Audio Deepfake Detection Using Linear Prediction Residual Cepstral Coefficients

Ritik Mahyavanshi[1], Aditya pusuluri,[1], Chandupatla Samyana Reddy[2], and Hemant A.Patil[1]

[1] Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DAIICT),
Gandhinagar, India.
{ritik_mahyavanshi, Aditya_ pusuluri , hemant_patil}@daiict.ac.in
[2] Department of Electronics and communication Engineering,
Koneru Lakshmaiah Education Foundation,
Hyderabad-500075, Telangana, India. 2210040032@klh.edu.in

Abstract. Deepfake audio synthesizes generate voices so advanced that they are almost impossible to identify as different or distinct than their real counterparts. Additionally, there are deep security and privacy risks of probably fooling voice recognition systems. In this context, we offer a novel audio deepfake detection technique of Order-Optimized Linear Frequency Residual Cepstral Coefficients (LFRCC). These LFRCC features are derived from the Linear Prediction residual in the cepstral domain, which captures small distortions from deepfake audio. We design an optimal LP order that maximizes the information content of the residual in this work, significantly enhancing the performance of detection.This work highlights that the deepfake characteristics are better captured in the fine spectral details and residual information that are often crucial for distinguishing subtle differences in audio signals in LFRCC.Compared to other conventional spectral feature sets, covering Linear Frequency Cepstral Coefficients (LFCC) and Mel Frequency Cepstral Coefficients (MFCC), our optimized LFRCC features significantly outperform others features in terms of accuracy and reliability for deepfake detection.The LFRCC features gave a best Equal Error Rate (EER) of 2.36 % and accuracy of 98.68 % for an optimal LP order 8 as compared to MFCC which gave 67.67 % accuracy and 15.30 % EER and LFCC gave 92.25 % accuracy and 7.62 % EER using TDNN classifier analysis of various dimensions of feature vectors.In addition, we performed experiments for babble noise at various levels.The statistical significance is evaluated using Equal Error Rate (EER) and F1-score Further, results are shown for the cross-database scenarios.We also performed experiments to compare with deep learning models baseline like whisper and Wav2vec2. Finally,the analysis of latency period with baseline feature sets suggests the potential applicability of LFRCC for deepfake detection.

Keywords: deepfake audio · Audio Deepfake Detection (ADD)· LFRCC · MFCC· LP residual· cepstral domain· MFCC· Whisper · Wav2vec2.

1 Introduction

The emergence of deepfake audio, driven by advanced machine learning algorithms, has introduced significant challenges in the field of audio manipulation and generation. Deepfake audio refers to the artificial synthesis of voices that closely mimic authentic human speech [1]. Initially used for entertainment, the applications of this technology have expanded, raising serious concerns about potential misuse. Malicious actors can exploit deepfake audio to create deceptive content, manipulate public opinion, or commit fraud by imitating trusted voices [2]. Consequently, developing robust methods for detecting and analyzing deepfake audio has become crucial [3].

Early approaches in the field of audio forensics and machine learning have explored various techniques for detecting deepfake content [3]. The eariler studies have highlighted the potential of cepstral features in identifying anomalies in audio recordings, including those generated by deepfake algorithms. For instance, MFCC have been widely adopted for their effectiveness in capturing the spectral characteristics of speech signals, particularly in speech recognition tasks [4]. Studies such as [5] have demonstrated the effectiveness of MFCC in detecting manipulated speech.

In addition to MFCC, LFCC have been proposed as they provide a more straightforward representation of the underlying linear filter characteristics of speech signals, improving robustness in noisy environments [6]. Methods based on LFCC for deepfake detection have been explored in [7]. Linear Prediction Cepstral Coefficients (LPCC) aim to capture the residual information using LP analysis, thereby enhancing the representation of speech signals by focusing on the excitation source component [8]. Despite these advancements, the specific application of LFRCCs in the context of deepfake audio analysis remains relatively unexplored [9].

In this paper, we address deepfake audio detection by introducing order-optimized LFRCCs, which combine the benefits of LFCC and LPCC by capturing both linear filter characteristics and residual excitation information [10]. This method enhances the representation of speech signals across multiple frequency bands, improving the discrimination between real *vs* fake audio. Our research builds on traditional methods like LFCC and MFCC, exploring the efficacy of LFRCCs in audio forensics and machine learning [11]. We provide a detailed technical description of LFRCC optimization, emphasizing the LP order to enhance discriminative power, and outline the experimental setup, including datasets, pre-processing, and parameters. The progression from MFCC to LFRCC reflects a refinement in capturing nuanced speech details, crucial for distinguishing authentic from synthetic audio [12]. This study aims to fill this gap in the literature w.r.t application of LFRCC for deepfake audio analysis, offering valuable insights into their utility for ADD task [13–16, 9, 6].

Additionally, we compared the performance of LFRCC with advanced deep learning models, such as Whisper Tiny, Whisper Base, Wav2Vec2.0 Large, and Wav2Vec2.0 Base. These comparisons allowed us to assess the relative effectiveness of LFRCCs against state-of-the-art deep learning approaches in deepfake

detection. By evaluating LFRCCs alongside these models, we aim to demonstrate the potential advantages and limitations of our proposed method in capturing and distinguishing between real *vs* fake audio content.

A key aspect of our approach involves the careful optimization of the LP order to modulate the information content within the LP residual signal. Through this process, we aim to enhance the discriminative power of the LFRCC, thereby improving the performance of deepfake detection algorithms.

The rest of the paper is organized as follows: Section 2 represents brief technical details of proposed LFRCC feature set, whereas Section 3 gives the details of experimental setup used for performance evaluation of LFRCC. Section 4 presents results using LFRCC w.r.t various LP order, comparison with existing feature sets (i.e., MFCC and LFCC), and robustness against signal degradation conditions for various Signal-to-Noise (SNR) levels of additive babble noise. Finally, Section 5 concludes the paper along with potential future research directions.

2 Proposed Approach

Linear Prediction (LP) analysis is a method, where a speech sample at the n^{th} time instant, $s(n)$, is approximated as a linear combination of the past p speech samples [5]. Motivated by it's succes in system identification literature, LP analysis is widely used in speech coding, estimating glottal closure instants (GCIs), and the fundamental frequency (F_0) [11]. The method separates the excitation source from the vocal tract system information, which are crucial for speech production [17].

In LP analysis, each speech signal sample is represented as a linear combination of the previous p samples, with p being the order of linear prediction [5, **?**]. The linear combinations are associated with weight parameters called Linear Prediction Coefficients (LPCs). The predicted speech sample, $\tilde{s}(n)$, is given by:

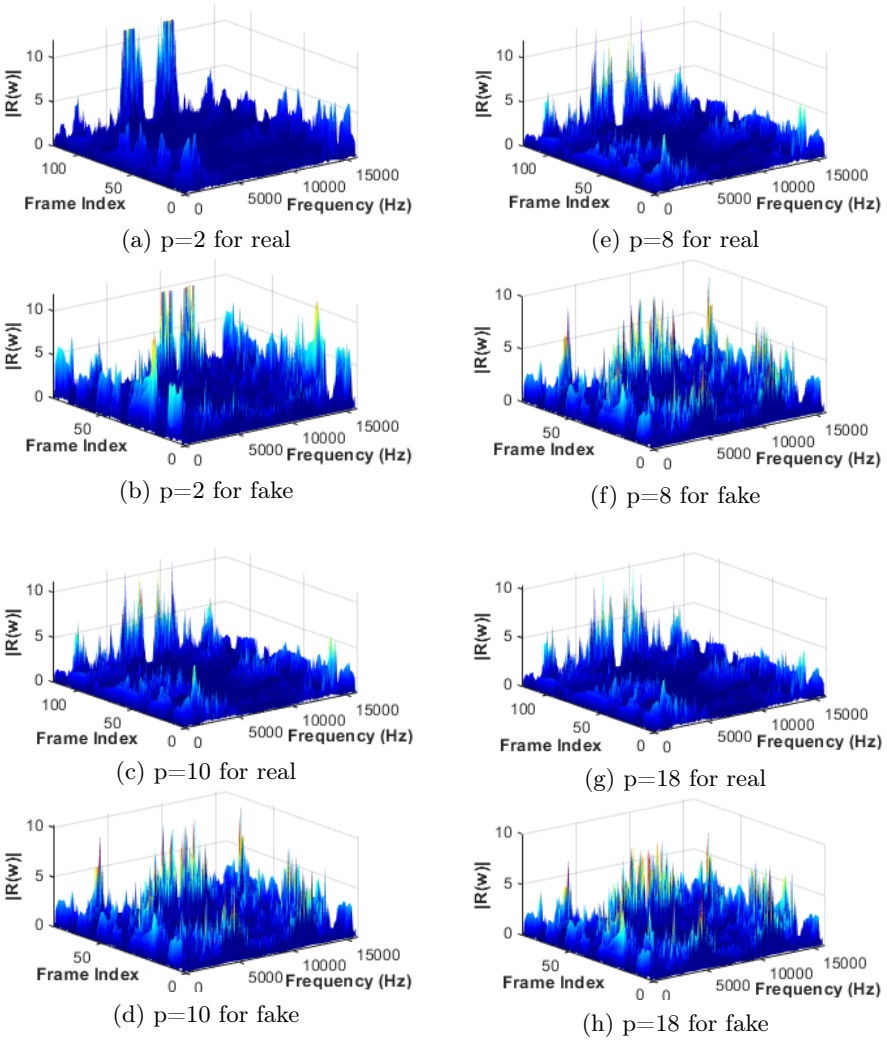
$$\tilde{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k), \tag{1}$$

where α_k are the LPC. The prediction error, known as the LP residual, carries the excitation source component of the speech signal and is defined as [5]:

$$r(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^{p} \alpha_k s(n-k). \tag{2}$$

In LP analysis, an all-pole filter is applied to the speech signal, represented by:

$$F(z) = 1 + \sum_{k=1}^{p} \alpha_k z^{-k}, \tag{3}$$



(a) p=2 for real (e) p=8 for real

(b) p=2 for fake (f) p=8 for fake

(c) p=10 for real (g) p=18 for real

(d) p=10 for fake (h) p=18 for fake

Fig. 1: Plots of framewise magnitude spectrum of LP residual, i.e., $|R(\omega)|$ of the LP residual of real *vs.* fake speech, for different LP order p.

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{4}$$

where $F(z)$ is the inverse filter corresponding to the all-pole LP filter $H(z)$, and G is the gain term in the LP model [5]. The LP residual spectrum captures the excitation source information by filtering out the system information, with peaks and valleys indicating the GCIs and Glottal Open Instances (GOIs) during speech production mechanism [18].

Optimizing the LP order is crucial to adjust the information content within the LP residual signal, enhancing the discriminative power of Linear Frequency Residual Cepstral Coefficients (LFRCCs) and improving the performance of deepfake detection algorithms [5].

2.1 Analysis of the LP Residual and LFRCC

LP Residual for ADD : The LP residual is known to capture the speakers' special characteristics from the speech signal. The LP spectrum is expected to represent the reinforcement of spectral energy in the vocal tract system in the form of resonances, and the nature of the excitation source is represented by the pitch (x) and its harmonics in the form of LP residual [19]. However, source characteristics are captured aptly when the linear prediction is done with *minimized* L_2 norm of LP residual, which further leads to the optimal LP order as $(f_s/1000) + 2$, where f_s is the sampling frequency of the speech signal [20]. Therefore, for a speech with 16 kHz sampling frequency, the signified optimal order for correct prediction is around 18. Note that this order makes sense only if we are interested in accurate prediction of speech samples. In the case of the ADD task, though, this is not the case for using LP residual information; our goal for the ADD task is to achieve optimally classified real vs. deepfake. To this effect, the amount of information that the LP order will impose varies, and thus, the LP residual will vary with the LP order. We have discussed, in this sub-section, the effect of varying LP orders on the LP residual.

Specifically, Fig. 1 shows the framewise magnitude envelope of the LP residual for real *vs.* fake speech for a range of LP orderings. It was observed that the difference between the magnitudes of the spectra for real *vs.* fake speech is maximum when $p = 8$.

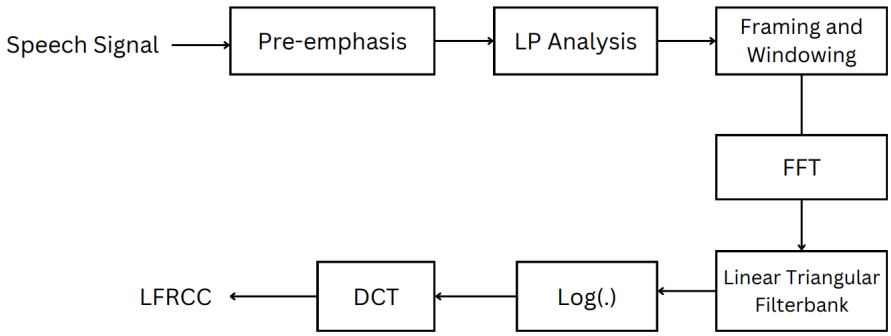
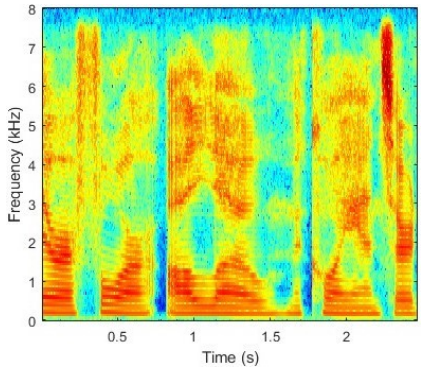
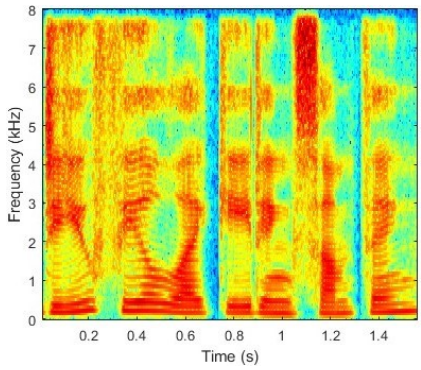


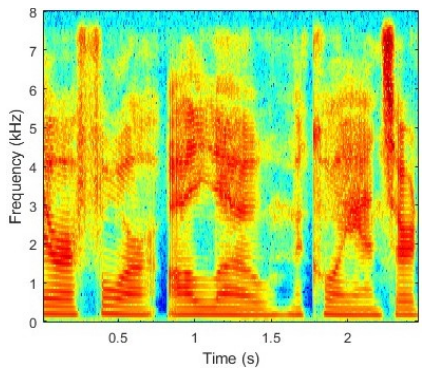
Fig. 2: Functional block diagram of LFRCC feature extraction [9].



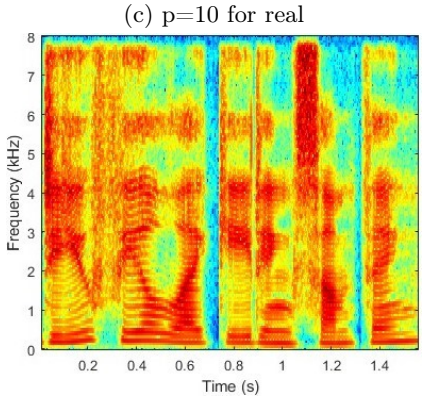
(a) p=2 for real



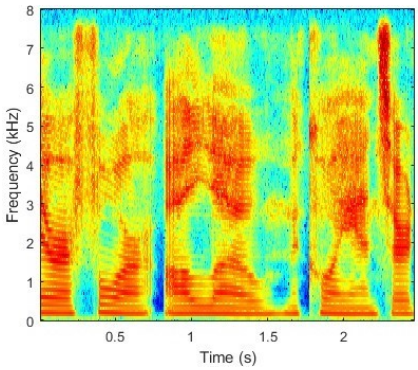
(b) p=2 for fake



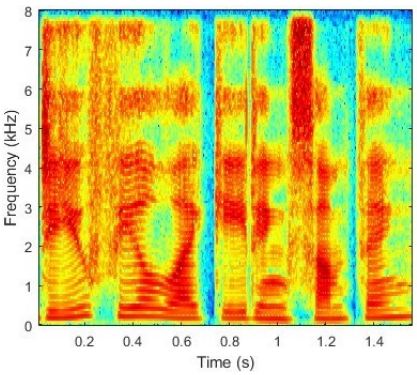
(c) p=10 for real



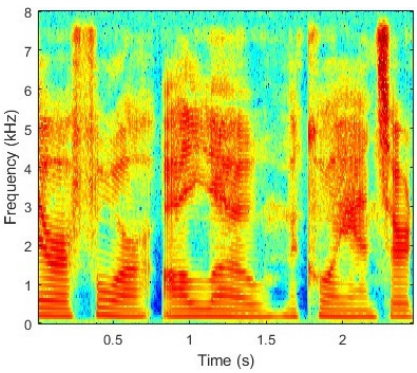
(d) p=10 for fake



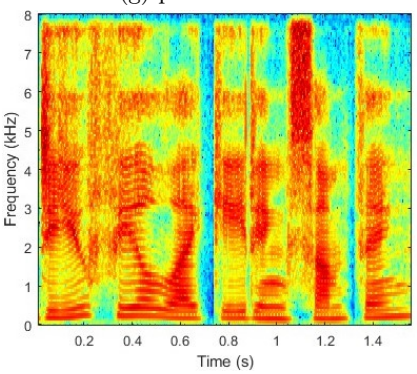
(e) p=8 for real



(f) p=8 for fake



(g) p=18 for real



(h) p=18 for fake

Fig. 3: Spectrogram plots for different LP orders of real vs. fake speech.

LFRCC Feature Extraction :

LFRCC features from LP residual like performing pre-emphasis (highpass) filter is applied to the speech signal to emphsize its high frequency components. The features extracted from high frequency regions carry more significance for the task of Spoofed Speech Detection (SSD) [16], particularly in the context of discerning between genuine *vs* replayed speech. The speech signal, after undergoing pre-emphasis, is further processed through a lowpass analysis block. This stage aims to extract LP residual waveform from the pre-emphasized speech.Window function divides the signal into short frames and apply window function,these processes enable a more localized analysis of the signal over time.[9] The use of speech frames makes it easier to capture variations in pitch (f_0), spectral content, and other characteristics that, may involve with shorter durations. Following the framing, windowing, and obtaining the LP residual, the next step involves computing the power spectrum for each of these LP residual frames [21]. Additionally, framing and windowing contribute to the efficiency of spectral analysis techniques like the Fast Fourier Transform (FFT). Breaking the signal into frames allows FFT to be applied more efficiently to each speech segment, reducing the computational cost compared to analyzing the entire signal at once. The power spectrum obtained from the LP residual frames is subjected to further processing by passing it through linearly-spaced triangular filterbank [9].This filterbank act like bank of frequency-specific subband filters that help to capture the energy distribution across different frequency bands. The bandwidth of each triangular subband filter is also equally distributed. This linear spacing ensures that the filters consistently cover the entire frequency spectrum. To reduce correlations within the feature set, a Discrete Cosine Transform (DCT) is employed to compute a low-dimensional representation of the features. The DCT helps transform the original feature set into a new set of coefficients, emphasizing the most important information while minimizing redundancy [7].Following the DCT, only a few initial cepstral coefficients are retained. This reduction in dimensionality helps in simplifying the representation while preserving essential information. In sum, DCT is applied for feature vector dimensionality reduction, energy compaction, and feature decorrelation [16].

3 Experimental Setup

3.1 Dataset used

For the experiments, we utilized the standard and statistically meaningful Fake-or-Real (FoR) dataset [15]. This dataset contains more than 195,000 utterances from both real humans and computer-generated speech. It's useful for training classifiers to detect synthetic speech. The FoR dataset includes over 87,000 synthetic utterances and more than 111,000 real utterances, making it valuable for research in synthetic speech detection [21]. There are four versions of this dataset in total, namely, For-original, For-norm, For-2sec, and For-rerec. In this work, we specifically employed the "for-norm" version of the dataset, which balances gender and class distribution and normalizes aspects, such as sampling rate,

volume, and number of channels. The "for-norm" version of the FoR dataset ensures consistency in acoustic properties across both synthetic and real utterances, thereby mitigating biases and enhancing generalization capabilities of our deepfake detection models [22]. We also utilized the standard and statistically meaningful ASVspoof 2021 dataset and FOR-original for cross-database analysis [4]. The FoR-norm dataset served as a benchmark for rigorously evaluating our trained deepfake detection models, ensuring their robustness and effectiveness in diverse real-world scenarios. This dataset contains more than 218,000 utterances from both real humans and computer-generated speech [23]. It's useful for training classifiers to detect synthetic speech. The ASVSpoof 2021 dataset includes over 158,000 synthetic utterances and more than 60,000 real utterances, making it valuable for research in synthetic speech detection. There are three main subsets of this dataset: the Logical Access (LA) subset, the Physical Access (PA) subset, and the DeepFake (DF) subset [24]. In this work, we specifically employed the LA subset, which provides a comprehensive evaluation of logical access attacks, and includes balanced class distribution and normalized aspects such as sampling rate, volume, and number of channels [23].

The inclusion of the ASVspoof 2021 dataset, particularly the LA subset, facilitated a comprehensive cross-database analysis, validating our models' efficacy across different attack scenarios and enhancing their applicability in diverse security contexts.

Table 1: Distribution of Samples in FOR-NORM Dataset

	Fake	Real
Training	26,927	26,940
Validation	5,398	5,400
Testing	2,370	2,264

Table 2: Distribution of Samples in ASVspoof 2021 Dataset

	Fake	Real
Training	20,478	25,380
Validation	6,059	7,355
Testing	2,25,000	2,35,000

3.2 BaseLine Features

The MFCC and LFCC features were used as baselines for comparison with the proposed LFRCC. To construct a comprehensive feature vector, we incorporated static (original), delta (first-order derivative), and double-delta (second-order derivative) parameters [24], capturing both *spectral* properties and *temporal* dynamics of the audio signal. For feature extraction, we used a 30 ms window length, 40 subband filters, and a *512*-point Fast Fourier Transform (NFFT) for fine-grained frequency resolution [14]. The linear prediction (LP) order was varied from 2 to 30, with an optimal order of 8 yielding the best EER results compared to MFCC and LFCC. The dimension of the LP order 8 feature was varied from 10 to 40 for static, delta, and double-delta parameters [24]. Additionally, we compared these features against baseline models, including Whisper Tiny, Whisper Base, Wav2vec2 Large, and Wav2vec2 Base, to evaluate the robustness and accuracy of LFRCC in capturing speaker-specific and residual information crucial for detecting audio deepfakes.

3.3 Classifier

For the classification task of audio deepfakes, a Time-Delay Neural Network (TDNN) was employed. TDNNs are specifically designed for processing sequential data, such as speech signals, excelling in capturing temporal dependencies and patterns in sequences such as LFRCC and LP residual. LFRCC compactly represent spectral properties of residual signals derived from LPC, highlighting critical frequency components for speech processing tasks. Meanwhile, LP residuals capture essential excitation source information, crucial for distinguishing voiced and unvoiced sounds.

Compared to architectures like Whisper960 tiny, Whisper960 base, CNNs, Wav2Vec large, Wav2Vec2 base, and RawNet2, TDNNs are preferred for their ability to model temporal relationships effectively across sequential data. TDNNs incorporate layers designed with delays to capture time dependencies comprehensively, making them well-suited for tasks, where understanding temporal context in speech is vital.

The TDNN architecture used for the classification task of audio deepfakes consisted of layers structured to process sequential data, with specific configurations optimized for temporal feature extraction and classification. We used a batch size of 32 and trained the network over 15 epochs, with a maximum latency period of 600 ms, to enhance its capability in learning sequential patterns and temporal nuances in audio data. This approach ensured robust performance in distinguishing authentic from manipulated audio clips.

3.4 Performance Metrics Used

Equal Error Rate (EER) : In evaluating the effectiveness of our proposed LFRCC for ADD, we employ the EER as the primary performance metric. EER is a widely recognized measure within the field of biometric authentication and pattern recognition, offering a balanced assessment of the system's accuracy by determining the point at which the rates of false acceptance and false rejection are equal. This metric serves as a pivotal criterion for evaluating the robustness and reliability of our system in distinguishing between genuine and impostor attempts [25]. The estimation of EER involves plotting the Receiver Operating Characteristics (ROC) curve and identifying the threshold, where the false acceptance rate (FAR) equals the false rejection rate (FRR).In paticular,

$$EER = \frac{FAR + FRR}{2}. \tag{5}$$

The EER value represents the error rate at this intersection point, providing a brief summary of the system's performance without favoring either type of error. Interpretation of the EER entails striving for a lower value, signifying a more accurate and dependable authentication system [25].

F1-score : The F1-score is a metric used to evaluate the performance of a classification model, particularly when dealing with imbalanced datasets. It is

the harmonic mean of precision and recall, providing a balance between the two. Precision measures the accuracy of positive predictions, defined as the ratio of true positive predictions to the total number of positive predictions made. Recall, on the other hand, measures the ability of the model to identify all relevant instances, defined as the ratio of true positive predictions to the total number of actual positive instances [25]. The F1-score combines these two metrics into a single score, calculated as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

This score is particularly useful when the cost of false positives and false negatives is high, and it provides a more comprehensive measure of model performance than accuracy alone [25].

4 Experimental Results

This section evaluates proposed LFRCC feature set first by doing LP order optimization w.r.t ADD performance,comparison with existing cepstral feature sets (i.e., MFCC and LFCC), and robustness under signal degradation conditions.We performed cross-dataset performance, which involves evaluating a model trained on one dataset against another dataset to assess its generalization and robustness and to check data compatibility. To show the robustness to noise of our model, nine different experiments were conducted for babble noise.

4.1 Optimization of LP Order

We conducted a series of experiments to enhance the performance of a TDNN classifier on detecting fake *vs* real utterances using features extracted via LFRCC. To achieve this, we varied the LP order from 2 to 30 in static vectors, with a window length of 30ms, for static 20 coefficients, and an NFFT (Number of points in FFT) of 512. We evaluated the system's performance based on EER, accuracy, and F1-score.

After thorough experimentation, we found that an LP order of 8 yielded the most impressive results. The system demonstrated exceptional performance with an impressive EER of 5.33 %, highlighting its effectiveness in reducing both false positives and false negatives. Additionally, the classifier demonstrated remarkable accuracy, reaching 98.63 %, and an impressive F1-score of 98.61 %, highlighting its balanced precision and recall.

These results underscore the significance of selecting an optimal LP order for feature extraction in enhancing the robustness of TDNN classifiers for fake speech detection.

4.2 Optimization of Feature Dimension

We conducted a series of experiments to enhance the performance of a TDNN classifier for detecting fake *vs* real utterances using features extracted via LFRCC.

We experimented with different dimensional of static feature vectors, initially varying the coefficients from 10 to 40 in static vectors. In this setup, we achieved the best EER of 5.33 % with 12 coefficients.

We then extended our experiments to include static, delta, and double delta vectors. This approach yielded even more impressive results, with an EER of 2.35 %, an accuracy of 97.65 %, and an F1 score of 97.62 % for 12 coefficients.

Based on these findings, we determined that the optimal dimension vector for our reasearch is the combination of static, delta, and double-delta features with 12 coefficients. This configuration significantly enhances the classifier's ability to accurately distinguish between fake *vs* real utterances.

4.3 Performance Comparison with Existing Feature Sets

We compared the effectiveness of three feature extraction techniques, namely LFCC, MFCC, and LFRCC for detecting real *vs* fake audio samples. LFRCC emerged as the superior method, achieving an impressive EER of 2.35 %, an accuracy of 97.65 %, and an F1 score of 97.62 %. In contrast, LFCC yielded a higher EER of 7.62 %, with an accuracy of 92.25 % and an F1 score of 92.26 %. Similarly, MFCC produced an EER of 15.30 %, an accuracy of 67.67 %, and an F1-score of 67.69 %. These results underscore the superior performance of LFRCC over LFCC and MFCC in distinguishing between real *vs* synthetic speech. Our experiments also highlighted significant performance variations among different architectures: the whisper960 tiny model achieved an EER of 6.38 %, while the wav2Vec2 base model showed a significantly higher EER of 32.55 %. The wav2Vec2 large model outperformed the base version with an EER of 19.03 %, and the Whisper960 base model achieved an EER of 5.87 %.

4.4 Results of Cross-dataset analysis

We conducted a cross-dataset analysis using the FoR-original and ASVspoof 2021 datasets to evaluate the robustness and generalizability of our audio deepfake detection model. The results for the FoR-original dataset were particularly insightful, revealing an accuracy of 61.73 %, an F1 score of 67.72 %, and an EER of 37.40 %.

In contrast, our experiments with the FoR-norm dataset, which we previously used for training and evaluation, demonstrated significantly better performance with an EER of just 2.36 %. Additionally, the results for the ASVspoof 2021 dataset showed an accuracy of 62.63 %, an F1 score of 51.11 %, and an EER of 58.13 %. These findings underscore the challenges and complexities of cross-dataset evaluation, where models must adapt to diverse data characteristics and nuances. Despite these hurdles, our analysis offers valuable insights into the model's performance across different datasets, highlighting the importance of continuous refinement and adaptation to achieve optimal results in various audio environments.

Table 3: cross-dataset analysis

Dataset Used	Accuracy	F-score	EER
FOR-orignal	61.73%	67.72%	37.40%
ASVspoof 2021	62.63%	57.11%	58.14%
FOR-NORM	97.65%	97.82%	2.35%

4.5 Effect of Additive Babble Noise

We conducted a series of experiments to evaluate the robustness of LFCC, LFRCC, and MFCC feature extraction techniques under various levels of babble ranging from -20 dB to 20 dB. The goal was to determine how well each method performs in noisy environments, a critical factor for real-world applications.

Table 4: Comparison of noise levels and their effects on LRCC, MFCC, and LFCC

Babble Noise			
SNR level (dB)	LRCC	MFCC	LFCC
-20 dB	24.53	34.94	33.44
-15 dB	38.00	40.46	30.16
-10 dB	16.71	34.60	28.97
-5dB	9.84	33.72	21.34
0 dB	11.50	25.39	6.52
5 dB	6.73	24.19	6.57
10 dB	6.44	19.40	6.98
15 dB	7.87	16.56	12.23
20 dB	3.82	36.36	10.44

These findings highlight that LFRCC consistently outperforms LFCC and MFCC in noisy conditions, particularly at higher signal-to-noise ratio (SNR). This robustness makes LFRCC an excellent choice for applications in challenging auditory environments, enhancing the reliability of speech-based systems under adverse conditions practical.

4.6 Analysis ofLatency period

We conducted a series of experiments, and varied the latency period from 50 ms to 600 ms to assess its impact on the performance of our audio deepfake detection model. The results were particularly compelling for the LFRCC feature extraction method, which outperformed the baseline across all the latency periods tested. By systematically adjusting the latency period, we observed that LFRCC consistently delivered superior performance, highlighting its robustness

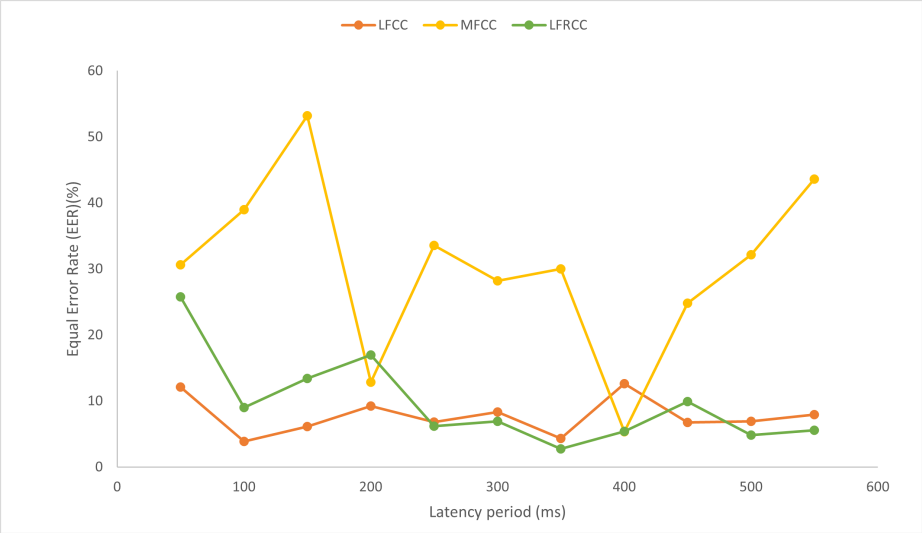


Fig. 4: Analysis of Latency period in comparision to baseline

and adaptability in capturing critical temporal features essential for distinguishing real from the fake audio. These findings not only reinforce the effectiveness of LFRCC but also underscore the importance of optimizing latency to enhance the detection capabilities of audio deepfake classifiers.

5 Summary and Conclusions

In the evolving domain of deepfake audio detection, our study introduces a cutting-edge method leveraging Order-Optimized LFRCC. These features, derived from the LP residual in the cepstral-domain, effectively capture the subtle distortions that characterize deepfake audio, making them more distinguishable from genuine counterparts. Our optimized LFRCC features achieved reduced EER of 2.35 % and an accuracy of 97.65 % using the TDNN classifier, significantly outperforming conventional spectral features, such as MFCC (67.67 % accuracy, 15.30 % EER) and LFCC (92.25 % accuracy, 7.62 % EER). Robust cross-dataset analysis further demonstrated LFRCC's superior performance, with accuracy and F1 scores of 97.65 % and an EER of 2.35 % on the FOR-Norm dataset. Additionally, LFRCC features proved resilient in various noise conditions and across different latency periods, consistently outperforming LFCC and MFCC in terms of EER. Among the deep learning models tested, TDNN combined with LFRCC showed the best results, confirming its robustness and reliability. In summary, this research highlights the significant advancement that LFRCC features bring to deepfake audio detection, offering enhanced accuracy, reduced error rates, and robust performance, thereby presenting a promising tool for mitigating the security and privacy risks posed by sophisticated audio deepfakes.

6 Acknowledgements

The authors would also like to thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for partial support from a consortium project titled 'Speech Technologies in Indian Languages' under 'National Language Translation Mission (NLTM): BHASHINI', subtitling 'Building Assistive Speech Technologies for the Challenged' (Grant ID: 11(1)2022-HCC(TDIL)). We also thank the authorities of DA-IICT Gandhinagar, India and Dr.Priyanka Gupta (LNMIT, jaipur) for their kind support and cooperation to carry out this research work.

References

1. S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting deepfake voice using explainable deep learning techniques," *Applied Sciences*, vol. 12, no. 8, 2022.
2. Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, 2022.
3. Z. Khanjani, G. Watson, and V. P. Janeja, "How deep are the fakes? focusing on audio deepfake: A survey," *arXiv preprint arXiv:2111.14203*, 2021 {Last Accessed date : 4^{th} January, 2024}.
4. B. Zhang, J. P. Zhou, I. Shumailov, and N. Papernot, "On attribution of deepfakes," *arXiv preprint arXiv:2008.09194*, 2020 Last Accesed : January 2024.
5. J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
6. G. Fant, *Speech Acoustics and Phonetics: Selected Writings*. Springer Science & Business Media, 2004, vol. 24.
7. Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, and V. Vimal, "Deepfake generation and detection: Case study and challenges," *IEEE Access*, vol. 11, pp. 143 296–143 323, 2023,2024.
8. N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition," *International Journal for Advance Research in Engineering and Technology*, vol. 1, no. 6, pp. 1–4, 2013.
9. H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection," in *INTERSPEECH*, 2018, Hyderabad, India, pp. 726–730.
10. S. P. Dewi, A. L. Prasasti, and B. Irawan, "Analysis of lfcc feature extraction in baby crying classification using knn," in *IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), Hong Kong*, 2019, pp. 86–91.
11. R. S. Strichartz, "Lp harmonic analysis and radon transforms on the heisenberg group," *Journal of Functional Analysis*, vol. 96, no. 2, pp. 350–406, 1991.
12. S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deep-fake videos from appearance and behavior," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6.
13. C. Hanilçi, "Speaker verification anti-spoofing using linear prediction residual phase features," in *25^{th} European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, August 8 - September 2, 2017, pp. 96–100.
14. E. Jokinen, R. Saeidi, T. Kinnunen, and P. Alku, "Vocal effort compensation for mfcc feature extraction in a shouted versus normal speaker recognition task," *Computer Speech & Language*, vol. 53, pp. 1–11, 2019.

15. A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.
16. B. S. Hora, S. Uthiraa, and H. A. Patil, "Linear frequency residual cepstral coefficients for speech emotion recognition," in *International Conference on Speech and Computer,*. Springer, 2023, pp. 116–129.
17. J. J. Albers, H. Kennedy, and S. M. Marcovina, "Evidence that lp [a] contains one molecule of apo [a] and one molecule of apob: evaluation of amino acid analysis data." *Journal of Lipid Research*, vol. 37, no. 1, pp. 192–196, 1996.
18. T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2011.
19. S. Cheedella S. Gupta and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.
20. B. S. Atal, "Automatic speaker recognition based on pitch contours," *The Journal of the Acoustical Society of America (JASA)*, vol. 52, no. 6B, pp. 1687–1697, 1972.
21. P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 7-10 December 2020, Auckland, New Zealand, pp. 543–549.
22. T. Freitas dos Santos, N. Osman, M. w. p. a. t. s. I. C. o. A. A. Schorlemmer, and M. S. A. in, "Ensemble and incremental learning for norm violation detection," 2022.
23. J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVSpoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.
24. C. Hanilçi and F. Ertas, "Optimizing acoustic features for source cell-phone recognition using speech signals," in *Proceedings of the first ACM workshop on Information Hiding and Multimedia Security*, 2013, pp. 141–148.
25. K. Emerson and T. Nabatchi, "Evaluating the productivity of collaborative governance regimes: A performance matrix," *Public Performance & Management Review*, vol. 38, no. 4, pp. 717–747, 2015.