2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) | 979-8-3315-0603-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/O-COCOSDA64382.2024.10800195

Analysis of Pathological Features for Spoof Detection

Myat Aye Aye Aung
Faculty of Computer Science
University of Computer Studies,Yangon
Yangon, Myanmar
myatayeayeaung@ucsy.edu.mm

Hay Mar Soe Naing
Faculty of Computer Science
University of Computer Studies,Yangon
Yangon, Myanmar
haymarsoenaing@ucsy.edu.mm

Aye Mya Hlaing
Faculty of Computer Science
University of Computer Stydies,Yangon
Yangon, Myanmar
ayemyahlaing@ucsy.edu.mm

Win Pa Pa
Faculty of Computer Science
University of Computer Studies,Yangon
Yangon, Myanmar
winpapa@ucsy.edu.mm

Kasorn Galajit
NECTEC, National Science and Technology Development Agency
Pathum Thani, Thailand
kasorn.galajit@nectec.or.th

Candy Olivia Mawalim
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa 923–1292 Japan
candylim@jaist.ac.jp

***Abstract*—Deepfake speech, an advanced use of speech synthesis technology, presents a considerable challenge due to its highly realistic sound and the complexities involved in detecting it. The selection and analysis of effective features are essential for enhancing spoof detection capabilities. This study focuses on analyzing pathological features within the Myanmar Spoof Dataset. Spoofed speech in the dataset is created using five distinct techniques: vocoder methods with HiFiGAN and Parallel WaveGAN, pre-trained voice conversion via FreeVC, and GMM-based and Differential GMM-based voice conversion (GMMVC_DIFF) methods. In this paper, we perform a comparative analysis of pathological features, including Harmonics-to-Noise Ratio (HNR), six types of jitter, and seven shimmer features. Additionally, Cepstral Peak Prominence (CPP) features were assessed under both voiced and unvoiced speech conditions. The analysis demonstrates that these features show substantial variations across different spoofing techniques. Specifically, the voice conversion methods GMMVC_DIFF show notable differences in these features. These results highlight the pivotal role of these pathological features in enhancing the precision of future spoof detection systems.**

Keywords—spoof detection, speech synthesis, Myanmar spoof dataset, pathological features, cepstral peak prominence features

I. INTRODUCTION

Deepfake speech involves the unethical use of advanced speech technologies, such as voice conversion and text-to-speech methods, to create synthetic and deceptive audio content [1-3]. Due to its highly realistic nature and the significant challenges associated with detection, deepfake speech poses a substantial threat to economic systems and societal stability. Fake Audio Detection (FAD) technologies are essential in protecting Automatic Speaker Verification (ASV) systems from spoofing attacks, including voice conversion, replay, and text-to-speech. These technologies identify and filter out synthetic or manipulated audio that could compromise ASV integrity. In FAD, the analysis of pathological features and CPP is crucial, alongside advancements in deep learning and acoustic feature extraction.

In the literature, acoustic features such as mel-frequency cepstrum coefficients (MFCC), modified group delay function, and cos-phase have been key in anti-spoofing efforts [4]. LFCC features [5] provide a linear frequency perspective, while constant-Q cepstral coefficients (CQCC) [6] offer detailed frequency analysis, especially in the low-frequency range. The authors [7] proposed using pathological features to analyze deepfake speech, suggesting that these features could reveal similarities to disordered voices. This paper focused on six key features: jitter, shimmer, HNR, Cepstral Harmonics-to-Noise Ratio (CHNR), Normalized Noise Energy (NNE), and Glottal-to-Noise Excitation Ratio (GNE).

We investigated fourteen pathological features, including six jitter metrics, seven shimmer metrics, HNR, and two CPP features, considering both voiced and unvoiced speech in the Myanmar spoof dataset. These features were selected for their ability to capture subtle variations in pitch, amplitude, and noise characteristics, indicative of synthetic manipulation in deepfake speech. By conducting a comparative analysis, we aimed to identify the most effective features for detecting nuanced alterations in speech patterns, which is crucial for developing robust countermeasures against deepfake speech and protecting the integrity of ASV systems. This paper is part of the ASEAN IVO 2023 project, "Spoof Detection for Automatic Speaker Verification," which aims to apply reliability of speaker verification using the new UCSY Spoof dataset and significant features for the Myanmar language.

The contributions of this study can be outlined as follows:

1. Comprehensive analysis that performed an in-depth evaluation of fourteen pathological features.
2. Assessment of CPP features in detecting subtle alterations in speech patterns.
3. Identification of significant feature variations: discovered notable variations across different spoofing techniques, enhancing the precision of spoof detection systems.

The organization of this paper is as follows: Section 2 provides a comprehensive overview of the five different Myanmar Spoof Dataset, collectively named UCSYSpoof. Section 3 details the system design employed for the feature analysis process. Section 4 discusses the implementation of pathological features and CPP features. Section 5 presents and analyzes the results obtained from these features, accompanied by relevant discussions. Finally, Section 6 concludes the study by summarizing the key findings and contributions.

II. UCSYSPOOF DATASET

This section outlines an overview of the UCSYSpoof dataset, which includes five distinct subsets tailored for spoofing detection in ASV tasks. The dataset features both genuine and spoofed speech samples. The genuine portion comprises 12,000 utterances, equally distributed among three female speakers, with 4,000 utterances per speaker. The

spoofed portion, consisting of 63,932 utterances, is generated using five sophisticated techniques. These techniques involve vocoder methods powered by HiFi-GAN and parallel WaveGAN [21], voice conversion through the pre-trained FreeVC model, and both GMM-based and Differential GMM-based voice conversion strategies. Summary of each technique used in the UCSYSpoof dataset is provided in Table I.

TABLE I.

DETAILED STATISTIC OF UCSYSPOOF DATASET

Label	Dataset Type	No. of Utterances
Genuine	Genuine dataset	12,000
Spoofed	HifiGAN	11,966
	Parallel WaveGAN	11,966
	FreeVC pretrained dataset	24,000
	GMM VC dataset	8,000
	GMM DIFFVC dataset	8,000

To develop a vocoder-based dataset, two GAN-based neural vocoders, HiFi-GAN and Parallel WaveGAN, were utilized, both specifically trained on Myanmar speech data [21]. HiFi-GAN [8] comprises a fully convolutional neural network generator with multi-scale and multi-period discriminators, enabling efficient, high-fidelity speech synthesis. Meanwhile, Parallel WaveGAN [9] offers a lightweight, fast waveform generation approach that avoids distillation, achieving realistic synthesis by optimizing adversarial loss in the waveform domain alongside a multi-resolution short-time Fourier transform (STFT) loss, thus eliminating the need for complex probability density distillation.

The FreeVC-based dataset, presented in TABLE I, was generated using FreeVC [10], a text-free, one-shot voice conversion system. FreeVC leverages a pre-trained WaveLM [11] to extract content information by applying an information bottleneck, without requiring text annotation, and adopts the end-to-end architecture of VITS. GMM-based voice conversion (VC) methods use parallel speech utterances from source and target speakers. This section highlights two common approaches: the maximum likelihood parameter generation (MLPG) with global variance (GV) and the vocoder-free log-spectral differentiation method (DIFFVC). Typically, the VC process involves two stages: training and conversion. The objective is to learn a mapping function from training data that transforms test features of the source speech (including prosodic and spectral features) into the acoustic space of the target speech [12].

III. Features Analysis

This paper presents five datasets developed specifically for the Myanmar Spoof Dataset, aimed at facilitating the analysis of deepfake speech. Figure 1 illustrates the system architecture, outlining the stages of data preprocessing, feature extraction, and classification. These stages are crucial for understanding how the extracted features contribute to identifying spoofed speech. For instance, during the feature extraction stage, this experiment computes measures such as jitter, shimmer, Harmonics-to-Noise Ratio (HNR), and Cepstral Peak Prominence (CPP). These features are sensitive indicators of vocal quality, with jitter, shimmer, and HNR highlighting irregularities in pitch and amplitude, while CPP assesses the clarity and prominence of the speech harmonic structure. Together, they can reveal subtle inconsistencies indicative of deepfake manipulation.

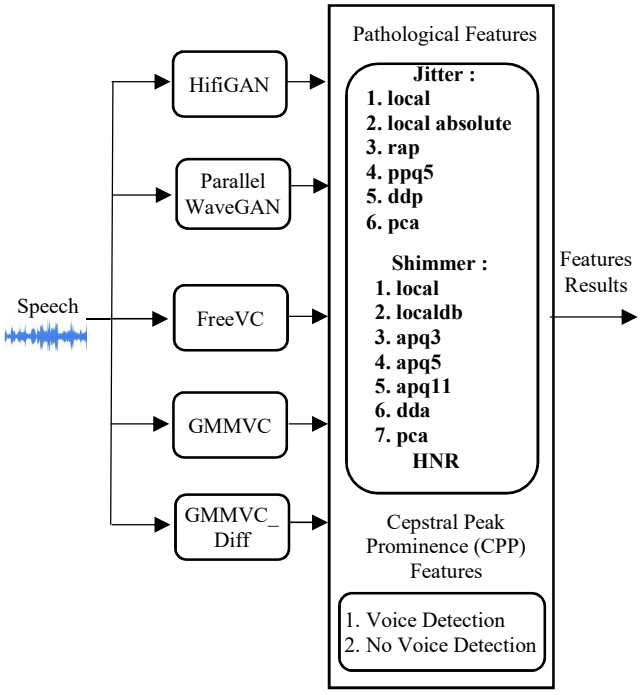


Figure 1. System Design for Comparing Datasets and Features

IV. Pathological Features and CPP features

This section details the implementation of pathological and CPP features. Pathological features are vocal characteristics that reveal abnormalities or disorders [13]. In spoof datasets, these features help identify differences between genuine and manipulated speech. They include irregular pitch, hoarseness, vocal tremor, reduced loudness,and altered vocal quality. Analyzing these deviations can detect if a voice signal has been altered or synthesized. In this comparative experiment, we utilized three pathological features: HNR, jitter and shimmer [14],[15]. These features are crucial for differentiating between authentic voice recordings and deepfake speech, thereby enhancing voice authentication systems.

A. Harmonics-to-noise ratio (HNR)

HNR provides insight into the periodicity and stability of the voice, which can be useful for detecting voice abnormalities or manipulations, such as deepfake speech. The noise component (ιEn) is calculated as the energy of the residual obtained by subtracting the mean waveform from each cycle. The harmonic energy (γEn) is derived from the energy of the average waveform constructed synchronously over ten consecutive glottal cycles within a frame pitch. Therefore, this feature relies on a prior estimation of f0 [20]. Here are key HNR features:

$$HNR = 20\ log\left(\frac{\gamma En}{\iota En}\right) \quad (1)$$

B. Jitter Features

Jitter is a fundamental parameter in voice analysis, particularly effective in identifying irregularities in speech signals. It evaluates the fluctuations in the fundamental

frequency from one cycle to the next, which may indicate voice instability or synthetic manipulation [16],[17]. This research focuses on six distinct types of jitter, as detailed below.

1) Jitter (local) quantifies the average absolute variation in fundamental frequency between consecutive periods within a short speech segment. The calculation is given by:

$$Jitter\ (local) = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100 \quad (2)$$

where T_i represents the fundamental frequency period lengths, and N is the number of periods analyzed.

2) Local Absolute Jitter measures the average absolute deviation in fundamental frequency between consecutive cycles, which may indicate voice manipulation or spoofing. The calculation is as follows:

$$Jitter\ (local) = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100 \quad (3)$$

where T_i denotes the period lengths of the fundamental frequency, and N represents the total number of periods analyzed.

3) Jitter (rap) is calculated as the average absolute difference between a given period and its two neighboring periods, normalized by the average period. It is defined as:

$$RAP = \frac{1}{N}\sum_{i=1}^{N}\frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \quad (4)$$

where A_i represents the amplitude of the *i-th* period, and N is the number of periods analyzed.

4) Jitter (ppq5) evaluates as the average absolute deviation between a period and its average, including the five nearest periods, divided by the average period. It is defined as:

$$PPQ5 = \frac{1}{N-5}\sum_{i=1}^{N-5}\frac{|T_i - T_{i+5}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100 \quad (5)$$

where T_i denotes the fundamental frequency periods, and N represents the total number of periods analyzed.

5) Jitter (DDP) measures pitch variability by analyzing the changes between consecutive pitch variations, making it sensitive to subtle irregularities from spoofing techniques. The calculation is given by:

$$DDP = \frac{1}{N-2}\sum_{i=1}^{N-2}|((T_{i+1} - T_i) - (T_{i+2} - T_{i+1}))| \quad (6)$$

where T_i represents the fundamental frequency periods, and N is the number of periods analyzed.

6) Jitter (PCA) uses dimensionality reduction to identify principal components that capture the most variance in pitch data, aiding in distinguishing normal speech from spoofed speech by revealing key patterns in jitter features.

C. Shimmer Features

Shimmer features evaluate amplitude variations in voice signals, providing critical insights into vocal intensity stability and facilitating the detection of irregularities that may signal pathological conditions or manipulations, such as deepfake speech. This study emphasizes seven specific shimmer features, detailed as follows:

1) Local Shimmer measures the average absolute difference in amplitude between consecutive pitch periods within a short segment of speech. It is calculated as:

$$Shimmer\ (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \quad (7)$$

where A_i represents the amplitude at the *i-th* period, and N is the number of periods.

2) Local Shimmer (dB): Local shimmer in decibels normalizes the amplitude differences by converting them into a logarithmic scale. It is calculated as:

$$Shimmer\ (local, dB) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}20\ \log_{10}\left(\frac{A_i}{A_{i+1}}\right)}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \quad (8)$$

3) APQ3 (Average Perturbation Quotient 3) measures amplitude perturbations over a window of three consecutive periods. It is computed as:

$$APQ3 = \frac{1}{N-2}\sum_{i=1}^{N-2}\frac{|A_i - 2A_{i+1} + A_{i+2}|}{A_i + A_{i+1} + A_{i+2}} \times 100 \quad (9)$$

where A_i represents the amplitude at the *i-th* period.

4) APQ5 (Average Perturbation Quotient 5) extends the concept of APQ3 to a window of five consecutive periods.

5) APQ11 (Average Perturbation Quotient 11) analyzes amplitude perturbations across eleven consecutive periods, offering a detailed view of longer-term variations in voice intensity.

6) Shimmer (DDA) calculates the differences between consecutive amplitude differences. It is expressed as:

$$DDA = \frac{1}{N-2}\sum_{i=1}^{N-2}|(A_{i+1} - A_i) - (A_{i+2} - A_{i+1})| \quad (10)$$

7) Shimmer (PCA) transforms the original features into uncorrelated components, simplifying the analysis and interpretation of complex shimmer patterns.

Cepstral Peak Prominence (CPP) is an important feature in voice analysis, providing insights into the periodicity and quality of speech signals [18]. It is particularly useful for differentiating between voiced and non-voiced segments. A brief overview of CPP for voice and non-voice detection as follows:

1) CPP for Voice Detection

CPP quantifies the prominence of the cepstral peak relative to the surrounding cepstral coefficients.

It is computed as follows:

$$CPP = \frac{P_{peak} - Mean_{sidebands}}{Standard\ Deviation_{sidebands}} \quad (11)$$

where P_{peak} represents the amplitude of the cepstral peak, and $Mean_{sidebands}$ and $Standard\ Deviation_{sidebands}$ are the mean and standard deviation of the cepstral coefficients in the regions adjacent to the peak (sidebands).

2) CPP for Non-Voice Detection

The CPP is calculated similarly, focusing on lower values to assess periodicity. Low CPP values suggest minimal cepstral peak prominence, indicating non-voiced segments like silence or background noise. A lower CPP threshold helps identify areas with minimal voice activity.

V. Features Results and Discussion

This analysis is conducted on 100 randomly selected samples using five methods: HifiGAN, ParallelWaveGAN, and three voice conversion techniques. Each voice conversion

method is further divided into two types: VC12 (converting from speaker_1 as the source to speaker_2 as the target) and VC13 (converting from speaker_1 as the source to speaker_3 as the target). Additionally, the study incorporates two variants of GMMVC (VC12 and VC13) as well as two variants of GMMVC_Diff (VC12 and VC13). In total, the study examines eight distinct dataset types. Detailed information about the datasets used in the experiment is provided in Table II.

TABLE II.

DETAILED EXPERIMENT OF DATASETS

Label	Dataset Type
Spoofed	HifiGAN
	ParallelWaveGAN
	VC12 (FreeVC)
	VC13 (FreeVC)
	GMMVC12 (GMMVC)
	GMMVC13 (GMMVC)
	GMMVC12_DiffVC
	GMMVC13_DiffVC

A. Experiment results

The comparative analysis utilized a total of sixteen features: fourteen pathological features and two CPP features, extracted from eight distinct dataset types. For the extraction of jitter, shimmer, and HNR features, we employed Python for analysis. The extraction of the CPP features was conducted using Praat [19].

Fig. 2 presents comparative results of Harmonics-to-Noise Ratio (HNR) features across eight datasets. The analysis reveals noticeable variations in HNR values among the different datasets. Notably, while HNR values for the HifiGAN and ParallelWaveGAN datasets exhibit some differences, these variations are not statistically significant, suggesting that both methods produce similar levels of harmonic clarity.

In contrast, the voice conversion techniques, particularly GMMVC_DIFF applied to VC12 and VC13, demonstrate substantial differences in HNR features. The GMMVC_DIFF method yields higher HNR values, indicating improved vocal quality and a more robust ability to preserve harmonic content compared to the other techniques. This suggests that GMMVC_DIFF is particularly effective in enhancing the naturalness and intelligibility of the converted speech.

Fig. 3 shows Jitter features, including Jitter (local), Jitter (local absolute), Jitter (RAP), Jitter (PPQ5), Jitter (DDP), and Jitter (PCA), as illustrated. In Fig. 3 displays significant features results, with GMMVC showing notable significance compared to other methods. Jitter (local) and Jitter (local absolute) on VC12 and VC13 reflect the short-term frequency instability in the speech signal, indicating minor variations in vocal fold vibrations. Jitter (PPQ5) shows increased significance, highlighting the average perturbation across multiple cycles, which is crucial in detecting subtle voice abnormalities. The most pronounced differences are observed in Jitter (RAP) and Jitter (DDP), which measure relative average perturbation and absolute cycle-to-cycle variation, respectively, providing deeper insights into irregularities in vocal fold vibration. Lastly, Jitter (PCA) demonstrates its relevance by capturing the principal components of jitter variations, summarizing complex patterns in the voice data.

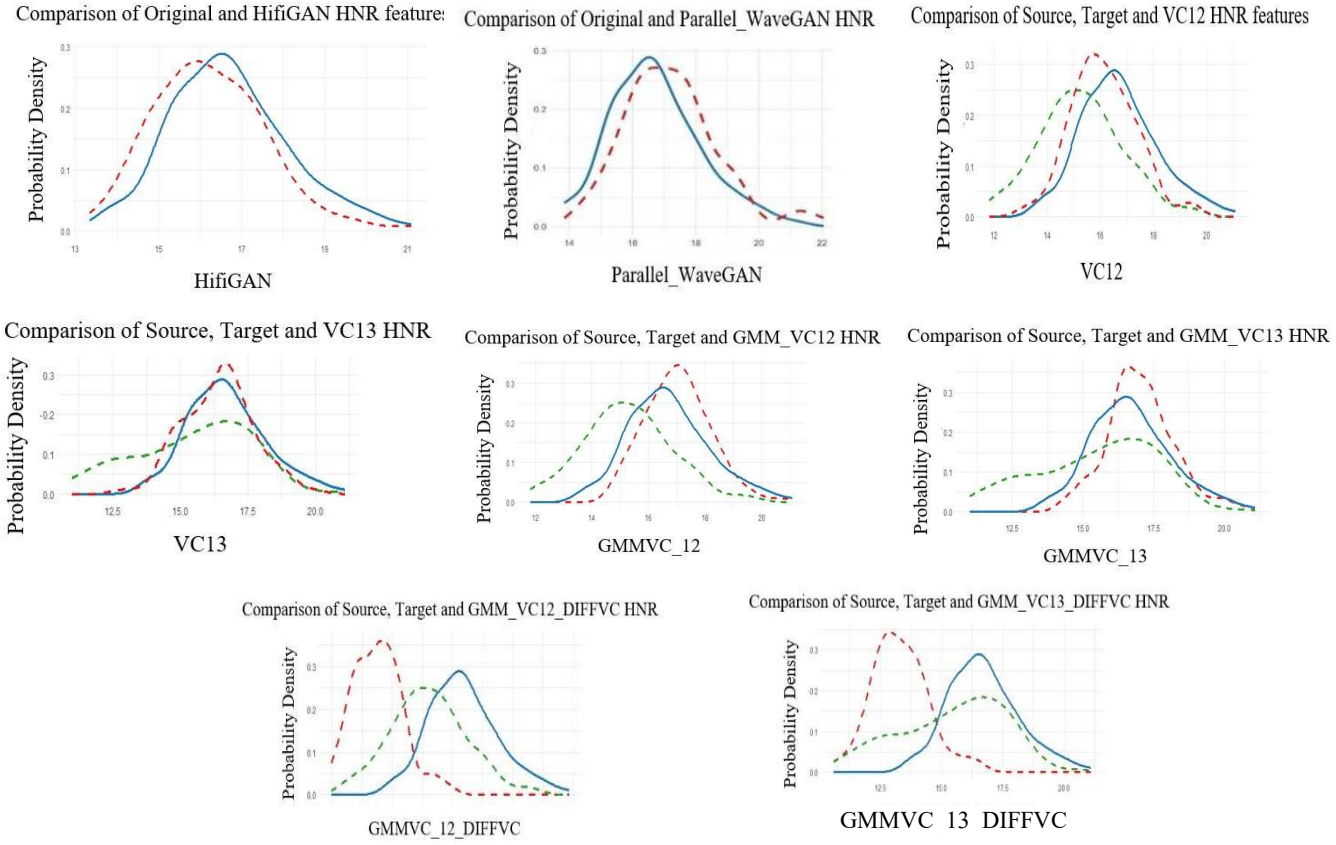


Figure 2. Comparative results for HNR features

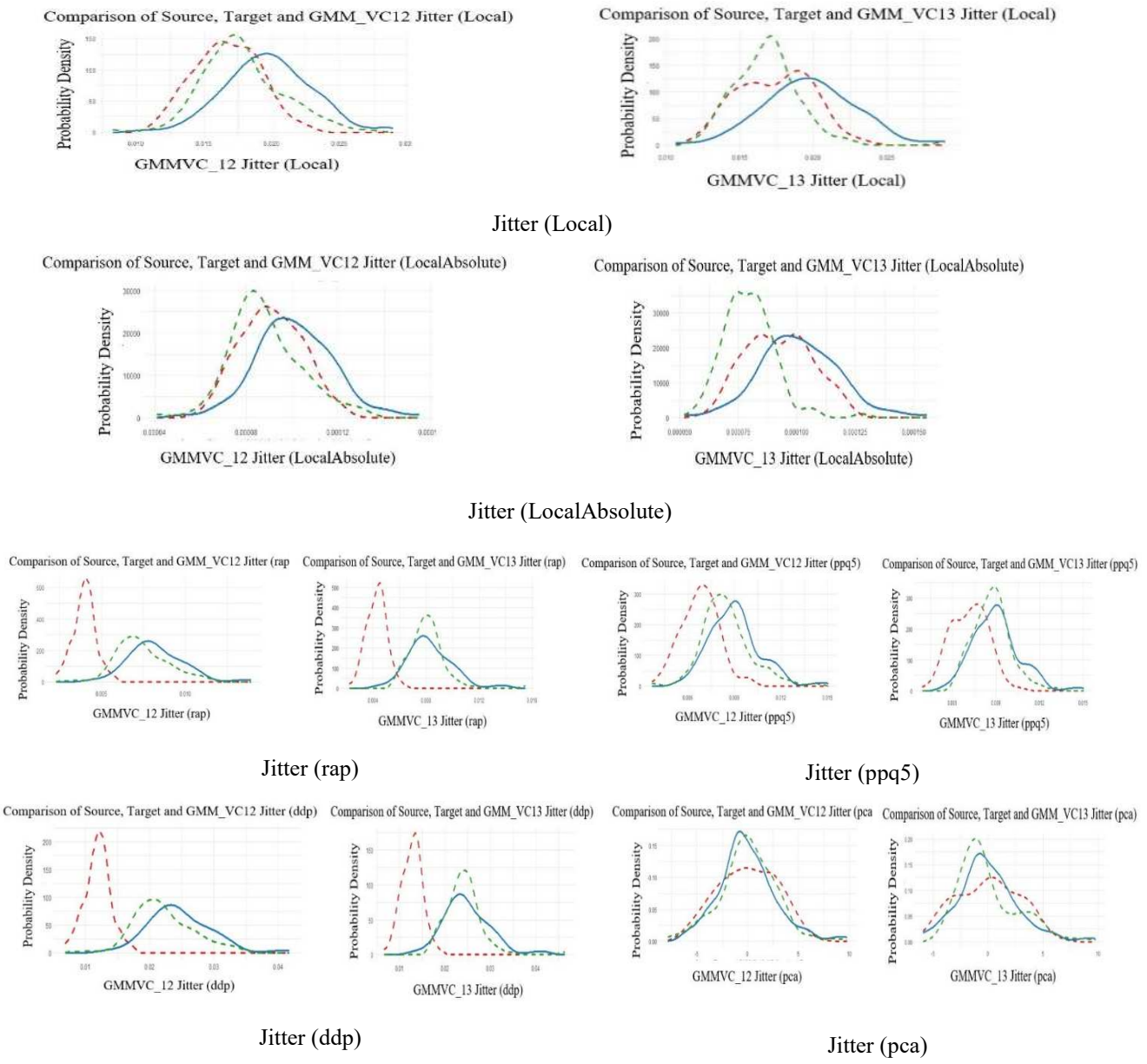
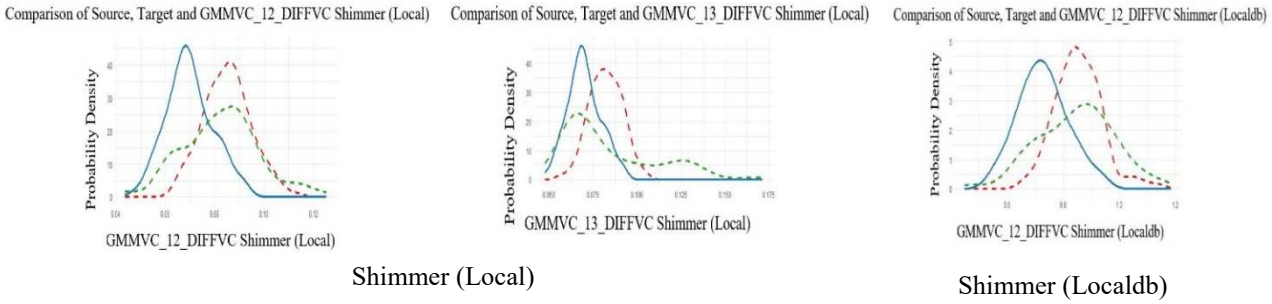


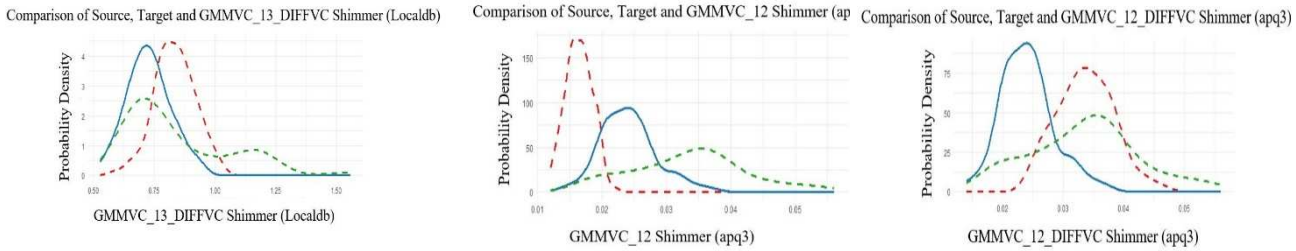
Figure 3. Significant features results for Jitter features

Fig. 4 presents line graphs for the key Shimmer features, including Shimmer (local), Shimmer (local dB), Shimmer (APQ3), Shimmer (APQ5), Shimmer (APQ11), Shimmer (DDA), and Shimmer (PCA), as shown. In Fig. 4 illustrates significant features results for Shimmer features, highlighting the prominence of GMMVC_Diff compared to other methods. Shimmer (local) and Shimmer (local dB) in VC12 and VC13 capture short-term amplitude variability in the speech signal, revealing fluctuations in vocal intensity. Shimmer (APQ3) shows even greater significance in both GMMVC and GMMVC_Diff, reflecting the averaged amplitude perturbations over three cycles, which is crucial for identifying subtle irregularities in voice quality. The most substantial differences are observed in Shimmer (APQ5) within the HifiGAN and GMMVC_Diff methods, representing the average amplitude perturbation over five cycles, and highlighting the sensitivity of these methods to voice disturbances. Additionally, Shimmer (APQ11) and Shimmer (DDA) in GMMVC_Diff exhibit notable significance, with Shimmer (APQ11) measuring amplitude variability over eleven cycles, and Shimmer (DDA) focusing on the average difference between the absolute differences of amplitudes. Finally, Shimmer (PCA) captures the principal components of Shimmer variations, summarizing complex amplitude perturbations within the voice data, and also shows significant distinctions across the datasets.

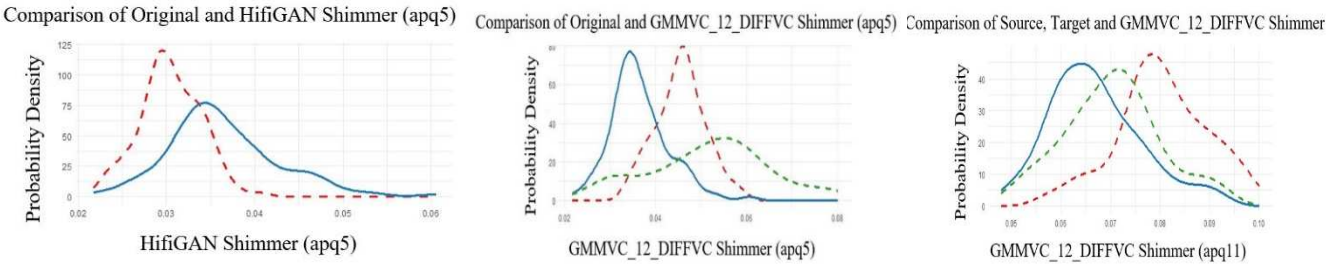
Fig. 5 presents significant features for CPP feature, comparing Voice Detection and No Voice Detection, as illustrated below.



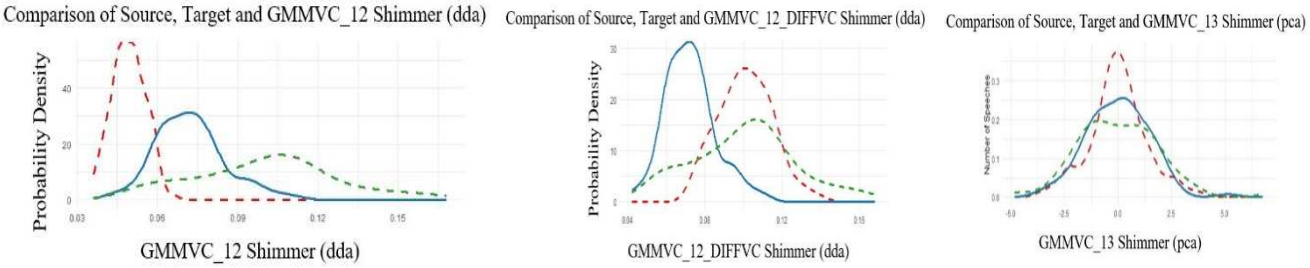
Shimmer (Local) Shimmer (Localdb)



Shimmer (Localdb) Shimmer (apq3)

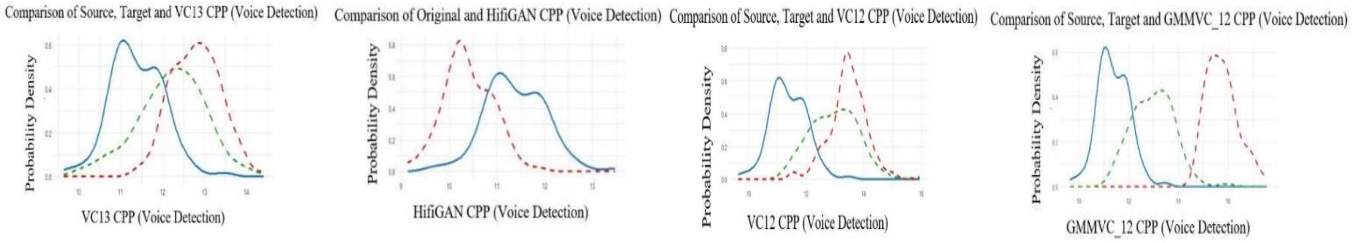


Shimmer (apq5) Shimmer (apq11)

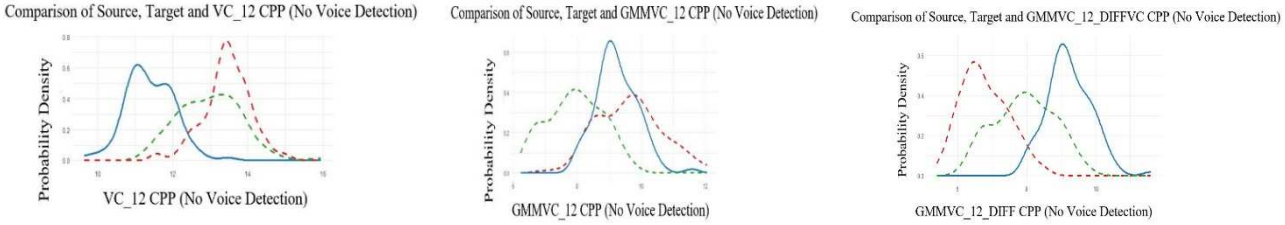


Shimmer (dda) Shimmer (pca)

Figure 4. Significant features results for Shimmer features



CPP (Voice Detection)



CPP (No Voice Detection)

Figure 5. Significant features results for CPP features

B. Discussion on Features of the Experiments

In the analysis of the Myanmar spoof datasets features, significant variations were observed in HNR, Jitter, Shimmer, and CPP features across different methods, highlighting their importance in spoof detection. While the differences between HifiGAN and ParallelWaveGAN were relatively minor, the voice conversion techniques, particularly GMMVC_DIFF (VC12 and VC13), exhibited notable distinctions. These features are critical in differentiating between genuine and spoofed speech signals.

Jitter and Shimmer features are particularly effective in detecting inconsistencies in vocal fold vibrations that are often indicative of spoofing. GMMVC_DIFF, in particular, displayed the most pronounced differences in these features, making it a powerful tool for identifying minor perturbations that might go unnoticed with other methods. Jitter features, such as Jitter (local) and Jitter (RAP), focus on variations in pitch, providing insights into the stability of vocal fold vibrations, while Shimmer features, such as Shimmer (local) and Shimmer (APQ5), highlight amplitude fluctuations that can signify irregularities in the vocal signal intensity.

CPP further enhances spoof detection by serving as a key indicator of voice quality. CPP effectively measures the prominence of the cepstral peak, which correlates with the perceived clarity and robustness of a voice. In both voice and no-voice detection scenarios, CPP exhibited significant results across various methods, particularly in the GMMVC and GMMVC_DIFF techniques, underscoring its critical role in identifying subtle differences in voice quality that are often exploited in spoofing attacks.

Collectively, these features provide a robust framework for detecting spoofed speech, enabling more accurate differentiation between authentic and synthetic audio.

VI. CONCLUSION

Based on the evaluation, this study highlights the importance of pathological features, specifically HNR, jitter, shimmer, and CPP, in effectively distinguishing genuine voice recordings from deepfake speech. The analysis demonstrates that these features show significant variations across various spoofing techniques. The experimental results reveal that voice conversion methods, particularly GMMVC_DIFF (VC12 and VC13), exhibit pronounced differences in these features. This indicates their effectiveness in detecting subtle anomalies that are indicative of spoofing. Jitter and shimmer features, which measure short-term pitch variations and amplitude fluctuations, respectively, are highly effective in identifying inconsistencies in vocal fold vibrations and voice intensity.These features are sensitive to the subtle deviations often introduced during spoofing processes. Meanwhile, CPP enhances spoof detection by evaluating voice quality and revealing subtle discrepancies that may be exploited in spoofing attacks. The ability of these features to discern subtle variations in speech signals is fundamental to the advancement of effective spoof detection systems.

The proposed features analysis effectively identifies and leverages significant features derived from five distinct spoofing techniques, demonstrating robust performance in differentiating between genuine and spoofed speech using pathological and CPP features. However, the efficacy of these features may diminish in certain contexts, particularly in environments characterized by significant background noise or poor audio quality. To address these challenges, future research should focus on the integration of additional features and the development of more advanced models to enhance the system's robustness and generalizability.

ACKNOWLEDGMENT

The authors express their gratitude to the "Spoofing Detection for Automatic Speaker Verification" project under ASEAN IVO (https://www.nict.go.jp/en/ asean_ivo/) and NICT (https://www.nict.go.jp/en/) for their financial support.

REFERENCES

[1] S. H. Mohammadi and A. Kain, ‘‘An overview of voice conversion systems,’’ Speech Commun., vol. 88, pp. 65–82, Apr. 2017.

[2] B. Sisman, J. Yamagishi, S. King, and H. Li, ‘‘An overview of voice conversion and its challenges: From statistical modeling to deep learning,’’ IEEE/ACM Trans. Audio, Speech, Language Process., vol. 29, pp. 132–157, 2021.

[3] Y. Ren, Y. Ruan, X. Tan, et al., “Fastspeech: Fast, robust and controllable text to speech,” Advances in neural information processing systems, vol. 32, 2019.

[4] Z. Wu, C. E. Siong, and H. Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition.” in INTERSPEECH, 2012.

[5] M. Sahidullah, T. Kinnunen, and C. Hanilc¸i, “A comparison of features for synthetic speech detection,” in INTERSPEECH, 2015.

[6] M. Todisco, H. Delgado, and N. W. Evans, “A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients.,” in Odyssey, vol. 2016, 2016, pp. 283–290.

[7] A. Chaiwongyen et al. “Deepfake-speech Detection with Pathological Features and Multilayer Perceptron Neural Network.” 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (2023): 2182-2188.

[8] J. Su, Z. Jin and A. Finkelstein, ”HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks”, arXiv preprint arXiv:2006.05694, 2020.

[9] R. Yamamoto, E. Song and J. M. Kim, ”Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram”, ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p.6199-6203, IEEE, 2020.

[10] J. Li, W. Tu, and L. Xiao, ”Freevc: Towards high-quality text-free oneshot voice conversion”, In ICASSP IEEE International Conference

on Acoustics, Speech and Signal Processing (ICASSP), p.1-5, IEEE, 2023.

[11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, et.al, "WavLM: Large-Scale Self-Supervised Pre-training for Full Stack Speech Processing", IEEE Journal of Selected Topics in Signal Processing, 16(6), pp.1505-1518, 2022.

[12] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical Voice Conversion with WaveNet-Based Waveform Generation", In Interspeech, pp.1138-1142, August 2017.

[13] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 1243–1247.

[14] I. R. Titze and H. Liang, ''Comparison of Fo extraction methods for high precision voice perturbation measurements,'' J. Speech, Lang., Hearing Res., vol. 36, no. 6, pp. 1120–1133, Dec. 1993.

[15] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, ''Bulbar ALS detection based on analysis of voice perturbation and vibrato,'' in Proc. Signal Process., Algorithms, Architectures, Arrangements, Appl. (SPA), Sep. 2019, pp. 267–272.

[16] M. Farrus, J. Hernando, and P. Ejarque, "Jitter and ´ shimmer measurements for speaker recognition," in 8th Annual Conference of the International Speech Communication Association, Aug. 27-31, 2007.

[17] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," Procedia Technology, vol. 9, pp. 1112–1122, International Conference on Health and Social Care Information Systems and Technologies, ISSN: 2212-0173, 2013.

[18] R. Fraile, J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," Circuits & Systems Engineering Department, ETSIS Telecomunicacion, Spain, 2014.

[19] E. S. Heller Murray, A. Chao and L. Colletti (2022) "A Practical Guide to Calculating Cepstral Peak Prominence in Praat," Journal of Voice. 10.1016/j.jvoice.2022.09.002.

[20] J. Gomez-Garcı ´ ea, L. Moro-Vel ´ azquez, J. D. Arias- ´ Londono, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part iii: Review of acoustic modelling strategies," Biomedical Signal Processing and Control, vol. 66, p. 102 049, 2021.

[21] A. M. Hlaing, W. P. Pa, "Generative Adversarial Network based Neural Vocoder for Myanmar End-to-End Speech Synthesis", In the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024) [Accepted]