the training set and assessed on both the development and evaluation sets, with the t-DCF employed as the primary performance metric. The experimental results for both subsets are summarized in Table IV, while Table III presents a comparison with cutting-edge methods. Among the individual models, those based on spectral features, particularly the log-spectrogram model (s1) demonstrated superior performance, underscoring the effectiveness of log-spectrogram features when combined with deep learning techniques. Additionally, various fusion models that integrate spectral and residual source features (s1+s2, s1+s3, s1+s4, and s1+s2+s3+s4) were evaluated. Notably, the s1+s2 fusion model achieved the best results, with t-DCF scores of 0.1447 and 0.0906 on the evaluation and development sets, respectively, outperforming the comparative models listed in Table III. These findings further confirm that integrating LPR source features with spectral features enhances the robustness and accuracy of replay spoofing detection systems. Figure 5 represents the comparisons of proposed model perfromances in terms of both EER and t-DCF.

V. Conclusion

Replay spoofing poses a serious threat to ASV systems. The deployment of ASV in public spaces necessitates robust countermeasures against replay attacks. Existing countermeasure systems mostly have utilized speech based features, only limited studies have explored source features for identifying the replay attacks. In the present work, we have explored LPR based source features, RCQCC, RMFCC and RPCQCC along with log-spectrograms for countermeasure development task. The DRNN algorithm is used as classifier and exclusively configured for LPR source features for better performance. The proposed models log-spectrogram-DRNN, RCQCC-DRNN, RMFCC-DRNN and RPCQCC-DRNN are evaluated with both 17PA and 19PA datasets. Further, various fusion models are evaluated. Among these, the combined model (s1+s2+s3+s4) achieved an EER of 8.40% on the 17PA dataset, while the s1+s2 fusion model delivered the best results on the development and evaluation splits of the 19PA dataset. Hence, a comparative analysis shows that our proposed combination outperforms many latest existing deep learning methods. Thus, this study demonstrates the usefulness of excitation source information in the context of replay attack detection. Experimental results show that combining both spectral features and source information-based features significantly improves spoofing detection performance, owing to their complementary characteristics. Furthermore, the integration of excitation source features with convolutional neural networks utilizing residual blocks proves especially effective for detecting replay attacks. It is suggested that much more improvement may be achieved by exploring such excitation source features with alternative deep learning solutions from future perspectives.

Acknowledgment

This study is conducted at the SPIC Engineering Laboratory, NIT Nagaland, with financial support from the Chips to Startup (C2S) project, funded by the MeitY, New Delhi.

References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] M. Singh and D. Pati, "Linear prediction residual based short-term cepstral features for replay attacks detection," in *Proc. Interspeech 2018*, 2018, pp. 751–755.

[3] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: Review and analysis," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 105–134, 2022.

[4] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification.," in *Proc. Interspeech*, 2013, pp. 925–929.

[5] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[6] L. Huang, Y. Gan, and H. Ye, "Audio-replay attacks spoofing detection for automatic speaker verification system," in *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, 2019, pp. 392–396.

[7] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, 2025.

[8] J. Mishra, M. Singh, and D. Pati, "Processing linear prediction residual signal to counter replay attacks," in *2018 international conference on signal processing and communications (SPCOM)*, IEEE, 2018, pp. 95–99.

[9] M. R. Kamble and H. A. Patil, "Detection of replay spoof speech using teager energy feature cues," *Computer Speech & Language*, vol. 65, p. 101 140, 2021.

[10] M. Singh and D. Pati, "Countermeasures to replay attacks: A review," *IETE Technical Review*, vol. 37, no. 6, pp. 599–614, 2020.

[11] C. S. Gupta, "Significance of source features for speaker recognition," *Master's thesis, Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India*, 2003.

[12] J. Yamagishi, M. Todisco, M. Sahidullah, *et al.*, "Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *ASV Spoof*, vol. 13, 2019.

[13] A. Chaudhari, D. Shedge, V. Bairagi, and A. Nanthaamornphong, "Replay attack detection using integrated glottal excitation based group delay function and cepstral features," *Symmetry*, vol. 16, no. 7, p. 788, 2024.

[14] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[15] B. Wickramasinghe, E. Ambikairajah, V. Sethu, J. Epps, H. Li, and T. Dang, "Dnn controlled adaptive front-end for replay attack detection systems," *Speech Communication*, vol. 154, p. 102 973, 2023.

[16] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Revitalizing convolutional network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[17] E. Hassan, M. S. Hossain, A. Saber, S. Elmougy, A. Ghoneim, and G. Muhammad, "A quantum convolutional network and resnet (50)-based classification architecture for the mnist medical dataset," *Biomedical Signal Processing and Control*, vol. 87, p. 105 560, 2024.

[18] S. Veesa and M. Singh, "Deep learning countermeasures for detecting replay speech attacks: A review," *International Journal of Speech Technology*, pp. 1–13, 2024.

[19] P. Gupta, P. K. Chodingala, and H. A. Patil, "Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components," *Computer Speech & Language*, vol. 77, p. 101 423, 2023.

[20] J. Mishra, M. Singh, and D. Pati, "Exploring linear prediction residual signal for developing countermeasures to playback attacks," in *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, IEEE, 2018, pp. 1–6.

[21] K. Dutta, M. Singh, and D. Pati, "Improved processing of lp-residual information for detection of replay signals," in *2019 IEEE 16th India Council International Conference (INDICON)*, IEEE, 2019, pp. 1–4.

[22] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[23] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[24] M. Singh and D. Pati, "Usefulness of linear prediction residual for replay attack detection," *AEU-International Journal of Electronics and Communications*, vol. 110, p. 152 837, 2019.

[25] S. Siddhartha, J. Mishra, and S. M. Prasanna, "Language specific information from lp residual signal using linear sub band filters," in *2020 National Conference on Communications (NCC)*, IEEE, 2020, pp. 1–5.

[26] R. Sharma, D. Govind, J. Mishra, A. Dubey, K. Deepak, and S. Prasanna, "Milestones in speaker recognition," *Artificial Intelligence Review*, vol. 57, no. 3, p. 58, 2024.

[27] M. Singh and D. Pati, "Combining evidences from hilbert envelope and residual phase for detecting replay attacks," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 313–326, 2019.

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

[29] B. Saritha, M. A. Laskar, R. H. Laskar, M. Choudhury, *et al.*, "Cacrn-net: A 3d log mel spectrogram based channel attention convolutional recurrent neural network for few-shot speaker identification," *Computers and Electrical Engineering*, vol. 115, p. 109 100, 2024.

[30] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *in Proc. Interspeech*, pp. 2–6, 2017.

[31] ASVspoof 2019: "Automatic speaker verification spoofing and countermeasures challenge," [Online]. Available: http://www.asvspoof.org/.