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-COCOSDA64382.2024.10800220

UCSYSpoof: A Myanmar Language Dataset for Voice Spoofing Detection

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

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

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

Myat Aye Aye Aung
Faculty of Computer Science
University of Computer Studies, Yangon
Yangon, Myanmar
myatayeayeaung@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*—Automatic Speaker Verification (ASV) is widely used in voice-based security mechanisms. It involves accepting or rejecting a person identification based on the individual's voice, a unique biometric feature. However, it faces many challenges and is vulnerable to direct or indirect attacks. Spoof voice detection is also an important component in secure voice authentication systems. Unfortunately, there is no spoof detection system using Myanmar language dataset. Spoof detection systems are important for many languages, including Myanmar, as they prevent fraud and misinformation, maintain trust, cultural and linguistic relevance, etc. Therefore, this paper proposes a Myanmar spoof voice dataset called UCSYSpoof, which contains both real and spoofed speech signals. End-to-end speech synthesis, vocoder-based speech reconstruction, and voice conversion techniques were used to generate the spoof speech based on 12,000 genuine speech signals. To demonstrate the impact of proposed dataset, a simple spoof detection model is implemented using long short-term memory (LSTM) and convolutional neural network (CNN) classifiers with linear frequency cepstral coefficients (LFCC) and Mel frequency cepstral coefficients (MFCC) features. Based on the empirical results, using CNN with LFCC and MFCC features achieves the comparable results on proposed dataset. The results show that the detection model has F1-score of 0.99 and an equal error rate (EER) of 0.004, respectively.**

***Index Terms*—ASV, spoof detection, UCSYSpoof, LFCC, MFCC, CNN, speech synthesis, voice conversion.**

I. Introduction

Automatic Speaker Verification (ASV) is a voice-based biometric system that aims to identify the authentication of a real speaker [14]. Spoofing is the act of disguising a communication or identity to make it appear to be associated with a trusted source. Identical twins, impersonation, voice conversion (VC) and text-to-speech (TTS) can be vulnerable in various spoofing attacks. The main goal of spoofed voice detection is a voice-based security mechanisms [17].

To observe the fake or simulated voice, this paper considers a spoofed speech detection system. Many studies have been conducted to detect real or fake speech. As far as we know, many acoustic features are adopted in spoofed speech detection tasks, including power spectrum, Mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), pathological features, etc. Furthermore, there are many classifiers, such as Gaussian Mixture Models (GMM), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), etc. [2].

In literature, ThaiSpoof dataset was proposed [9]. Their dataset was generated using the synthesis speech, pitch shifting and modifying fundamental frequency approaches. To demonstrate the performance of detection process, a simple CNN with LFCC features were utilized. The voice anti-spoofing dataset HABLA was launched for Spanish, including Argentinian, Colombian, Peruvian, Venezuelan, and Chilean accents [13]. The generation of spoof samples focused on six different strategies: three VC algorithms: StarGAN, CycleGAN, Diffusion, a TTS system, and two TTS-VC combinations (TTS-StarGAN and TTS-Diff). For detection, LFCC and light CNN architectures were used. The effectiveness of pathological features in spoofing detection was introduced [1]. Six pathological features were proposed, namely jitter, shimmer, normalized noise energy, glottal to noise excitation ratio, harmonic to noise ratio (HNR), and cepstral harmonic to noise ratio (CHNR). These pathological features were combined with traditional LFCC and MLP neural networks to achieve the better performance in detection process. The dynamic acoustic features, namely constant-Q cepstral coefficients (CQCCs) were proposed [5]. The authors pointed out that these features are more desirable for the spoof detection task because they have variable resolution in both time and frequency domains. GMM and DNN are used to train as classifiers. Moreover, a novel human log-likelihoods (HLLs) scoring method was proposed.

In this study, we propose the Myanmar Spoof voice Dataset (UCSYSpoof) specifically designed for spoof detection task. The generated speech is converted through five approaches:

end-to-end speech synthesis, vocoder-based speech reconstruction, using pre-trained model of FreeVC, trajectory GMM-based and differential GMM-based voice conversion techniques. This dataset enables more focused research in low-resource languages like Myanmar, which is not commonly covered in existing spoof detection datasets. Additionally, a simple spoof detection model is implemented to demonstrate the usefulness of this dataset. We evaluated the performance based on two classifiers CNN and LSTM using two feature extraction techniques-LFCC and MFCC. This paper is part of the ASEAN-IVO 2023 project "Spoofing Detection for Automatic Speaker Verification" which leads to facilitate the effective detection of spoofing attacks in Myanmar language.

The paper is structured as follows. Section 2 depicts the detailed development of spoofed dataset. Section 3 explains the implementation of the genuine or spoof detection model. Section 4 presents the experimental results and discussion, and finally, Section 5 concludes this study.

II. UCSYSpoof: Dataset Construction

This section provides the development of spoofed voices dataset. UCSYSpoof is a dataset for spoof detection in automatic speaker verification tasks which contains both genuine and spoofed utterances. The real or genuine subset consists of 12,000 utterances from three female speakers (4,000 utterances for each). The spoof dataset contains 71,932 utterances generated by five different approaches. It is constructed by speech synthesis, vocoder using parallel WaveGAN and HiFi-GAN, pre-trained voice conversion using FreeVC, GMM-based and Diffenential GMM-based voice conversion techniques. The detailed statistics of each technique in the UCSYSpoof dataset are expressed in Table I.

TABLE I
Detailed Statistic of UCSYSpoof Dataset

Label	Subset Type	No. of Speaker	No. of Utterance
Genuine	Genuine	3 (4K utts/each)	12,000
Spoofed	Vocoder-based [3]	3	23,932
	FreeVC-based	3	24,000
	Text-to-speech [3]	1	8,000
	GMM VC	2	8,000
	GMM DIFFVC	2	8,000

A. Genuine Dataset

The genuine dataset comes from the Basic Travel Expressions Corpus (BTEC), which is available for multiple languages including Myanmar [4]. This corpus is a textual multilingual corpus covering the travel domain. The phonetically balanced corpus is carefully constructed by selecting from the BTEC data. Here, three female native speakers participated in the recording, and each speaker recorded 4K sentences. The genuine dataset contains a total of 12,000 utterances and takes about 18.5 hours. The utterance is in wav file format, mono, 16 kHz sampling rate and 256 kbps bit rate.

B. Vocoder-based Dataset

For the preparation of vocoder based dataset, we used two Generative Adversarial Network (GAN) based neural vocoders, namely Parallel WaveGAN and HiFi-GAN vocoders which are specifically trained on the Myanmar language. The Parallel WaveGAN [15] is a distillation-free, lightweight, and rapid waveform generation method and it achieves realistic waveform synthesis by jointly optimizing an adversarial loss in the waveform domain and a multi-resolution short-time Fourier transform (STFT) loss, enabling the vocoder to produce high-quality speech without relying on complex probability density distillation methods. HiFi-GAN [8] is composed of one fully convolutional neural network based generator and two discriminators: multi-scale and multi-period discriminators, which can give efficient and high-fidelity speech synthesis. The 12,000 utterances of the three female speakers are reconstructed by applying the Parallel WaveGAN and HiFi-GAN vocoders, and generated 23,932 speeches are used as the fake speeches.

C. FreeVC-based Dataset

FreeVC-based dataset shown in TABLE I is generated by applying FreeVC [6], a text-free one-shot voice conversion system. FreeVC uses a pre-trained WaveLM [16] for extracting content information by imposing an information bottleneck without text annotation and then follows the end-to-end architecture of VITS. The six combinations of three female speakers such as speaker 1 as the source and speaker 2 as the target are prepared for generating converted speeches and 24,000 speeches are generated.

D. Text-to-Speech Dataset

Myanmar end-to-end speech synthesis based on Tacotron2 [7] with two waveform generation techniques was employed to generate high-quality speech. Tacotron2 model with traditional Griffin-Lim and trained HiFi-GAN vocoders are utilized for text to speech generation of 4,000 sentences with phonemes. At the synthesis time, the Tacotron2 model converts the input phoneme sequences to the corresponding mel-spectrograms and each Griffin-Lim and HiFi-GAN based vocoder generates the speech waveform according to the given mel-spectrograms and 8,000 synthesized speeches are produced.

E. Voice Conversion based on GMM

GMM-based VC methods utilize the parallel speech utterances of the source and target speakers. This section focuses on two typical conversion methods. The first is the maximum likelihood parameter generation (MLPG) based on GMM considering the global variance (GV), and the second is the vocoder-free VC using log-spectral differentiation (DIFFVC).

The goal is to learn a mapping function from training observations that can be used to map any test features of the source speech (including prosodic and spectral features) to the acoustic space of the target speech [10] [11].

In typical parallel voice conversion, time-aligned features of source and target speakers are required. This work uses Mel-frequency cepstrum as the spectral feature representation and attempts to convert the features of the source speaker into the features of the target speaker.

In GMM-based conversion, a joint feature matrix, which consists of the time-aligned between source and target features is needed. The following steps are performed in the training process [11]:

- Compute acoustic features including aperiodicity, F0 and mel cepstrum for each speaker
- Calculate acoustic feature statistics
- Use Dynamic Time Warping (DTW) to achieve time alignment between source and target feature vectors
- GMM modeling.

Figure 1 demonstrates a conversion process of VC based on GMM, and Figure 2 shows a conversion process of differential voice conversion (DIFFVC) based on differential GMM (DIFFGMM).

First, aperiodicity, F0, and Mel-cepstrum features are derived from the source utterance. In the GMM-based VC, F0 is linearly transformed frame by frame and the Mel-cepstrum is converted into the cepstrum of target speaker using MLPG. Afterwards, the GV post-filter is enforced to the converted Mel-cepstrum. Finally, based on the transformed F0 and the converted Mel-cepstrum, excitation generation and Mel-log spectral approximation (MLSA) filter are used to generate the converted speech.

In DIFFVC method, the parameters of trained GMM model are modified from the joint probability density of source and target features to a joint probability density of the source feature and a feature differential between the source and target features. Then, the MLPG and GV post-filter are used to estimate the mel-cepstral derivatives from the source mel-cepstrum. At last, the converted speech is generated by filtering the Mel-cepstrum differential, where the MLSA filter is also utilized [11].

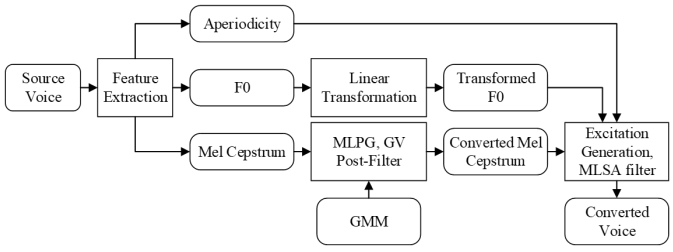

Fig. 1. Conversion process of VC based on GMM.

III. Implementaton of Spoof Detection Model

This section provides the implementation of a spoof detection model that classifies the genuine or spoofed voice.

A. Dataset

In this experiment, 71,932 utterances are used to detect genuine or spoofed utterances, of which 12,000 utterances

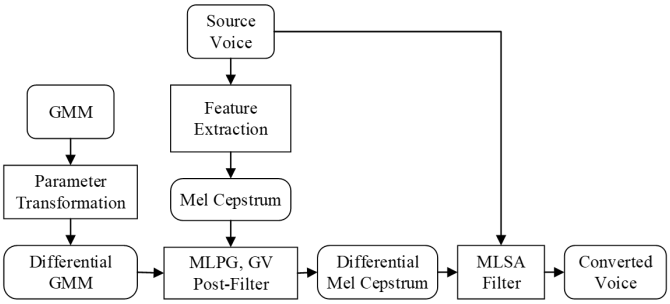

Fig. 2. Conversion process of DIFFVC based on differential GMM.

are genuine and the rest are spoofs from various spoofing techniques. During the detection process, we randomly generate training and test utterances. 75% of the entire dataset is used for training and 25% is used for testing. The training set contains 67,146 audio files and the evaluation set contains 16,785 audio files. In other words, the training to test ratio is 4.0. The detailed statistics of data usage in the spoofing detection model are described in Table II.

TABLE II
Detailed Statistic of data usage in Spoof Detection

Label	Proportion	No. of Utterances
Training	75%	67,147
Testing	25%	16,785
Total		71,932

B. Experimental Conditions

The implementation of the spoofing detection model is based on two feature extraction techniques, LFCC and MFCC. Two classifiers, CNN and LSTM are also used.

Feature extraction techniques extract information from speech frames. LFCC and MFCC are commonly used acoustic features in many speech processing fields. The detailed calculation steps of MFCC are as follows [12] [18]. First, the speech signal is pre-emphasized and chunked into overlapping frames. Each frame is weighted using a Hamming window to minimize high-frequency effects and eliminate edge effects. Then, apply a First Fourier Transform (FFT) to each frame to convert the time domain to the frequency domain (spectrum). A Mel-scaled filter bank is performed on the output of the FFT to capture the energy distribution in different frequency bands. After that, the logarithmic function is taken onto energy. Finally, a discrete cosine transform (DCT) is applied to log Mel spectrum energy. MFCC and LFCC are almost the same in terms of coefficient extraction part. The only difference is in the filter bank process. LFCC filter bank coefficients cover all frequency ranges equally and consider them to have equal importance. In this experiment, 40 feature dimension were used in the MFCC and LFCC feature extraction process and the input MFCC/LFCC dimension was 128x40.

Shallow convolutional neural networks (CNN) generally refer to a network with simple architectures, typically composed of a few convolutional layers, pooling layers, and fully connected layers. The network structure is simple and has fewer network parameters, so training the model takes up less computing resources and memory. In CNN comprising of four convolutional layers, max pooling layers with a pool size of 2x2, and two fully connected layers. The output of last convolutional layer is flattened and forwarded to two connected layers. The output layer is a fully connected layer with two neurons, one for each class (genuine or spoofed), using a linear activation function to output the probability distribution between the two classes. The ReLU activation function is enforced. In our experiments, we use a batch size of 128, the number of random seeds is 42, and the maximum number of epochs to train is 50.

Long Short-Term Memory (LSTM) is a type of neural network architecture for processing sequences of data. LSTM can learn dependencies over long sequences and alleviate issues such as vanishing gradients that may occur with traditional recurrent neural networks (RNNs). A simple LSTM architecture was applied in this study. The dimension of the input feature size in the LSTM is 128x40. A bidirectional LSTM with two stacked LSTM layers is used to enhance the performance of the model in capturing temporal patterns. The number of cells in the LSTM layer is 972. After processing by the LSTM layer, the output can be passed through a dense layer. This layer helps in converting the output into the required format. The output layer provides the class of genuine or spoofed. In this experiment, the dropout value is 0.01 and the ReLU activation function is employed. The number of epoches, random seeds and batch size are the same as above CNN model.

C. Evaluation Metrics

The performance of real or fake detection system is usually measured using the equal error rate (EER). The EER is a metric used in biometric security systems to measure the effectiveness of a system in correctly identifying an individual. It is the point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal. Lower EER values indicate better performance in the spoofing detection task. Additionally, other evaluation metrics such as accuracy and F1 score are used to evaluate the system.

IV. Experimental Results

Four pairs of experiments based on feature extraction methods and classifiers are investigated to showcase the utilization of proposed UCSYSpoof dataset in spoof detection task. These are LFCC features with CNN classifier, LFCC features with LSTM classifier, MFCC with CNN classifier, and MFCC with LSTM classifier. EER, accuracy, and F1 score are used as the evaluation metrics.

First, the CNN model using LFCC features gave an EER of 0.016, and then the LSTM model using LFCC features yield an EER of 0.037, respectively. After that, the MFCC features with two classifiers were studied. Using the CNN classifier, the model achieved an EER of 0.004 and while using the LSTM classifier, this experiment achieved an EER of 0.008. Based on the experimental results, the best detection model with MFCC and CNN classifier achieves the F1-score of 0.99 and EER of 0.004, respectively.

TABLE III
Detailed Statistic of data usage in Spoof Detection

Features	Classifiers	Accuracy(%)	F1 score	EER
LFCC	CNN	99.70	0.989	0.016
LFCC	LSTM	97.97	0.931	0.037
MFCC	CNN	99.82	0.994	0.004
MFCC	LSTM	99.40	0.979	0.008

In this study, LFCC and MFCC features provide comparable performance in the Myanmar spoof detection task (except for the combination of LFCC and LSTM). MFCC is consistent with human hearing and is robust to speech variations. It can capture the perceptual characteristics to distinguish real speech from manipulated forms. LFCC renders more detailed frequency resolution, which helps to identify the subtle differences between gnuine and spoofed speech, but it may not be able to capture perceptual information as effectively as MFCC. In this exploration, both feature extraction techniques are applicable to detecting Myanmar spoofed speech signals and produce comparable results. Moreover, the CNN classifier has clear advantages over LSTM in terms of efficiency, ability to handle spatial patterns and complexity of computation. Table III shows the evaluation performance of spoof detection on the proposed UCSYSpoof dataset.

V. Conclusion

This paper highlighted on the construction of the Myanmar language UCSYSpoof dataset. The manipulated speeches are generated in five ways, end-to-end speech synthesis, vocoder-based conversion using parallel WaveGAN and HiFiGAN, using pre-trained FreeVC, and statistical GMM-based and differential GMM-based voice conversion techniques. To showcase the utilization of proposed dataset in spoofing detection, two acoustic feature extraction methods - MFCC and LFCC, are studied in this paper. These features are then classified using CNN and LSTM classifiers to detect whether the speech is real or fake. According to the experiments, LFCC and MFCC features yielded the comparable performance in spoof detection task while using CNN classifier. The results revealed that the detection model achieves an F1-score of 0.99 and EER of 0.0004, respectively. Existing dataset lacks sufficient diversity in terms of gender variability, dialects, and accent variations. The current dataset contains only female speakers. This fact may result in performing well only in specific situations and may not be applicable to all potential spoofing scenarios. A larger and more diverse dataset may lead to learning a wider range of variations and achieving better

accuracy and reliability. In future work, we will expand and improve the UCSYSpoof dataset using state-of-the-art voice conversion techniques.

ACKNOWLEDGMENT

The authors would like to evince their deep and sincere gratitude to the "Spoofing Detection for Automatic Speaker Verification" project under the ASEAN Institutional ICT Virtual Organization (ASEAN IVO) https://www.nict.go.jp/en/asean_ivo/ and the National Institute of Information and Communications Technology (NICT) https://www.nict.go.jp/en/ for their financial support in producing the content of this paper.

REFERENCES

[1] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon and M. Unoki, "Deepfake-speech detection with pathological features and multilayer perceptron neural network", In Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 2182-2188, IEEE, October 2023.

[2] 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.134018-134028, 2022.

[3] 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]

[4] H.M.S. Naing, A. M. Hlaing, W. P. Pa, X. Hu, Y. K. Thu, C. Hori, and H. Kawai, "A Myanmar large vocabulary continuous speech recognition system", In Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp.320-327, IEEE, December 2015.

[5] H. Yu, Z. H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features" IEEE transactions on neural networks and learning systems, Vol. 29(10), pp.4633-4644, 2017.

[6] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion", In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p.1-5, IEEE, 2023.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, et.al, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions", In IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779-4783, IEEE, 2018.

[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] K. Galajit, T. Kosolsriwiwat, M. Unoki, C. O. Mawalim, et al. "ThaiSpoof: A Database for Spoof Detection in Thai Language. 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), pp. 1-6, IEEE, November 2023.

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

[11] K. Kobayashi, and T. Toda, "sprocket: Open-Source Voice Conversion Software". In Odyssey, pp. 203-210, June 2018.

[12] M. Mohammadi, H.R.S. Mohammadi, "Robust features fusion for text independent speaker verification enhancement in noisy environments", In Iranian Conference on Electrical Engineering (ICEE), pp. 1863-1868, IEEE, May 2017.

[13] P. A. T. Flórez, R. Manrique, and B. P. Nunes, "HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing", INTERSPEECH 2023, 20-24 August 2023, Dublin, Ireland.

[14] P. Gupta, H. A. Patil, and R. C. Guido, "Vulnerability issues in Automatic Speaker Verification (ASV) systems", EURASIP Journal on Audio, Speech, and Music Processing, p.10, 2024(1).

[15] 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.

[16] 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.

[17] T. Kinnunen, Z. Wu, E. Nicholas Evans, and J. Yamagishi, "Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database", 2018.

[18] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. E. Wilson, S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition", In IEEE workshop on automatic speech recognition & understanding, pp.559-564, IEEE, December 2011.