UCSYSpoof: A Myanmar Language Dataset for Voice Spoofing Detection

Win Pa Pa, Aye Mya Hlaing, Myat Aye Aye Aung
Kasorn Galajit, Candy Olivia Mawalim

Presented By :

Hay Mar Soe Naing

University of Computer Studies, Yangon

Myanmar

Date : 17th, Oct 2024

Outlines

- Introduction
- Problem Statements
- Objectives
- UCSYSPOOF: Dataset Construction
- Implementing Spoof Detection Model
- Experimental Results
- Result Discussion
- Conclusion

Introduction

- Automatic Speaker Verification (ASV) is Accepting or rejecting a person identification based on the individual's voice, a unique biometric feature.
- Used in voice-based security mechanisms.
- Vulnerable to malicious attack to spoof the system.

Introduction (Cont’d.)

- Spoof voice detection is an important component in secure voice authentication.
- Some key applications:
  - Biometric Security Systems
  - Virtual Assistants and Smart Devices
  - Law Enforcement and Forensics

Challenges in Spoofed Detection

- ❑ Evolving Attack Techniques
- ❑ Variability in Voice
- ❑ Data Quality and Diversity
- ❑ Resource Constraints
- ❑ Real-Time Processing

Problem Statements

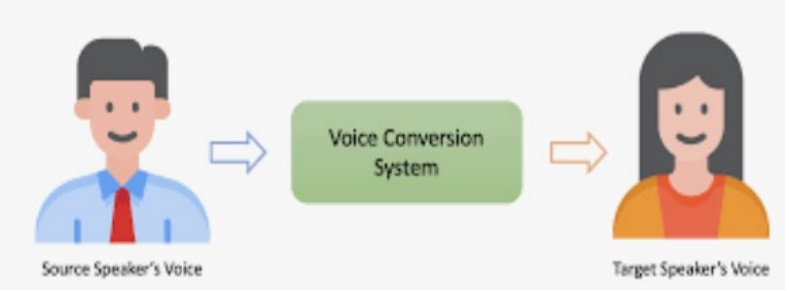
- There is no spoof detection system using Myanmar language dataset.
- To prevent fraud and misinformation, maintain trust, cultural and linguistic relevance, etc.

Objectives

- To propose Myanmar Spoof voice Dataset (UCSYSpoof) specifically designed for spoof detection task.
- To implement a simple spoof detection model to demonstrate the usefulness.
- To evaluate the performance based on two classifiers CNN and LSTM using LFCC and MFCC.

UCSYSPOOF: Dataset Construction

- Spoof dataset is generated by using five different approaches.
  - End-to-End speech synthesis
  - Vocoder using parallel WaveGAN and HiFi-GAN,
  - Pre-trained voice conversion using FreeVC
  - GMM-based voice conversion
  - Differential GMM-based voice conversion



UCSYSPOOF Dataset

Genuine Dataset

- Genuine dataset comes from **Basic Travel Expressions Corpus (BTEC)**
- A textual multilingual corpus covering the travel domain.
- **Three female native speakers** participated in the recording.
- Each speaker recorded 4K sentences (3 x 4K = 12,000 utts, takes about 18.5 hours)
- Utterance is wav file format, mono, 16 kHz sampling rate and 256 kbps bit rate.

(1) End-to-End Speech Synthesis

- Employed Myanmar end-to-end speech synthesis based on Tacotron2.
- Converts the input 4,000 phoneme sequences to corresponding mel-spectrograms.
- Utilized two waveform generation techniques.
  - traditional Griffin-Lim
  - trained HiFi-GAN vocoders
- Produced **8,000 synthesized speeches.**

(2) Vocoder-based Dataset

- Neural vocoders, namely Parallel WaveGAN and HiFi-GAN vocoders are specifically trained.
- **12,000 utterances** of three female speakers are reconstructed by **Parallel WaveGAN**.
- **23,932 utterances** are generated by **HiFi-GAN vocoders**.

(3) FreeVC-based Dataset

- Apply FreeVC, a text-free one-shot VC system.
- Uses a pre-trained WaveLM for extracting content information and then follows the end-to-end architecture of VITS.
- Six combinations of three female speakers. (i.e., Speaker 1 as source and Speaker 2 as target)
- Totally, **24,000 speeches** are generated.

(4) GMM-based Voice Conversion

- Steps to perform in the training process:
  - 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.

(4) GMM-based Voice Conversion (Cont'd.)

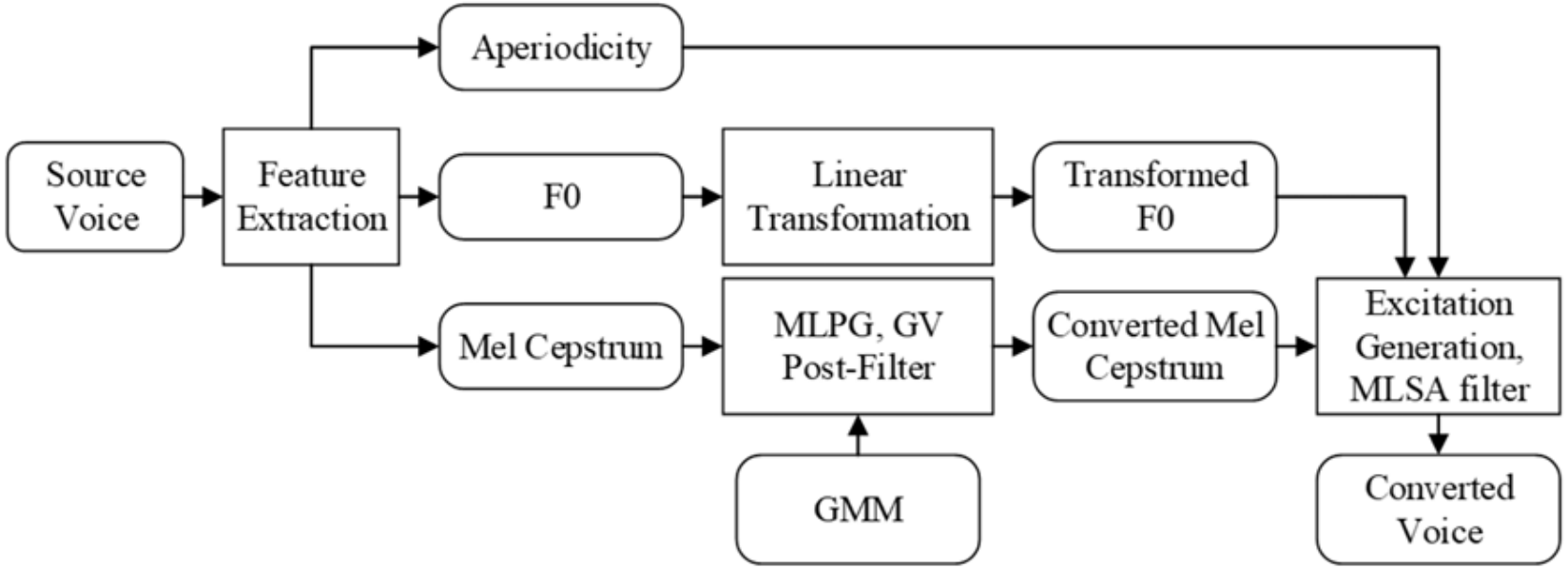


Fig 1 : Conversion process of VC based on GMM.

(5) Differential GMM-based Voice Conversion

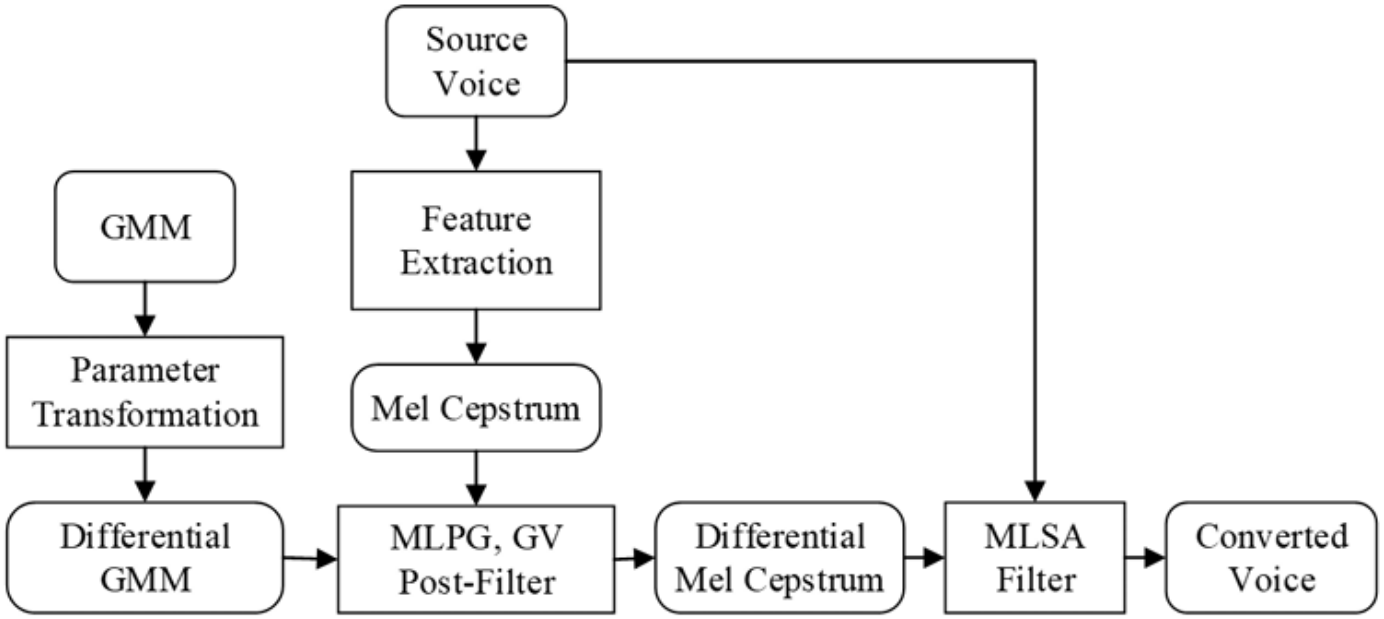


Fig 2 : Conversion process of DIFFVC based on differential GMM

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	23,932
	FreeVC-based	3	24,000
	Text-to-speech	1	8,000
	GMM VC	2	8,000
	GMM DIFFVC	2	8,000

✓ Totally, UCSYspoof dataset contains 71,932 utterances.

Implementing Spoof Detection Model

- Split train and evaluation subsets, the training to test ratio is 4.0.

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

Evaluation Metrics

- The Equal Error Rate (EER) is a metric used in real or fake detection system.

$$EER = \frac{FAR + FRR}{2}$$

- Accuracy
- F1 Score

Experimental Results

- Showcase the utilization of UCSYSpoof dataset in spoof detection.
- LFCC and MFCC features provide comparable performance.

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

✓ MFCC and CNN classifier achieves the lowest EER of 0.004.

Discussion

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 genuine and spoofed speech.

It may not be able to capture perceptual information as effectively as MFCC.

✓ *Both are applicable to detect spoofed signals and produce comparable results.*

Discussion (Cont’d.)

CNN classifier has clear advantages over LSTM in terms of efficiency, ability to handle spatial patterns and complexity of computation.

Limitations

- Existing dataset lacks sufficient diversity in terms of gender variability.
- The current dataset contains only female speakers.
- Perform well only in specific situations and may not be applicable to all potential spoofing scenarios.

Conclusion

- Highlighted on the construction of UCSYSpoof dataset.
- Manipulated speeches are generated in five ways
- Used CNN and LSTM classifiers to detect whether the speech is real or fake.
- LFCC and MFCC features yielded the comparable performance in spoof detection task while using CNN classifier.
- Combination of MFCC and CNN achieves the lowest EER of 0.004.

Thank You