IFIP WG Meeting

DeepVoice Detection: A Practical Approach

Souhwan Jung

2024.06.30

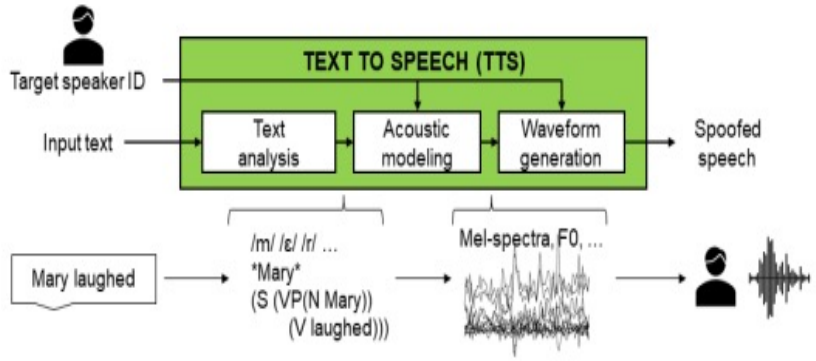
QUIZ : Which of the three is an AI-generated voice?

How is a DeepVoice generated?

- DeepVoice (Audio Deepfake) : DeepLearning + Voice
  - The creation of a DeepVoice is archived by using AI-powered TTS & VC technology
  - DeepVoice technology now enables the generation of deepfake audio in seconds
  - OpenAI Voice Engine requires only 15-seconds target voice sample for generating voice clones

TTS(Text-to-Speech)

- Convert any text into target voice



Diffusion

- Tortois-TTS
  - Two main technologies : Autoregressive Diffusion Decoder
  - Focuses on creating diverse, natural-sounding voices

Transformer

- Transformer-TTS
  - Parallel processing
  - Using Attention technology to learn array relationships

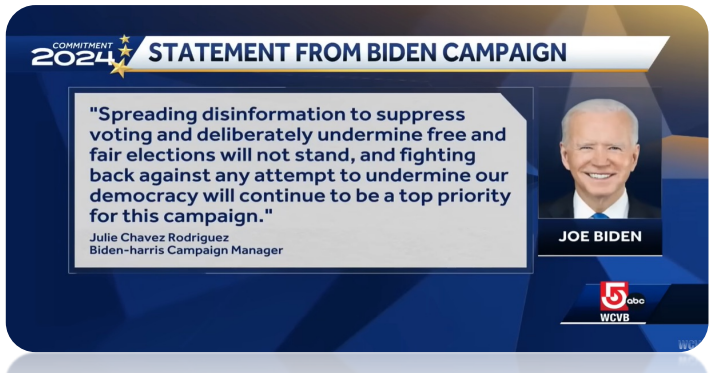
GAN

- GAN-TTS
  - Uses two competing networks to generate realistic synthetic speech.

Source: Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," in IEEE 2021

Fraud Cases using DeepVoice

- Recent fraud case using deepfake



DeepVoice Robocaller

- Someone using President Biden's voice
- Says "your vote makes a difference in November, not this Tuesday"

Fake Information/News

- Someone created a fake advertisement using Taylor Swift's image and voice
- This fake ad led to payments and caused harm to the fans

A DeepVoice Fraud Case in Korea

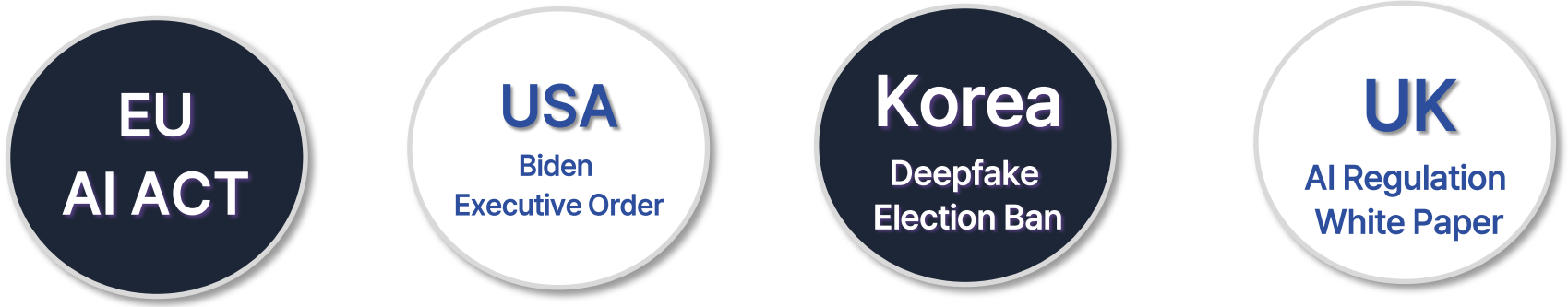
- The first voice phishing case using DeepVoice in Korea
  - Impersonating a daughter to her mother
  - Claiming kidnapped and asking for money to be sent
- Classified the mother's and daughter's speech with a detection model
  - The daughter's speech was false with an overall probability of 60% (individually, typically over 90%)
  - The mother's speech was classified to be less than 1% false

File Name	Fake (%)
1_딸_엄마.wav	87.98%
6_딸.wav	93.90%
7_엄마.wav	0.04%
12_딸_피싱범.wav	18.98%
13_딸_엄마.wav	39.02%
14_딸.wav	99.78%
15_딸_엄마.wav	0.41%
16_딸.wav	99.91%
17_딸.wav	0.01%
18_딸_피싱범.wav	0.05%
딸 목소리_전체모음.wav	60.94%

How can we Solve these Problems?

Approach 1 : Regulation of Watermarking on DeepFake

- Policies for regulating deepfakes in various countries



- IT Companies' Response to Generative AI Content

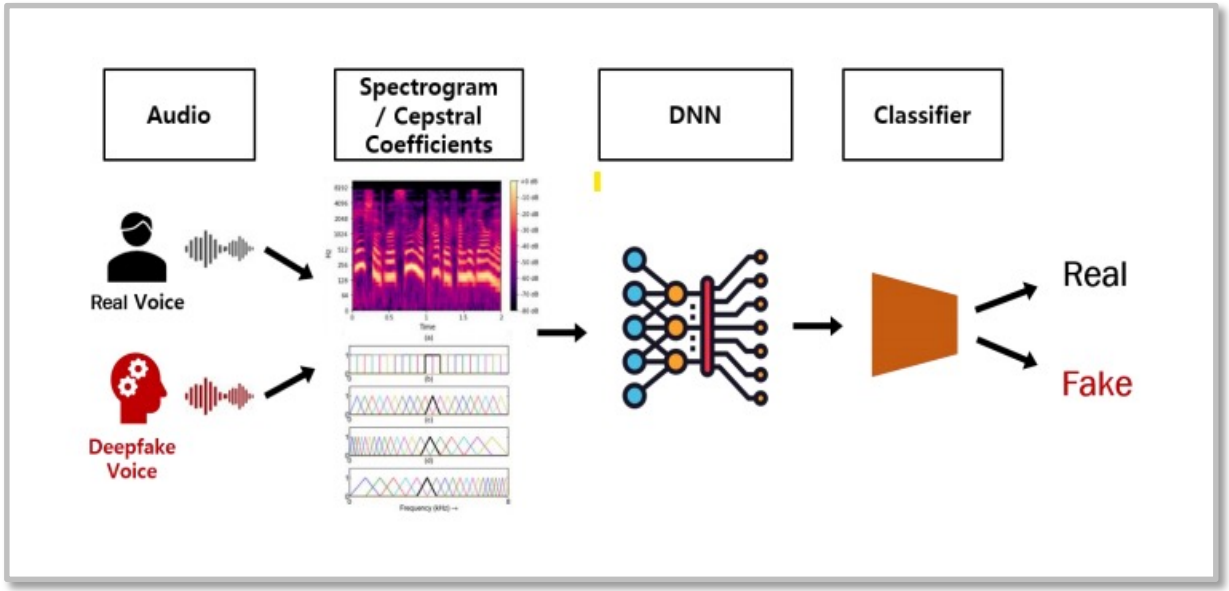
Company	Response
NAVER	Blocking Harmful Deepfakes in Real-time Using AI Filtering Technology '**GreenEye**'
Meta	Plans to **Identify AI-generated Content** on Instagram and Facebook
OpenAI	Inserting **Invisible Watermarks** in Images Created by the AI Image Generator 'DALL-E'

Approach 2 : DeepVoice Detection Technology

- Framework for DeepVoice Detection
  - General Framework

The diagram shows a process for detecting deepvoice

- Handcrafted Feature-based Models, DNN-based Models, End-to-End Models



Our works for DeepVoice Detection(1)

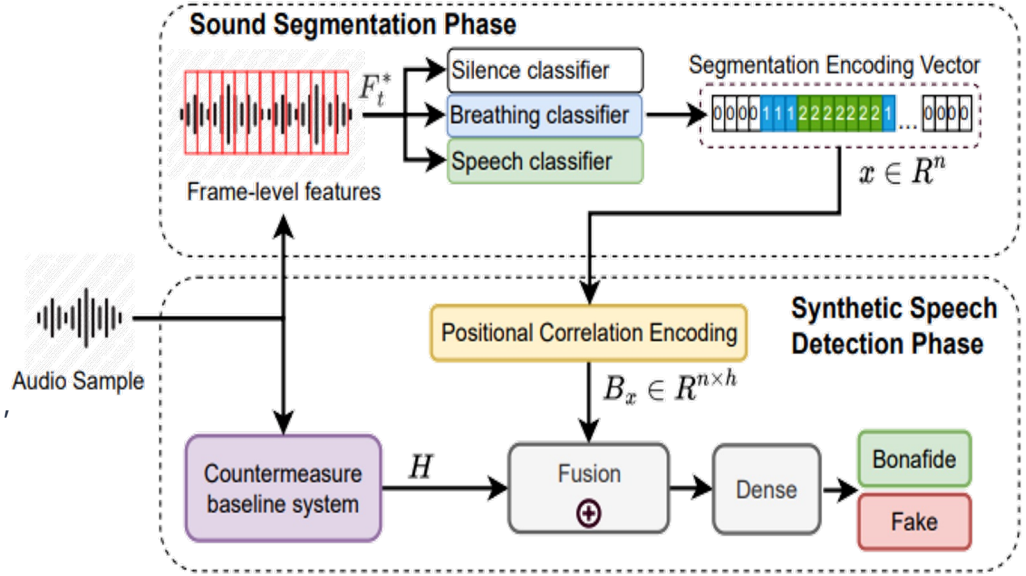
- BTS-E (Breathing-Talking-Silence Encoder)

  - TTS focuses on linguistics content

  - While training, all non-speech segments are removed

  - No biological signal in the synthesized speech waveform

  - Apply VAD to encode the positional of breathing, talking and silence signal at frame-level



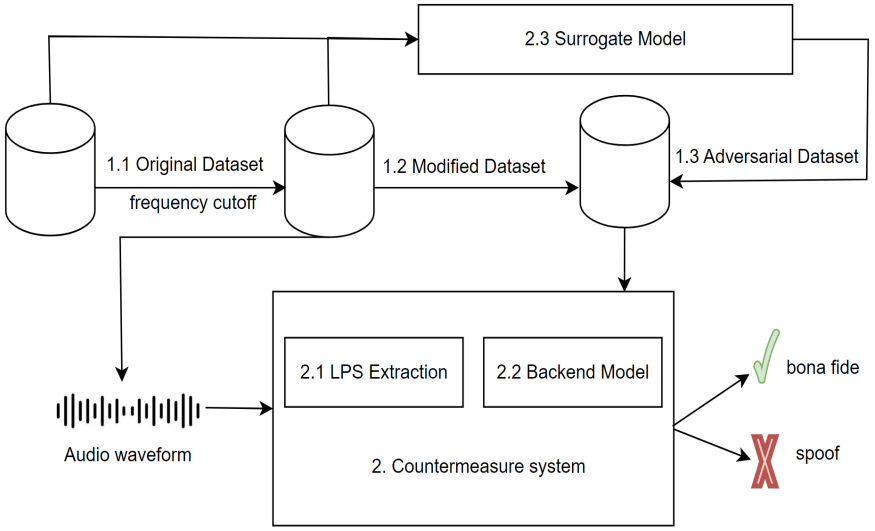
Source: T. -P. Doan, L. Nguyen-Vu, S. Jung and K. Hong, "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder," ICASSP 2023

Our works for DeepVoice Detection(2)

- Frequency cutoff technique for robust CM against Adversarial Attacks

  - Extract the subband only (remove high/low frequencies) since full bandwidth audio has noise
  - Train these with adversarial model (surrogate model) to add generated data to the dataset
  - Train the detections system with this dataset

- The model trained on subband is much more robust against adversarial attacks (up to 56% reduction in EER)



Source: L. Nguyen-Vu, T. -P. Doan, M. Bui, K. Hong and S. Jung, "On the Defense of Spoofing Countermeasures Against Adversarial Attacks," in IEEE Access, vol. 11, pp. 94563-94574, 2023

Our works for DeepVoice Detection(3)

- Balance, Multiple Augmentation, and Copy-synthesis Method

  Utilize **Supervised Contrastive Learning**, but in a better training strategy, which proactively set the number of samples in each training mini-batch as follows:

(1) Balances samples between real/fake

(2) Utilizes multiple augmentation methods

(3) Copy-synthesis to generate fake samples

Source: Thien-Phuc Doan, Long Nguyen-Vu, Kihun Hong, Souhwan Jung, "Balance, Multiple Augmentation, and Re-synthesis: A Triad Training Strategyfor Enhanced Audio Deepfake Detection," in Interspeech 2024, accepted

Our works for DeepVoice Detection(3)

- Why it's better?

01 **Deepfake Dataset is unbalanced**

So keep it balance for every training mini-batch improve the optimization progress

02 **Using Multi-augmentation methods to improve the model generalization**

03 **Copy-synthesis**

To generate more hard negative samples, which have the same linguistic content to real sample but have AI artifact, **useful to find better decision boundary.**

Table 5: *Comparison of our best system with SOTA models in ASVSpoof 2021 DF track. The result is shown in EER (%).*

	Methods	EER
[22]	Wav2Vec + lightDART	7.86
[23]	Wav2Vec + FeedForward (FF) + Atn.Pool	4.98
[24]	Wav2Vec + biLSTM	4.75
[25]	Wav2Vec + ViT-based + FF	3.18
[21]	Wav2Vec + AASIST	2.84
[7]	Wav2Vec + Conformer	2.58
	ours SCL conf-3 (Wav2Vec + Linear layers)	**2.17**

Source: Thien-Phuc Doan, Long Nguyen-Vu, Kihun Hong, Souhwan Jung, "Balance, Multiple Augmentation, and Re-synthesis: A Triad Training Strategyfor Enhanced Audio Deepfake Detection," in Interspeech 2024, accepted

Challenges to DeepVoice Detection

3 Core Challenges

Data Generalization & Model Robustness

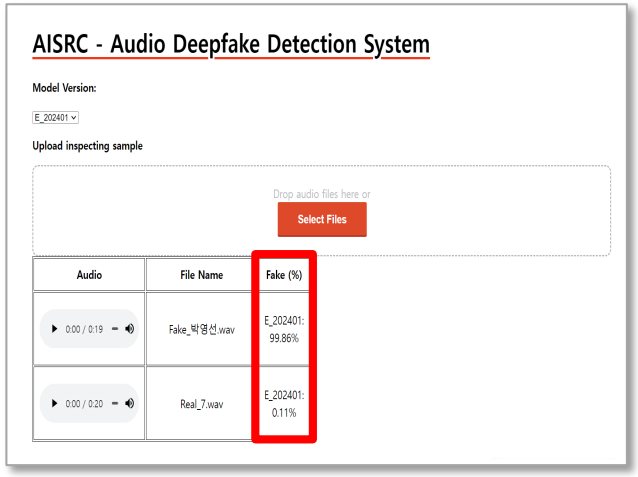
Partially Fake

Adversarial Attack

Practical DeepVoice Detection Systems

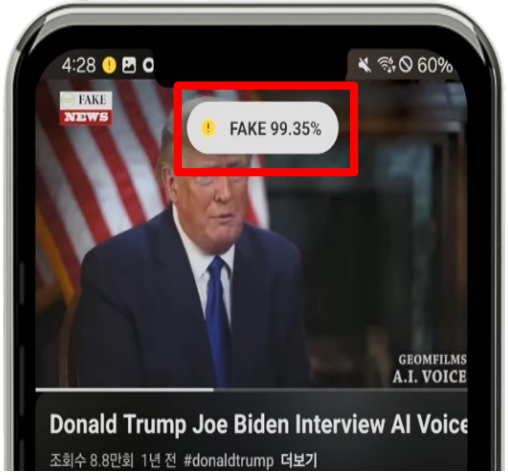
- Our DeepVoice Detection System

[Web-Based Detection System]



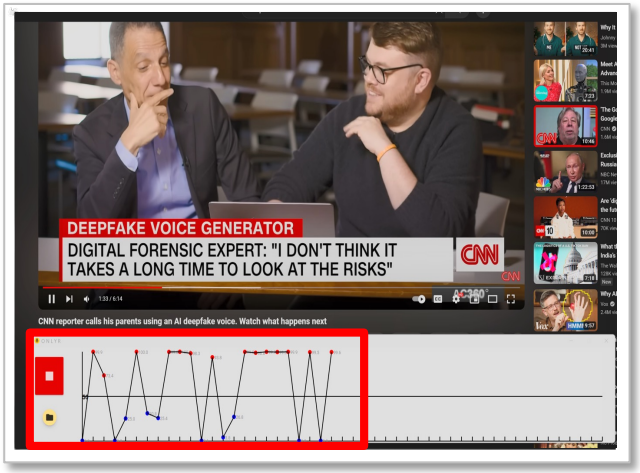
- Supervised Contrastive Learning (SCL)
- Breathing-Talking-Silence Encoding (BTS-E)
- User-Friendly Interface

[Smartphone-Based real-time Detection]



- Efficient Lightweight Model
- Real-time Monitoring
- Enhance Voice Activity Detection(VAD)

[PC-Based real-time Detection]



- Optimized Lightweight Model
- Continuous Real-Time Analysis
- Interactive Graphical Display
- Alert Notifications

Video Clip of DeepVoice Detection Systems

Take-Away

- You could be a victim anytime by your deepvoice in the near future.
- Two approaches to respond to these threats :
  Regulation(watermaking) and Technology(detection system)
- Government regulations are already going on.
- Urgent to develop a robust detection system
- Still many challenges to robust detection system in real-time environment

Thank You!

Any questions can be directed to my email
souhwanj@ssu.ac.kr