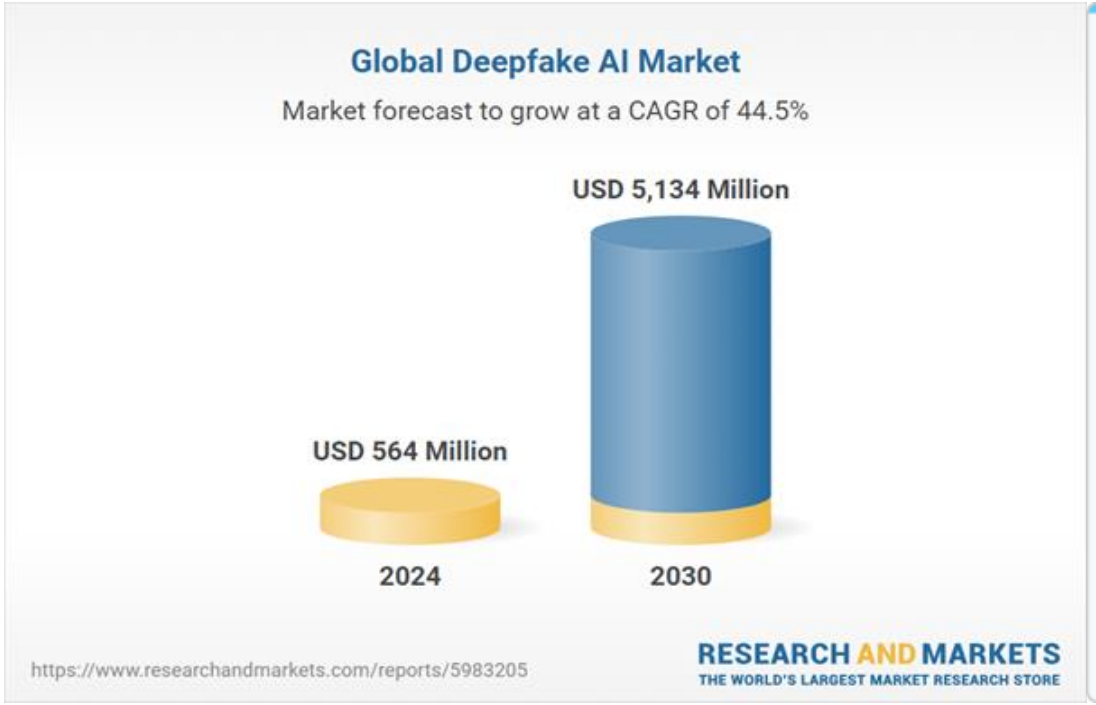
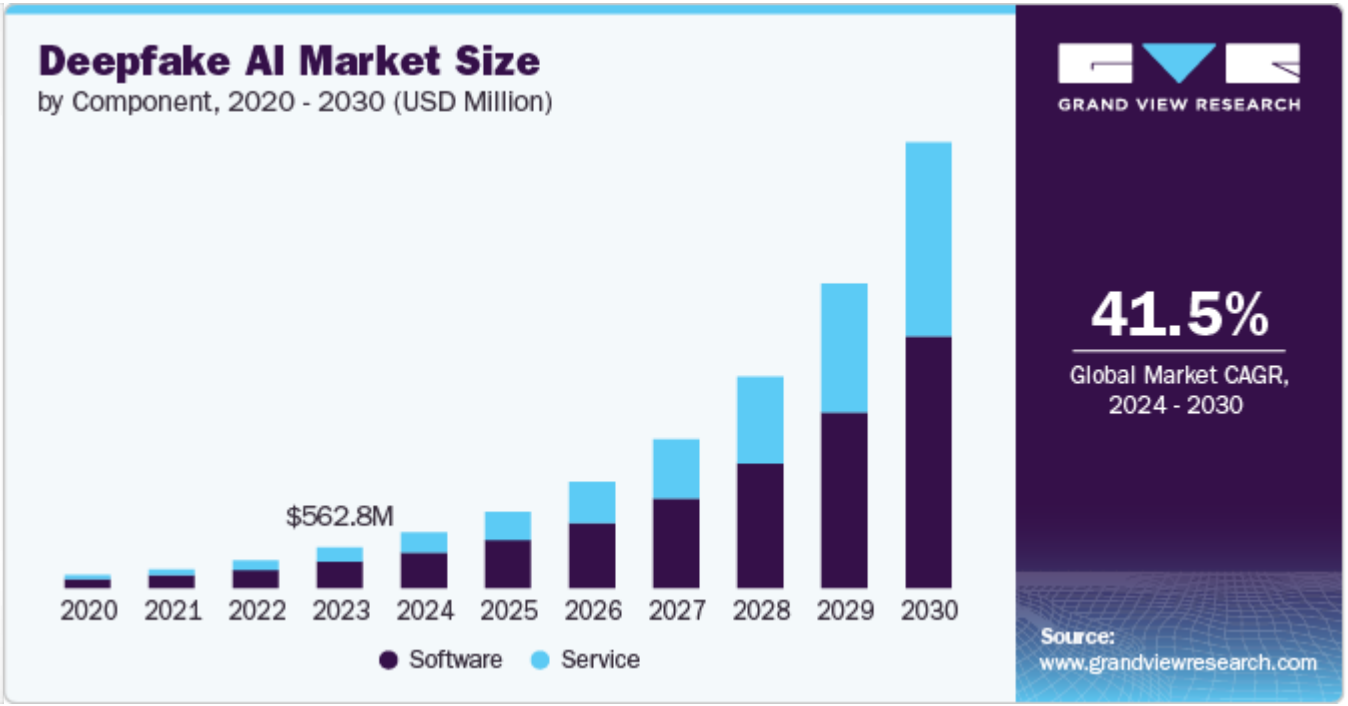
ThaiSpoof : An extension to current methods and database catering advanced spoof detection

Puntika Leepagorn, Khaing Zar Mon, Kasorn Galajit, Jaya Shree Hada, Navod Neranjan Thilakarathne and Jessada Karnjana

19th iSAI-NLP, Pattaya, Thailand Nov.11-15 2024

Deepfake AI Market





According to Google Trends, **searches for "free voice cloning software" rose 120 percent** between July 2023 and 2024.

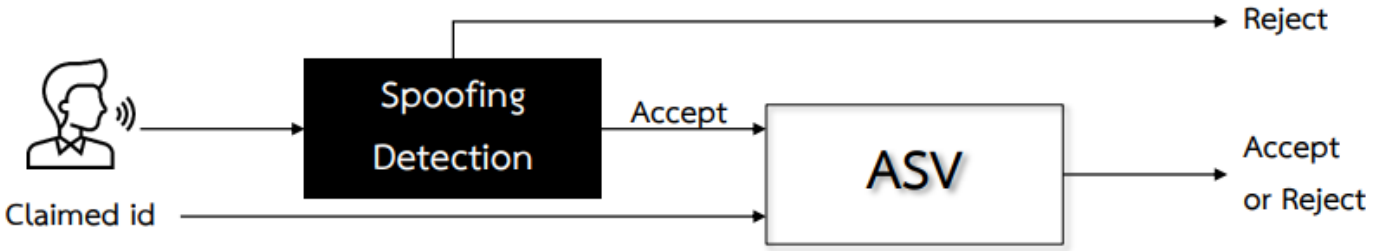
In January, a robocall impersonating U.S. President Joe Biden went out to New Hampshire voters, advising them not to vote in the state's presidential primary election.

- **Three seconds of audio** is sometimes all that's needed to produce an **85 percent voice match** from the original to a clone.
- According to a McAfee survey, **70 percent of people said they aren't confident that they can tell the difference between a real and cloned voice.**

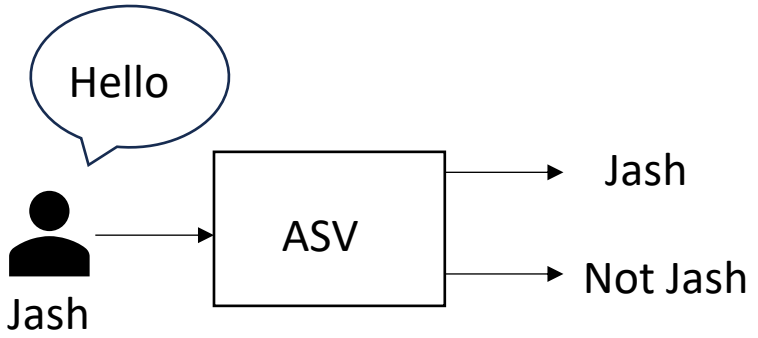
Spoof Detection for ASV System

Spoofing refers to a presentation attack using fake biometrics for a valid person.

Spoof Detection



Automatic Speaker Verification



AVAILABLE SPOOF DATASET

Dataset	Year	Accessibility	Language	Spoof Type	Environment
ASVSpoof 2015	2015	Yes	English	TTS, VC	Clean
ASVSpoof 2019 - LA	2019	Yes	English	TTS, VC	Clean
FoR - original	2019	Yes	English	TTS	Clean
ASVSpoof 2021 - LA	2021	Yes	English	TTS, VC	Codec
ASVSpoof 2021 - DF	2021	Yes	English	TTS, VC	Codec
FMFCC-A	2021	Yes	Chinese	TTS, VC	Noisy, Codec
WaveFake	2021	Yes	English, Japanese	TTS	Clean
ITW	2022	Yes	English	TTS	Noisy
TIMIT - TTS	2022	Yes	English	TTS	Noisy, Codec
CFAD	2023	Yes	Chinese	TTS, VC	Noisy, Codec
ThaiSpoof - 2023	2023	Yes	Thai	TTS	Clean, Noisy
MLAAD	2024	Yes	23 Languages	TTS	Clean

PREVIOUS WORK

THAI SPOOF DATA SET BY KASORN ET AL.

01

TEXT-TO-SPEECH: TTS

02

FUNDAMENTAL FREQUENCY MODIFICATION:F0

03

PITCH SHIFTING

DATABASE CONSTRUCTION

ADDING SOURCE

- Adding genuine voice from “Common Voice”
- Currently have Common Voice + LOTUS

SCREENING TASK

- The dataset is screen before modification and synthesizing.
- Cut off some low-quality speech and the speeches whose length is shorter than 5 seconds.

INTRODUCING MMS-TTS

- Add new data set from new technique which is “Massively Multilingual Speech Model (MMS)” developed by Meta AI

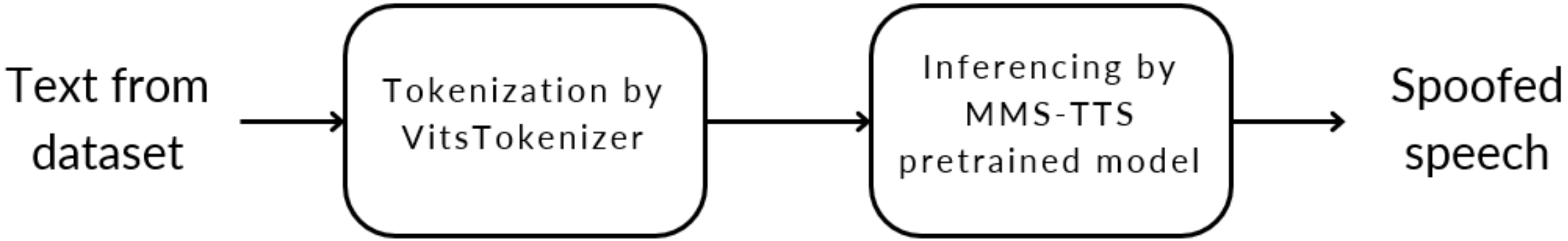
MMS-TTS

is built on advanced machine learning techniques, particularly deep learning, to replicate human-like speech from text inputs.

Key Features:

- Multilingual Capacity
- Natural Speech Output
- Advance Language Processing

GENERATING SPOOF SPEECH USING MMS-TTS



NEW SPOOF DATASET

Label	Dataset Type	Degree	Utterances
Genuine	Genuine Dataset	-	4,583
Spoof	Text to Speech - TTS	-	4,583
	F0 Modification	10 ch/oct	4,583
		40 ch/oct	4,583
		160 ch/oct	4,583
		320 ch/oct	4,583
	Pitch Shifting	+ 4%	4,583
		+ 10%	4,583
		+ 20%	4,583
		-4%	4,583
		-10%	4,583
		- 20%	4,583
	Massively Multilingual Speech - MMS	-	4,583

EXPERIMENT SET UP

utilized a **CNN model** to train and demonstrate performance using two distinct text-to-speech datasets: the **VAJA dataset and the MMS dataset.**

2 feature extraction

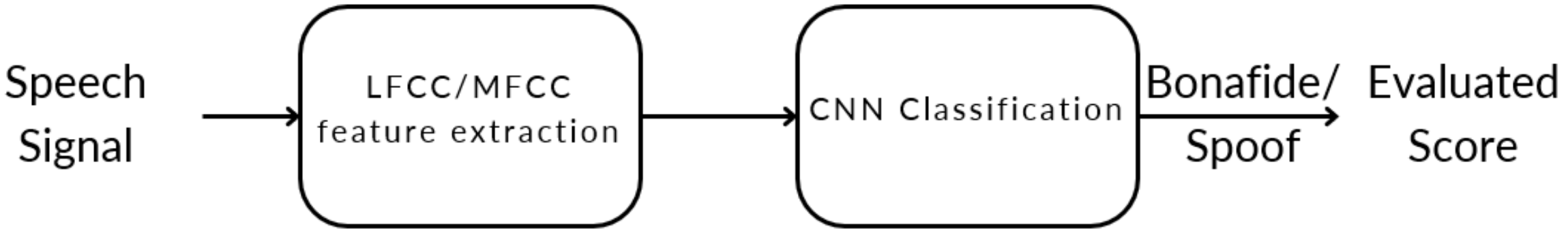
- LFCC

Linear Frequency Cepstral Coefficient

- MFCC

Mel-Frequency Cepstral Coefficient

SPOOF DETECTION MODEL DIAGRAM



EVALUATION MATRIX

01

EQUAL ERROR RATE (EER)

$$FAR = \frac{FP}{FP + TN}$$

$$FRR = \frac{FN}{FN + TP}$$

$$EER = FAR = FRR$$

02

ACCURACY

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

03

F1 SCORE

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

RESULTS

LFCC

Feature Extraction

Training Data	Test Data	EER (%)	Balanced Accuracy (%)	F1 Score
MMS + Genuine	MMS + Genuine	0.04	99.96	99.96
	VAJA + Genuine	2.98	97.02	96.93
VAJA + Genuine	MMS + Genuine	49.85	50.15	0.58
	VAJA + Genuine	0	100	100

MFCC

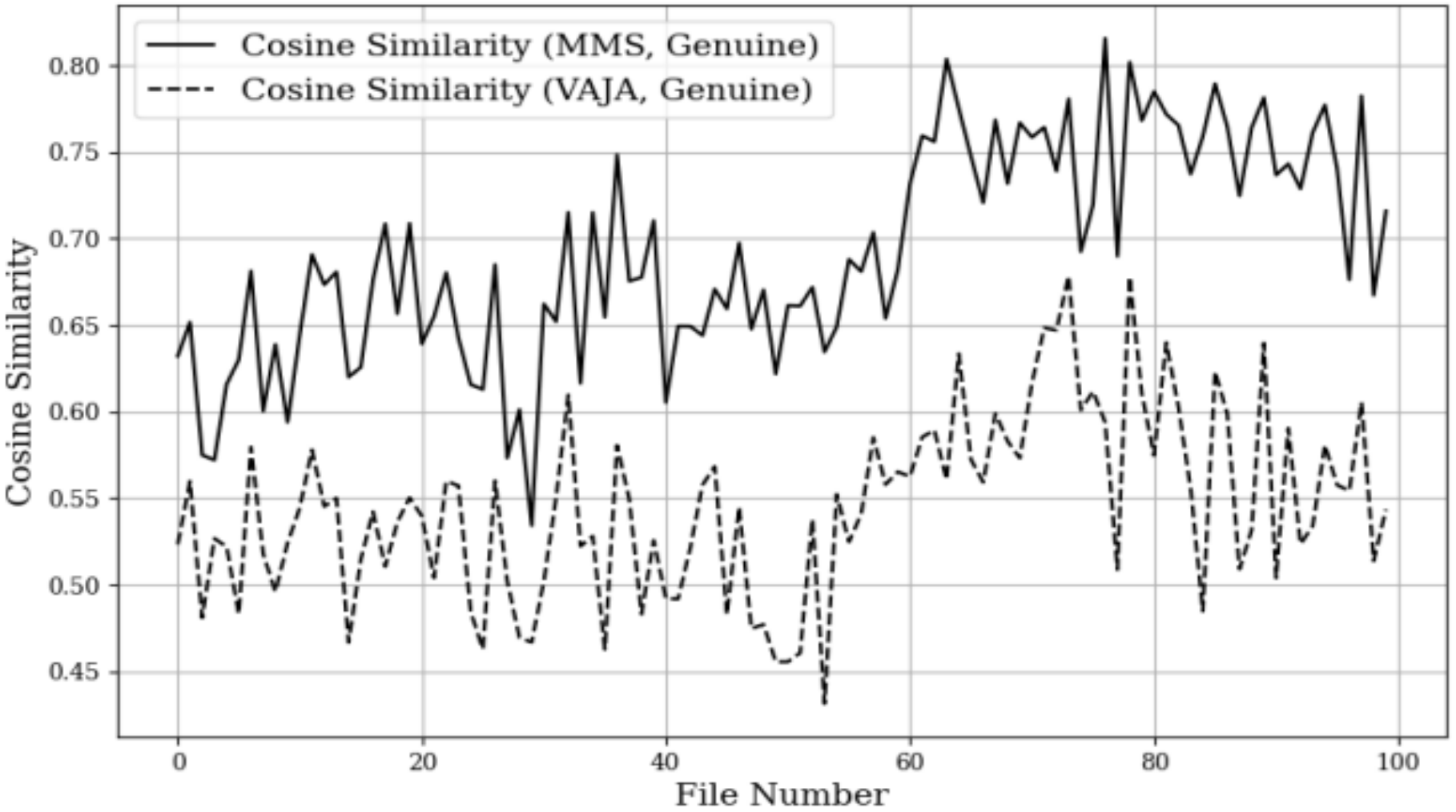
Feature Extraction

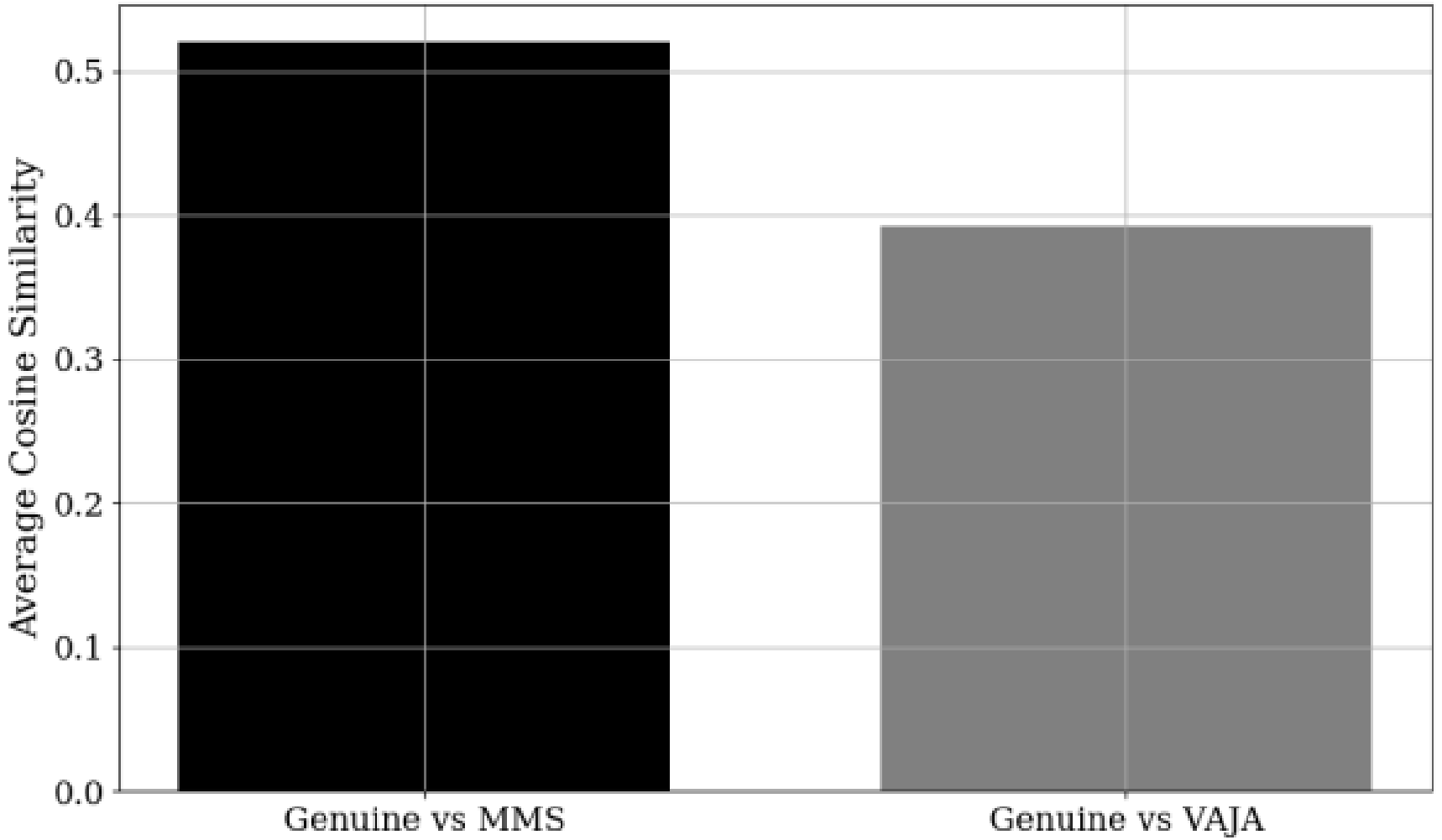
Training Data	Test Data	EER (%)	Balanced Accuracy (%)	F1 Score
MMS + Genuine	MMS + Genuine	0.07	99.93	99.93
	VAJA + Genuine	19.49	80.51	75.82
VAJA + Genuine	MMS + Genuine	47.35	52.65	10.21
	VAJA + Genuine	0.03	99.97	99.97

MMS Versus VAJA

Experiment Set Up

1. randomly select 100 speech signals ID (10 utterances from 10 speakers)
2. pull the selected speed signal from Genuine, MMS, and VAJA dataset
3. calculate cosine similarity of LFCC feature between 2 pairs, (Genuine, MMS) and (Genuine, VAJA)
4. compare the similarity of synthesis voice datasets and genuine voice dataset


Cosine Similarity (MMS, Genuine)
Cosine Similarity (VAJA, Genuine)
Cosine Similarity
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0
20
40
60
80
100
File Number


Average Cosine Similarity
0.0
0.1
0.2
0.3
0.4
0.5
Genuine vs MMS
Genuine vs VAJA

Thank you

Any question or comment is welcome