Analysis of Pathological Features for Spoof Detection

Myat Aye Aye Aung
Faculty of Computer Science
University of Computer Studies, Yangon, Myanmar

Hay Mar Soe Naing
Faculty of Computer Science
University of Computer Studies, Yangon, Myanmar

Aye Mya Hlaing
Faculty of Computer Science
University of Computer Studies, Yangon, Myanmar

Win Pa Pa
Faculty of Computer Science
University of Computer Studies, Yangon, Myanmar

Kasorn Galajit
NECTEC, National Science and Technology Development Agency
Pathum Thani, Thailand

Candy Olivia Mawalim
Japan Advanced Institute of Science and Technology,
Ishikawa, Japan

Outlines

- Abstract
- Introduction
- Key contributions
- UCSYSpoof Dataset
- Features analysis
- Pathological Features and CPP features
- Features Results and discussion
- Conclusion
- References

Abstract

- Deepfake speech presents significant challenges due to realistic sound and detection complexities.

- Effective feature selection and analysis are vital for improving spoof detection.

- This study focuses on analyzing pathological features within the Myanmar Spoof Dataset.

- Spoofed speech in the dataset is created using five distinct techniques:
  - **1. HiFiGAN (vocoder method)**
  - **2. Parallel WaveGAN (vocoder method)**
  - **3. FreeVC (pre-trained voice conversion)**
  - **4. GMM-based**
  - **5. Differential GMM-based voice conversion (GMMVC_DIFF)** methods.

- A comparative analysis is conducted on features, including Harmonics-to-Noise Ratio (HNR), jitter, shimmer features and Cepstral Peak Prominence (CPP) features.

- Results highlight the importance of these features for enhancing spoof detection precision.

Introduction

- In Fake Audio Detection (FAD) technologies, analyzing pathological features and Cepstral Peak Prominence (CPP) is essential.
- In the literature, acoustic features such as:
  - MFCC, LFCC features, constant-Q cepstral coefficients (CQCC) and then applied
  - pathological features: jitter, shimmer, HNR, Cepstral Harmonics-to-Noise Ratio (CHNR), Normalized Noise Energy (NNE), and Glottal-to-Noise Excitation Ratio (GNE).
- Investigated fourteen pathological features, including:
  - Six jitter metrics
  - Seven shimmer metrics
  - HNR
- Two CPP features
- Aimed to identify the most effective features for detecting subtle alterations in speech patterns.
- This identification is crucial for developing robust countermeasures against deepfake speech and ensuring the integrity of ASV systems.

Key contributions

- Conducted a comprehensive analysis evaluating fourteen pathological features.
- Assessment of CPP features in detecting subtle alterations in speech patterns.
- Identified significant feature variations across different spoofing techniques.
- Enhanced the precision of spoof detection systems through these findings.

UCSYSpoof Dataset

- Comprises five distinct subsets designed for spoofing detection in ASV tasks.
- Contains both genuine and spoofed speech samples.
- The spoofed portion, consisting of 63,932 audios, is generated using five sophisticated techniques.
- 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 Audios
Genuine	Genuine dataset	12,000
Spoofed	HifiGAN	11,966
	Parallel WaveGAN	11,966
	FreeVC	24,000
	GMM VC dataset	8,000
	GMM DIFFVC dataset	8,000

UCSYSpoof Dataset (cont'd)

- Vocoder-Based Dataset Development
  - Utilized two GAN-based neural vocoders: HiFi-GAN and Parallel WaveGAN, trained on Myanmar speech data.
  - **HiFi-GAN**:
  - Fully convolutional neural network generator.
  - **Parallel WaveGAN**:
  - Lightweight and fast waveform generation approach.
  - Achieves realistic synthesis without distillation.
- FreeVC-Based Dataset:
  - Leverages a pre-trained WaveLM for content extraction via an information bottleneck.
  - No text annotation required; adopts end-to-end architecture of VITS.
- GMM-Based Voice Conversion (VC) Methods:
  - Uses parallel speech utterances from source and target speakers.
  - Employs maximum likelihood parameter generation (MLPG) with global variance (GV) and vocoder-free log-spectral differentiation (DIFFVC).

Features analysis

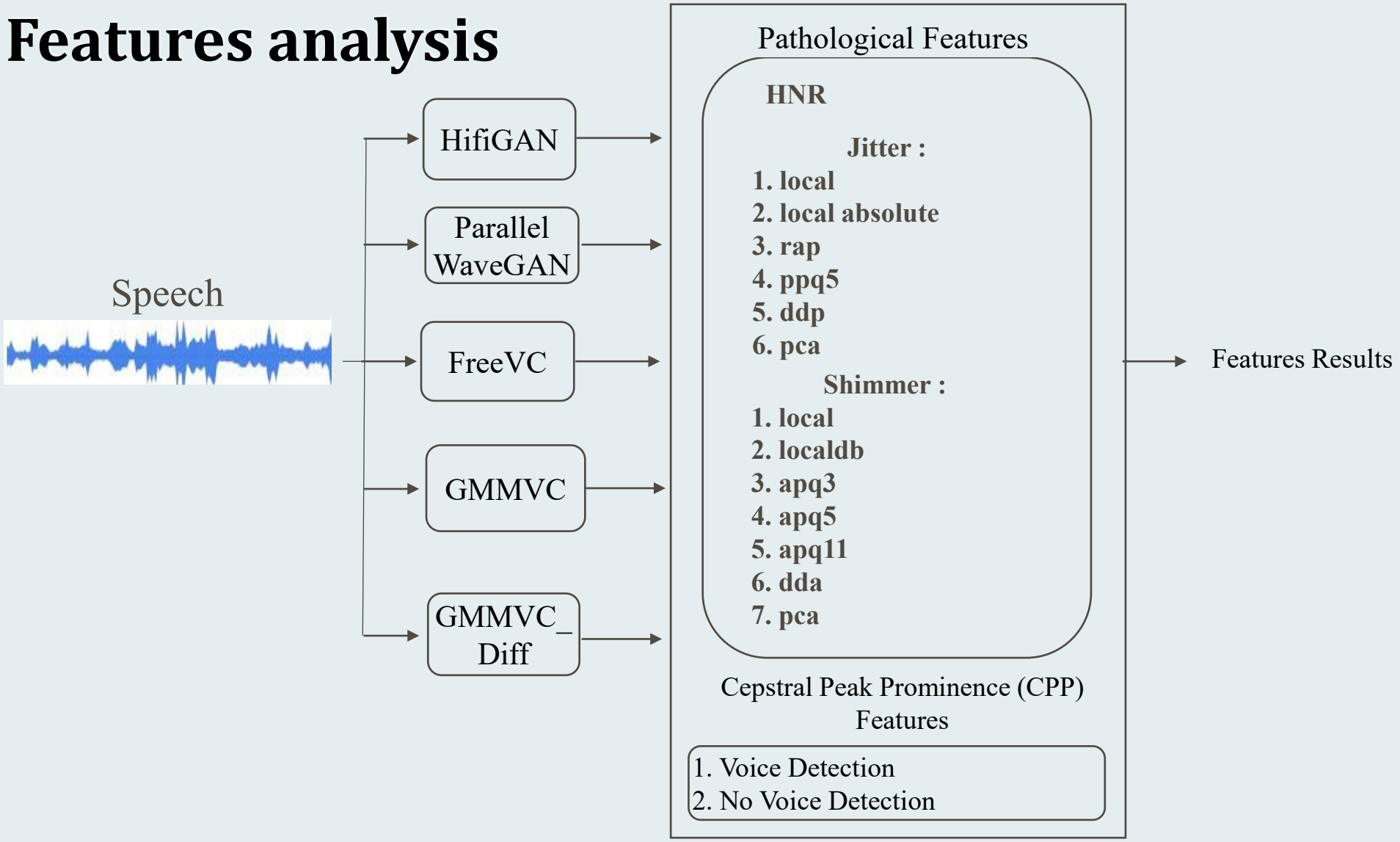


Figure 1. System Design for Comparing Datasets and Features

Pathological Features and CPP Features

- 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 alterations or syntheses in voice signals.
- Utilized three pathological features: HNR, jitter and shimmer.
- HNR:
  - Insights into the periodicity and stability.
- Jitter evaluates the fluctuations in the fundamental frequency from one cycle to the next, which may indicate voice instability or synthetic manipulation.
- Shimmer features provide critical insights into vocal intensity stability and facilitating the detection of irregularities that may signal pathological conditions or manipulations.
- CPP is an important feature in voice analysis, providing insights into the periodicity and quality of speech signals and useful for differentiating between voiced and non-voiced segments.

Features Results and Discussion

- 100 randomly selected samples for five methods.
- 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).
- 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

Features Results and Discussion (cont'd)

- Experiment results for HNR Features

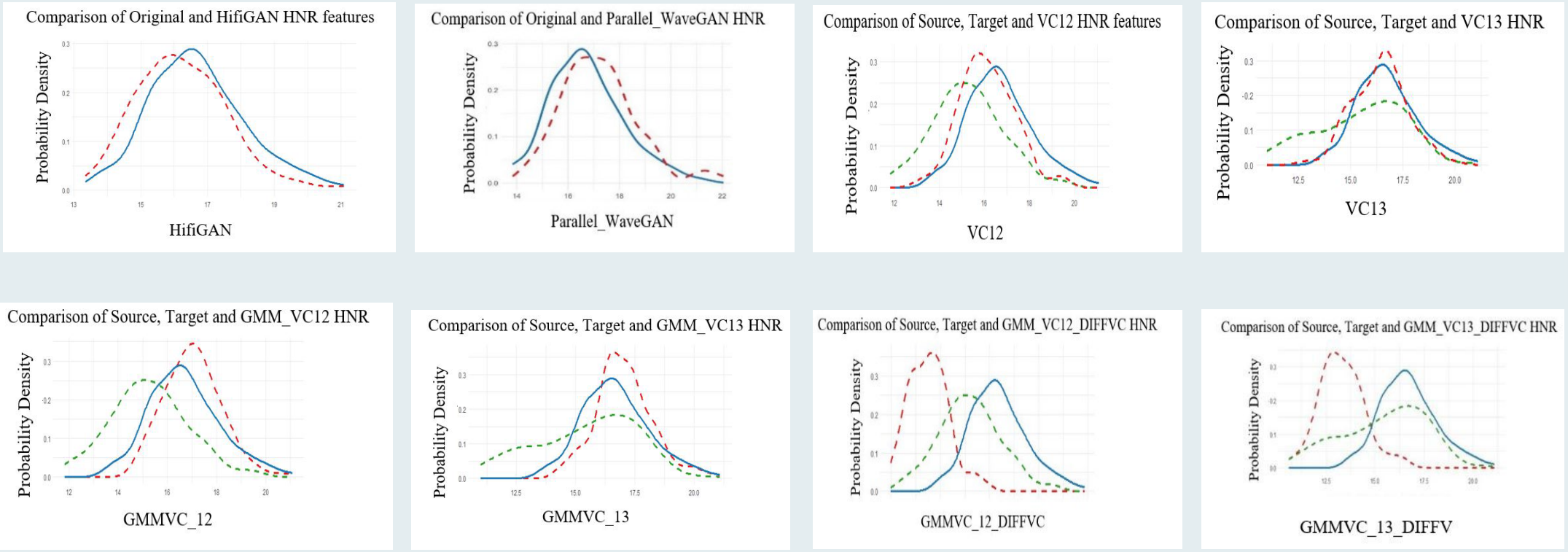
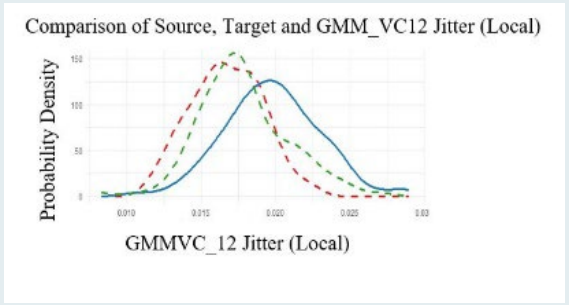
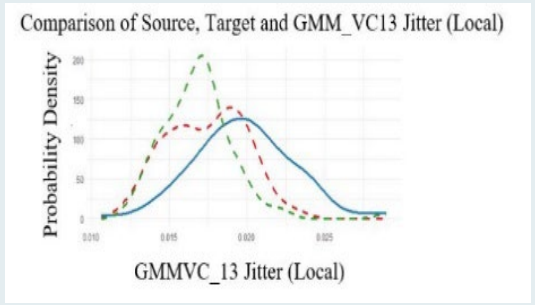


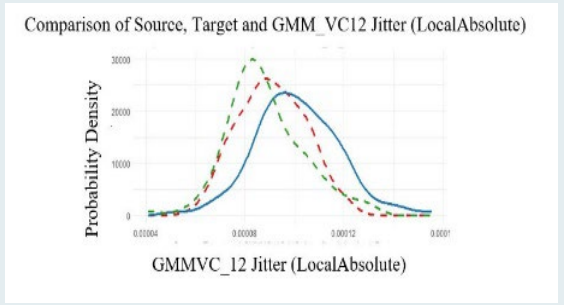
Figure 2. Comparative results for HNR features

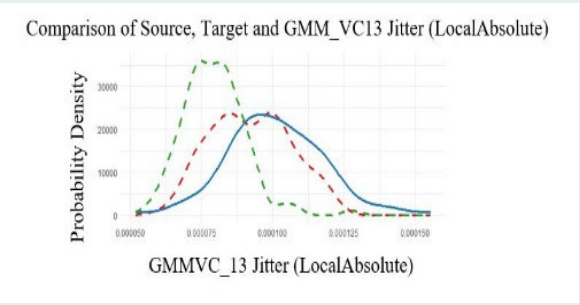
Figure 3. Significant features results for Jitter features



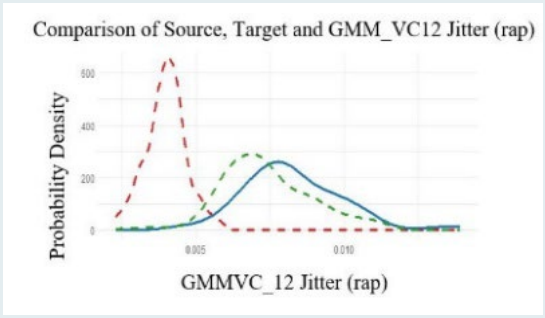


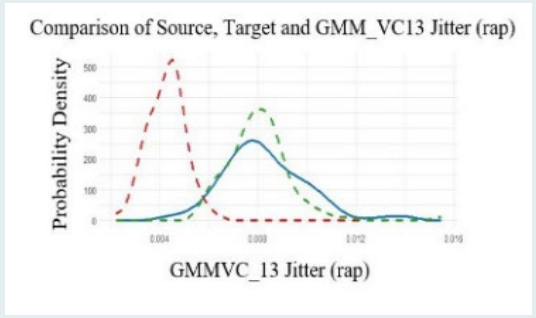
Jitter (Local)



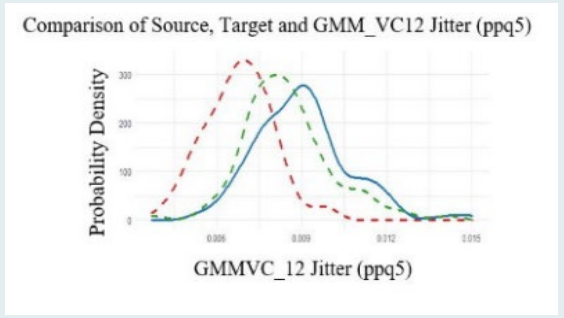


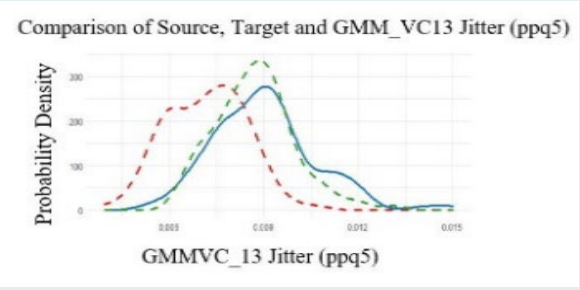
Jitter (Local Absolute)





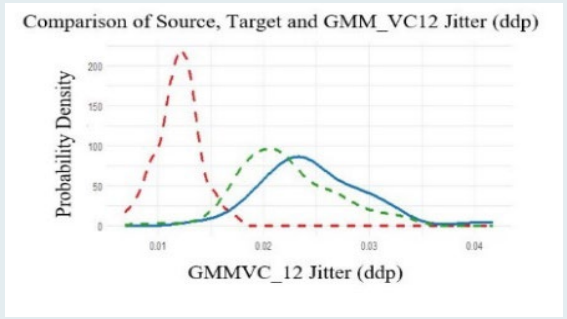
Jitter (rap)

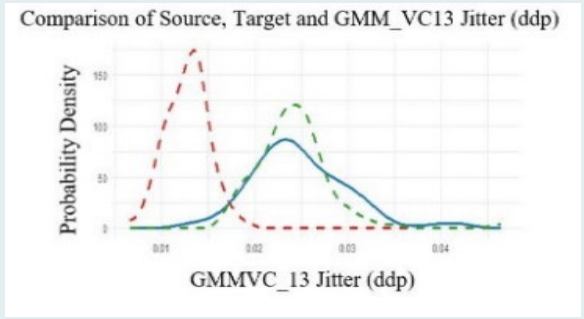




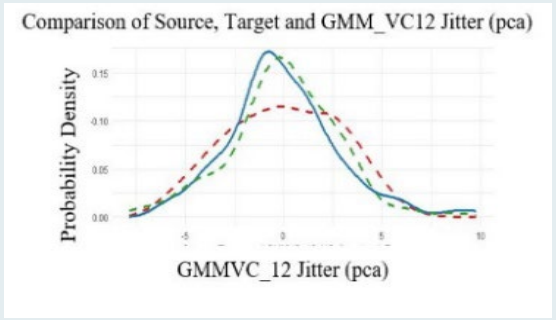
Jitter (ppq5)

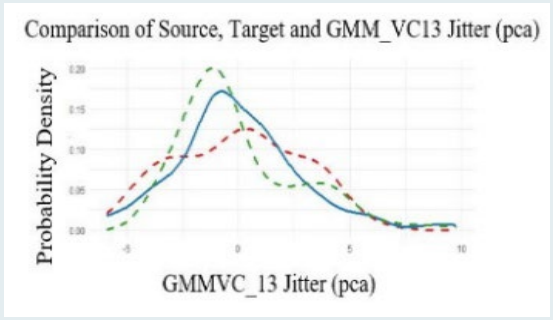
- Significant features results for Jitter Features (cont'd)





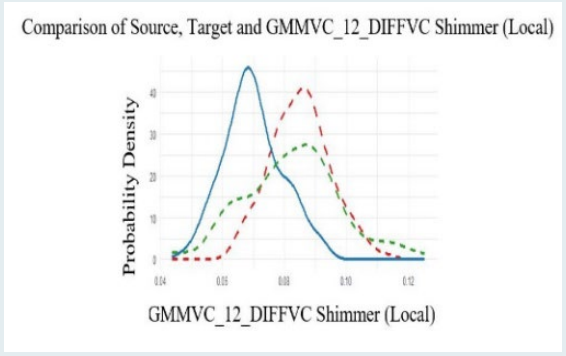
Jitter (ddp)

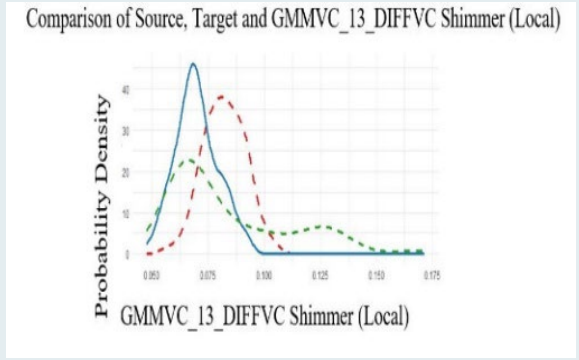




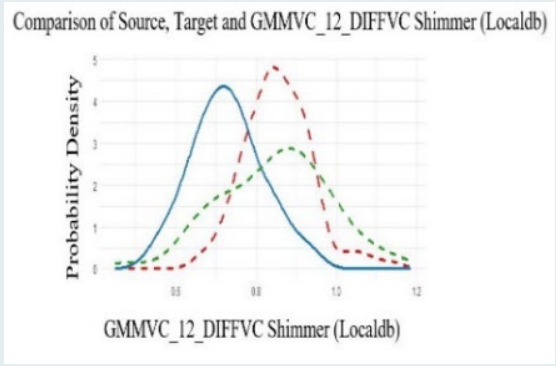
Jitter (pca)

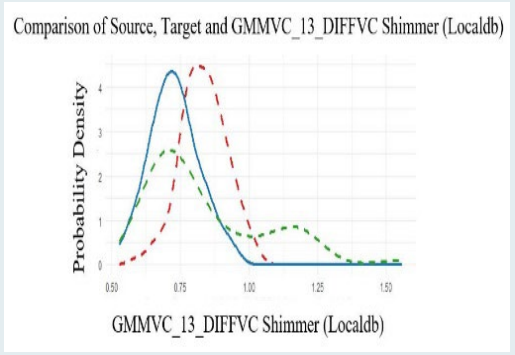
Figure 4. Significant features results for Shimmer features





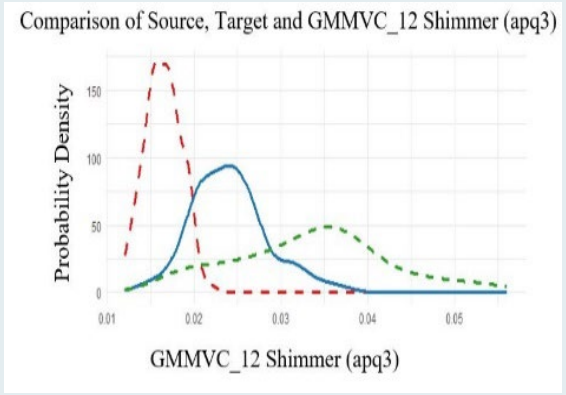
Shimmer (Local)

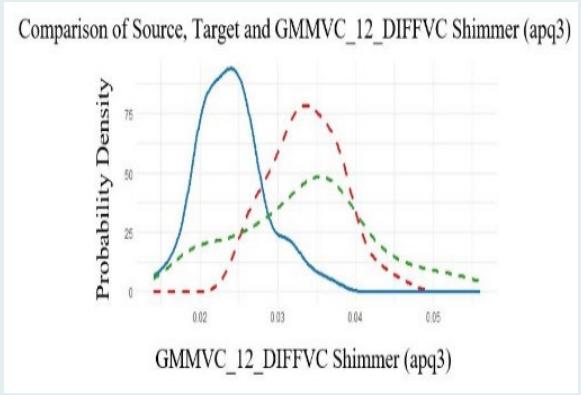




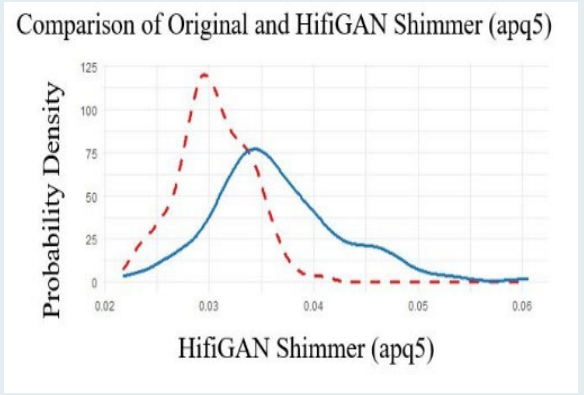
Shimmer (Localdb)

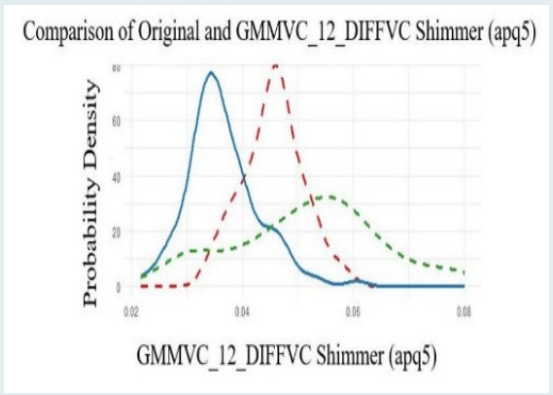
- Significant features results for Shimmer Features (cont'd)



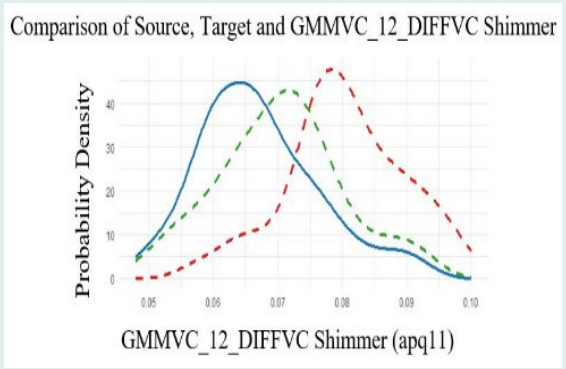


Shimmer (apq3)

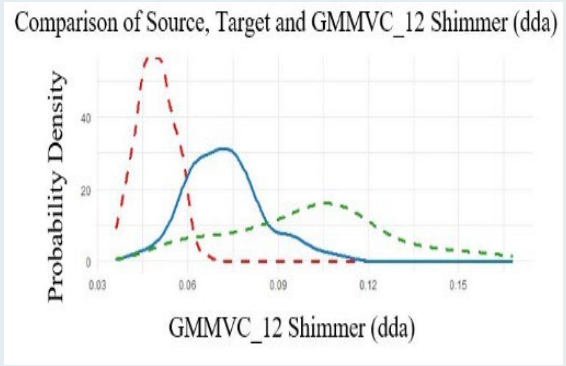




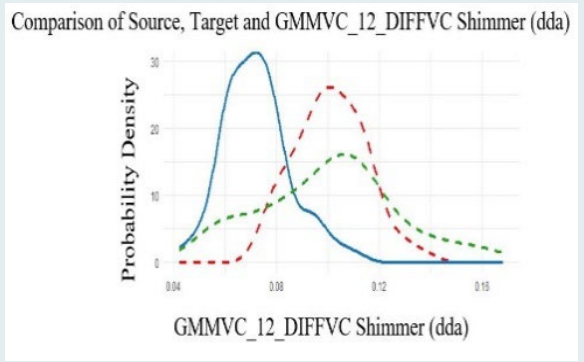
Shimmer (apq5)

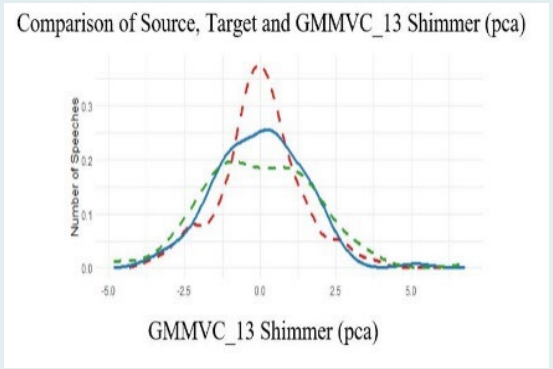


Shimmer (apq11)



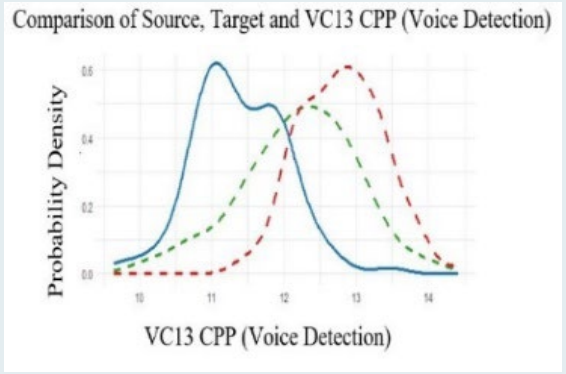
Shimmer (dda)

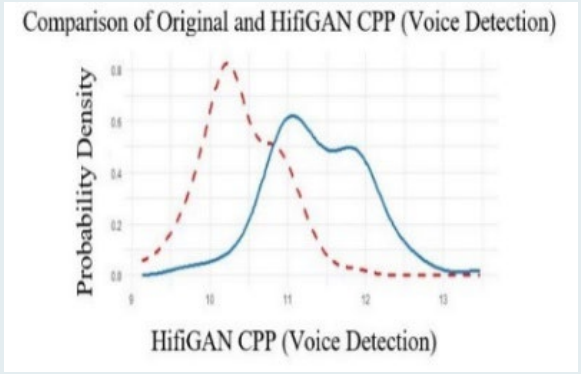


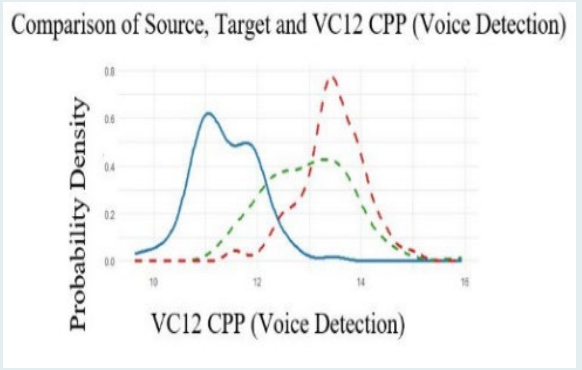


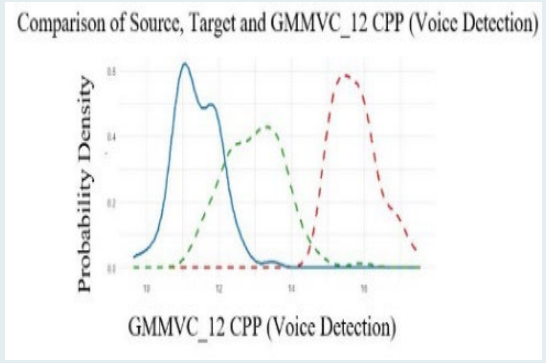
Shimmer (pca)

- Significant features results for CPP Features (cont'd)

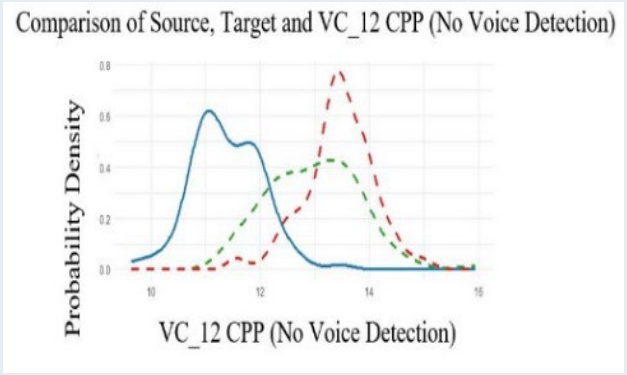


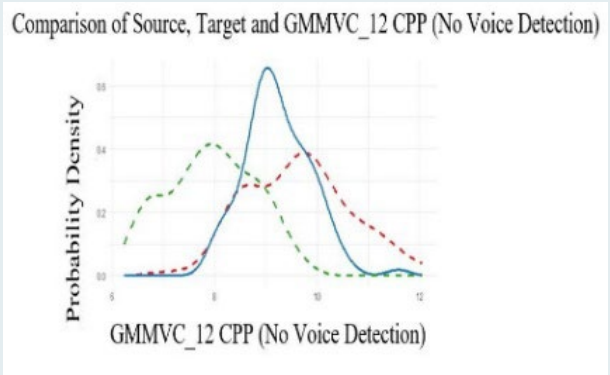


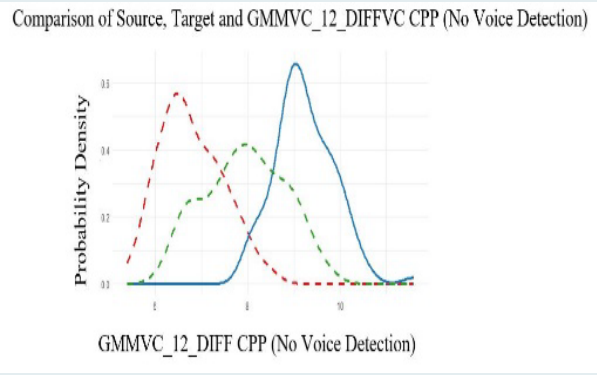




CPP (Voice Detection)







CPP (No Voice Detection)

Figure 5. Significant features results for CPP features

Discussion on Features of the Experiments

- Myanmar spoof datasets features, significant variations were observed in HNR, Jitter, Shimmer, and CPP features.
- The differences between HifiGAN and ParallelWaveGAN were relatively minor, the voice conversion techniques, particularly GMMVC_DIFF (VC12 and VC13), exhibited notable distinctions.
- 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.

Discussion on Features of the Experiments (cont'd)

- 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.

Conclusion

- 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 are highly effective in identifying inconsistencies in vocal fold vibrations and voice intensity. CPP enhances spoof detection by evaluating voice quality.
- The ability of these features to discern subtle variations in speech signals is fundamental to the advancement of effective spoof detection systems.
- However, the efficacy of these features may diminish in certain contexts, particularly in background noise or poor audio quality.
- 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.

References

- S. H. Mohammadi and A. Kain, ‘‘An overview of voice conversion systems,’’ Speech Commun., vol. 88, pp. 65–82, Apr. 2017.
- 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.
- 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.
- 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.
- M. Sahidullah, T. Kinnunen, and C. Hanilc¸i, “A comparison of features for synthetic speech detection,” in INTERSPEECH, 2015.
- 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.
- 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.

References (cont'd)

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

Thank You!!!