black hat USA 2022

Human or Not: Can You Really Detect the Fake Voices?

Liu Xin, Tan Yuan

School of Information Science and Engineering, Lanzhou University

Contents

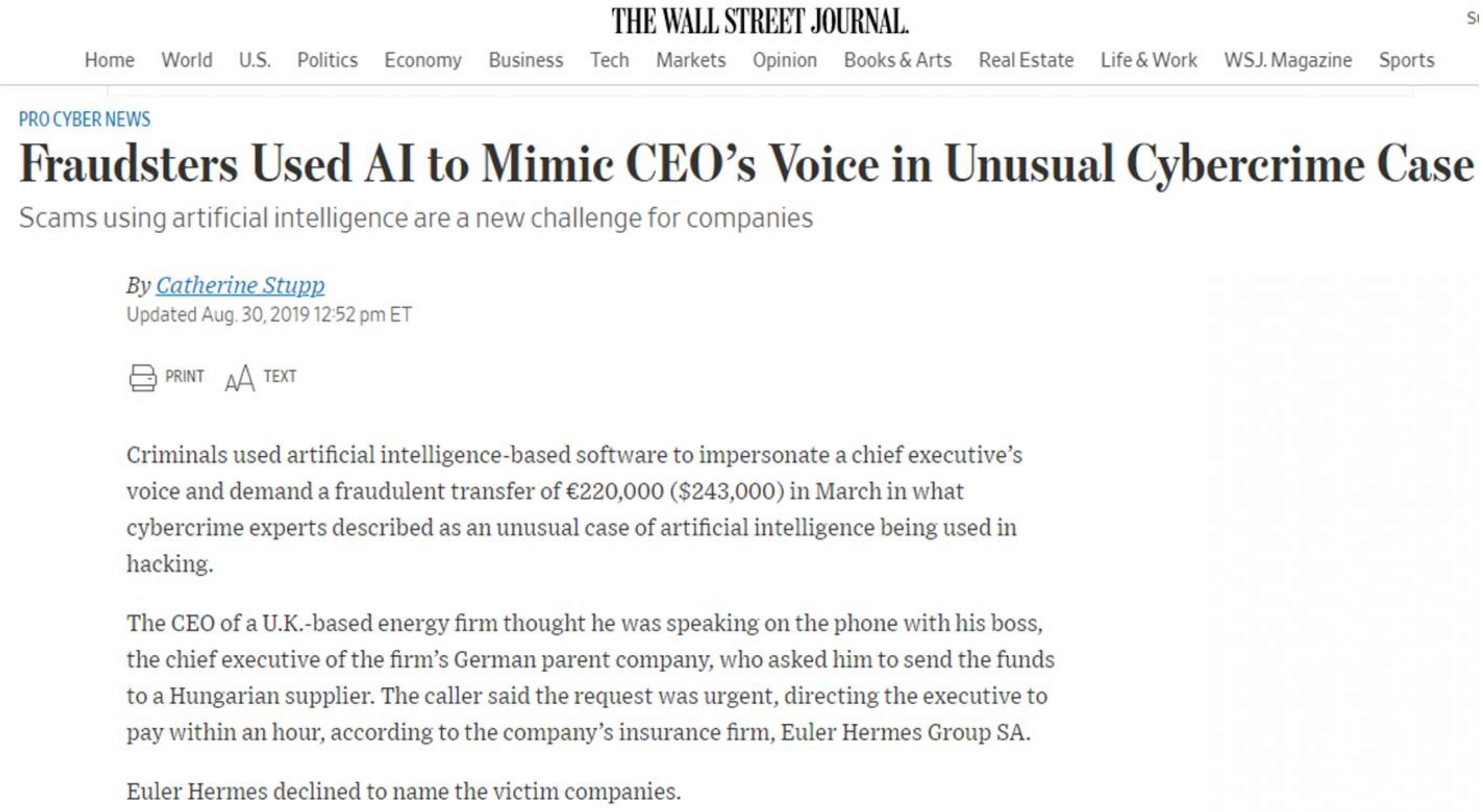
- **What's Fake Voice**
  - Introduction of AI-synthesized speeches
- **Existing Detectors**
  - Existing AI-synthesized Speech Detection Approaches
  - Problems in Existing Approaches
- **SiF-DeepVC**
  - Voice Clone based on Deep Learning and Speaker-irrelative Features
  - Detection Bypass using SiF-DeepVC
- **Evaluation**
  - Four experiments to prove our findings
- **Conclusion**
  - Takeaways
  - Open-source code and datasets

What’s Fake Voice

What's Fake Voice

- Novel fake voices are **AI-synthesized speeches**
- Commonly used for fraud, customer service, and authorization bypass
- Voice Clone (VC) is the most dangerous one

VC-based Crime



THE WALL STREET JOURNAL.

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

By Catherine Stupp

Updated Aug. 30, 2019 12:52 pm ET

Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him to send the funds to a Hungarian supplier. The caller said the request was urgent, directing the executive to pay within an hour, according to the company's insurance firm, Euler Hermes Group SA.

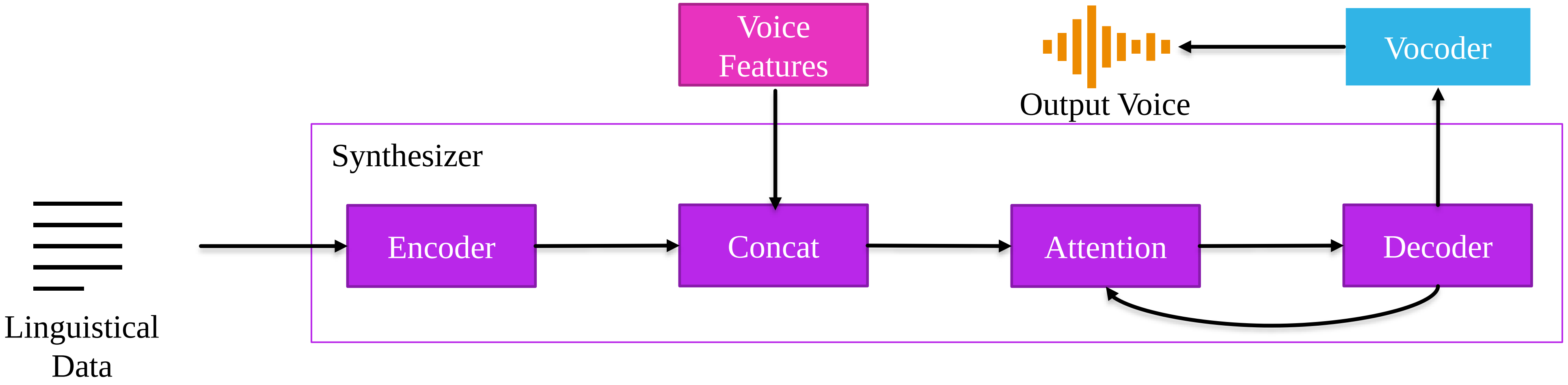
Euler Hermes declined to name the victim companies.

History of Speech Synthesis

- Old Days (Before 20th Century)
  - Simulate sounds with different machines
  - Very difficult to simulate human voice
- “Jigsaw Era” (Before 2010)
  - Automatic “unit selection”
  - Very poor coherence and easy to detect
- AI-synthesized speeches (Since 2010)
  - Smooth and natural
  - Difficult to detect

AI-synthesized Speeches

- Input: what you want to say, output: voice
- Voice Clone (VC): **Replace “Voice Features”!**



Existing Detectors

Existing Detection Approaches

- All existing approaches are reported **very promising performance**

- Computer Vision (CV)-based approaches
  - Inspired by image recognition techniques that are now quite mature
  - Convert **voice to image** and then use image techniques for classification
  - Most of existing approaches are using CV
  - Top conferences or journals
    - Deep4SNet (2021, ACC > 98%)
    - RES-EfficientCNN (2020, F1 > 97%)
    - Farid et al.(2019, AUC > 99%)

Existing Detection Approaches

- All existing approaches are reported **very promising performance**

- Neural Network Feature (NNF)-based approaches
  - Proposed in ACM MM 2020 (Top Conference)
  - Use neuronal activity in neural network as features
  - SOTA: DeepSonar (2020, ACC=100%)
- End-to-End (E2E)-based approaches
  - Novel approaches, commonly used in NLP problems
  - SOTA: RawNet2 (2021, Baseline for ASVspoof 2021, EER=6.1%)
- Statistical-based approaches
  - Traditional approaches. Not popular in recent years.
  - No publications on top conferences/journals in recent 3 years.

Problems in Existing Approaches

Unrealistic Datasets

Speaker-irrelative Features

Multiple Classifications

Speaker-irrelative Features

- Features that should **NOT** be used to determine "human or not"
  - Not a necessary part of the transmission content
  - Not related to the speaker
- Examples
  - Meaningless Silences: before and after the human voice
  - Background Noises: current sound, wind, and so on
  - Different Languages: English, Chinese, French, and so on

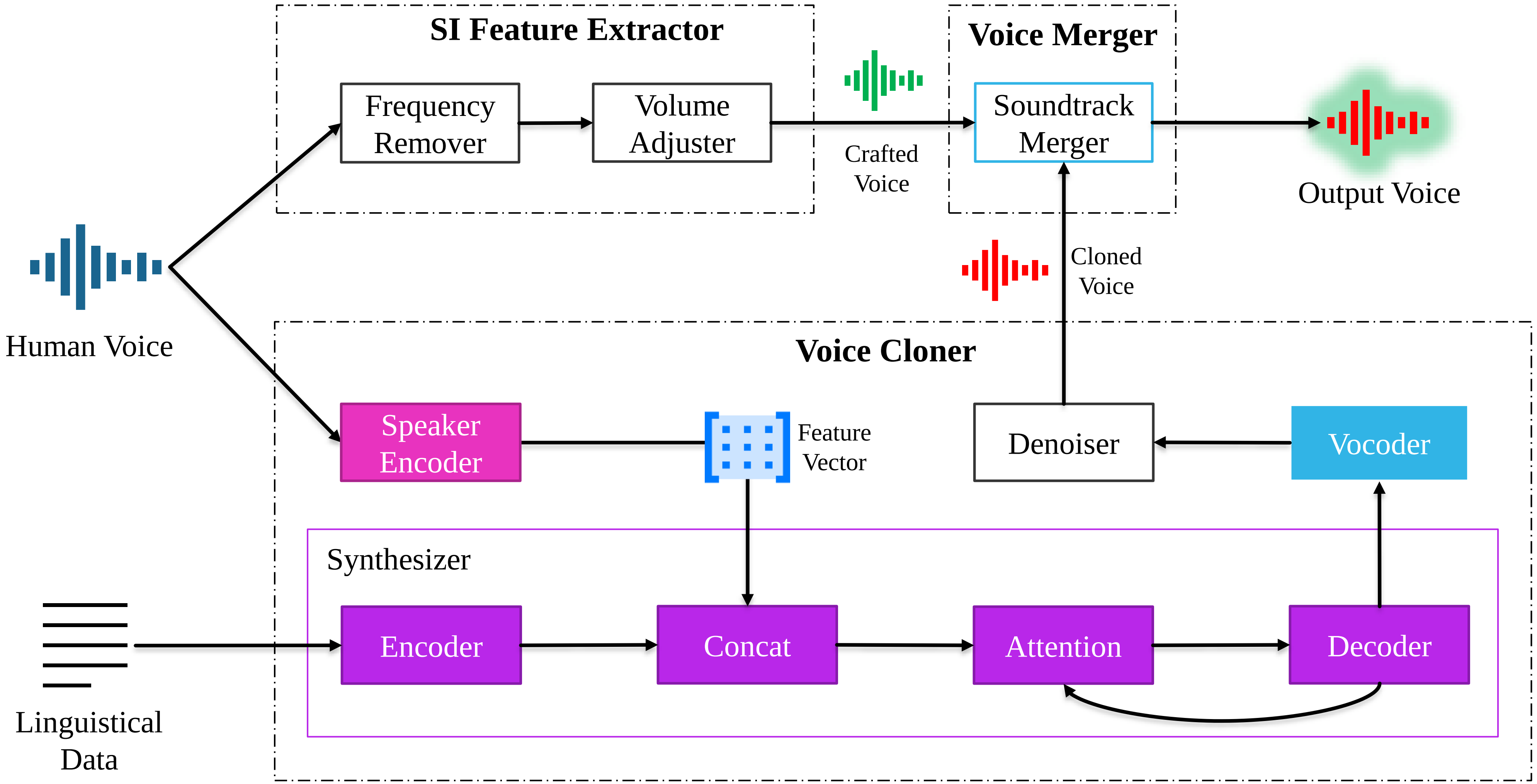
SiF-DeepVC

What’s SiF-DeepVC

- Voice Clone based on Deep Learning and Speaker-irrelative Features
- “SiF-DeepVC”
  - “SiF” stands for “Speaker-irrelative Feature”
  - “Deep” stands for “Deep Learning”
  - “VC” stands for “Voice Clone”

- **PWN detectors with Speaker-irrelative Features**

Overview

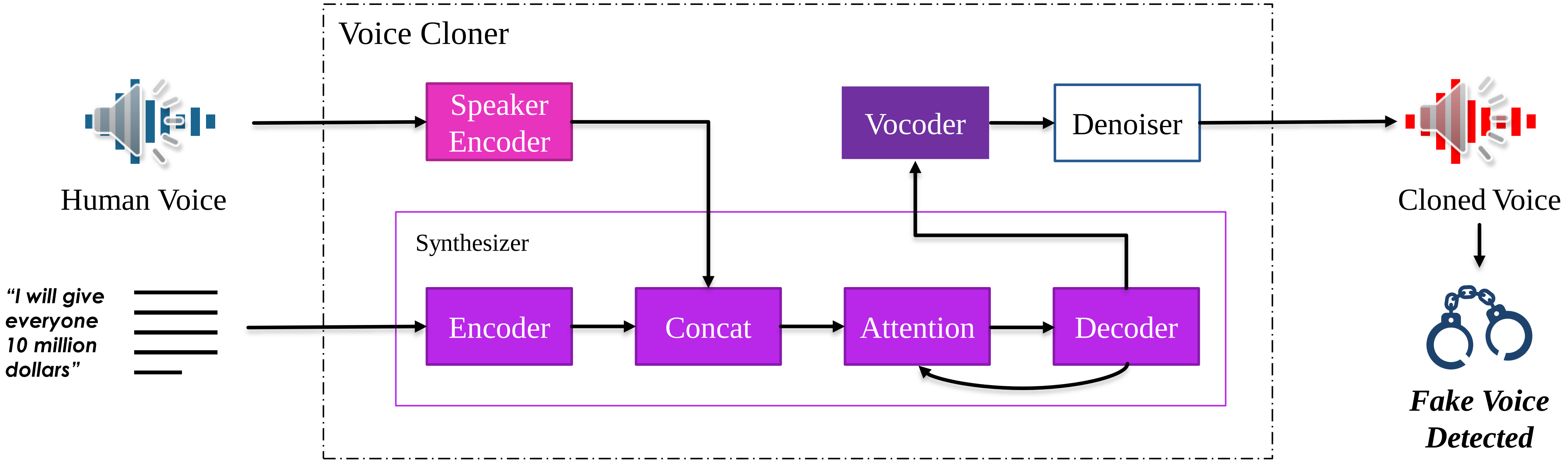

SI Feature Extractor
Frequency Remover
Volume Adjuster
Crafted Voice
Voice Merger
Soundtrack Merger
Output Voice
Human Voice
Cloned Voice
Voice Cloner
Speaker Encoder
Feature Vector
Denoiser
Vocoder
Synthesizer
Encoder
Concat
Attention
Decoder
Linguistical Data

Voice Cloner

- Speaker Encoder
  - Based on G2EE
  - It computes a fixed dimensional feature vector from the speech signal
- Synthesizer
  - Sequence-to-sequence and based on Tacotron implementation
  - It generates a mel spectrum under the constraint of speaker embedding vector
- Vocoder
  - Based on WaveRNN
  - It converts the mel spectrum generated by Synthesizer into time-domain waveforms
- Denoiser
  - It removes the noisy part of the voice generated by Synthesizer and Vocoder
  - high-frequency noise, current sound, etc.

A Demo: Mr. Musk

- Based on a recording of Mr. Elon Reeve Musk
  - Here are the original voice and the voice generated by Voice Cloner
- The latest top-journal published detector (Deep4SNet) still marks it as ***fake***



SI Feature Extractor

- VC attack
  - Convey the information we want to convey through speech
- A successful VC attack needs:
  - The intelligibility of cloned voice is acceptable
  - The size of output cannot be too large

- How to get the speaker-irrelative features from human speeches
  - Remove human audible sound (most of human voice are in 1 kHz~3 kHz)
  - Lower the volume to avoid to be “too noisy”

SI Feature Extractor

- First, for a specific audio file *W:*
  - Its amplitude set is *A,* timestamp set is *T,* frequency set is F
  - We have:

$$W = (A, T, F)$$

- Then, the amplitude *a* ∈ *A* at specific timestamp and frequency is:

$$a = v(W, t, f), t \in T, f \in F$$

SI Feature Extractor

- We denote the silence frequency range F_{sil}
- If $f \in F_{sil}$, we have:

$$W' = (A', T, F)$$

- Then, the amplitude at specific timestamp and frequency is:

$$a' = v'(W, t, f) = \begin{cases} 0, & f \in F_{sil} \\ v(W, t, f), & else \end{cases}$$

SI Feature Extractor

- We silence all the voices below 4 kHz (human voice mainly in 0.3 kHz~3 kHz)
- Since most of 4 kHz+ sounds are noise, we reduce their volume Δa
- We define:
  - $SiF(W)$: the processing function of SI Feature Extractor
  - W'': the audio generated by SI Feature Extractor
- We have:

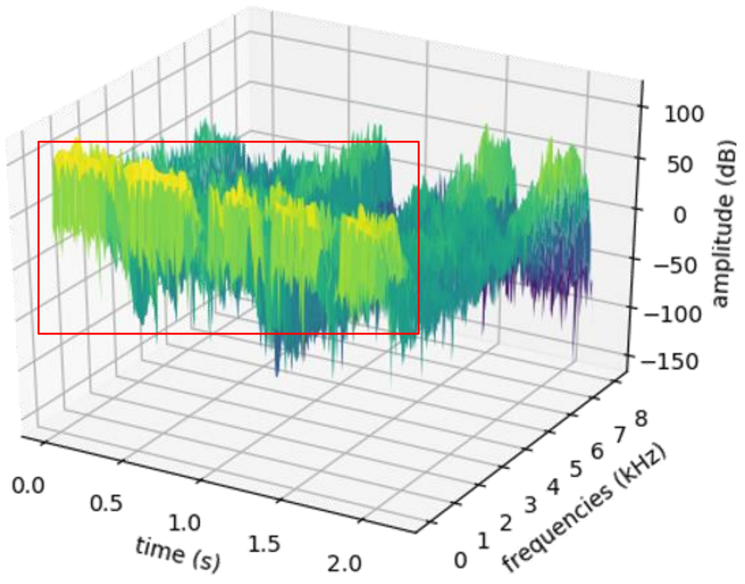
$$W'' = SiF(W) = (A'', T, F)$$

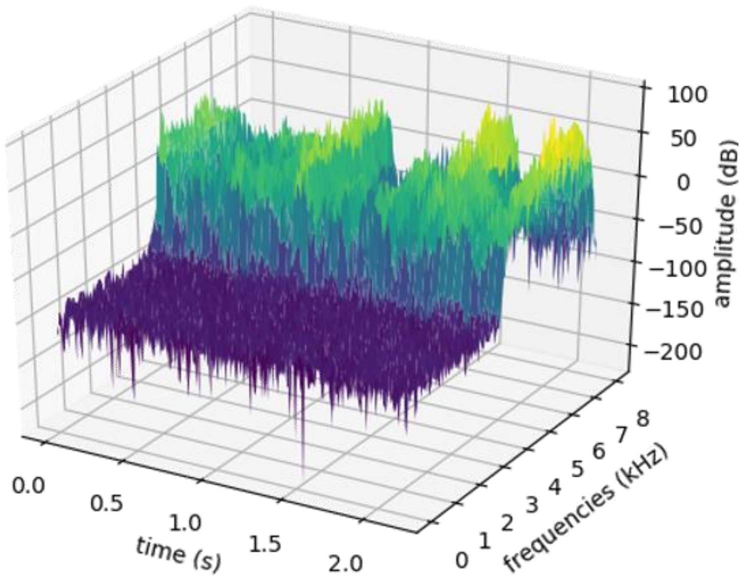
- Its amplitude:

$$a' = v'(W, t, f) = \begin{cases} 0, & f \in F_{sil} \\ v(W, t, f), & else \end{cases}$$

SI Feature Extractor

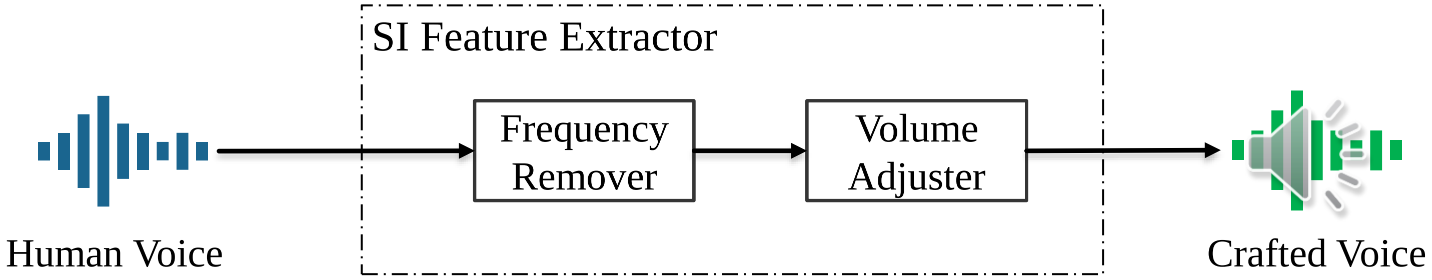
- Time-domain spectrums
  - Human Voice (Left) and Voice after Frequency-based Process (Right)
  - We can see most of high amplitudes are **below 3 kHz** in human voice





A Demo: Mr. Musk

- Based on a recording of Mr. Elon Reeve Musk
  - Target: “**I will give everyone 10 million dollars!**”
  - Here are the original voice and the voice extracted by SI Feature Extractor



Voice Merger

- It combines voices from SI Feature Extractor and Voice Cloner
  - W_{cr} : voice from SI Feature Extractor
  - W_{cl} : voice from Voice Cloner
  - If the length of W_{cr} is smaller or larger than W_{cl}
    - Repeat or crop W_{cr} until its length is the same as W_{cl}
- Its output can be denoted as W_{ou} :

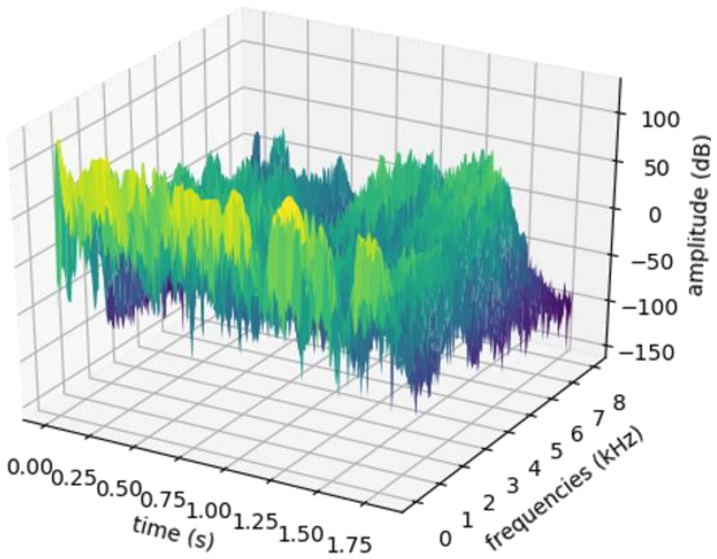
$$
\begin{aligned}
T_{ou} &= T_{cr} \cup T_{cl} \\
F_{ou} &= F_{cr} \cup F_{cl} \\
W_{ou} &= (A_{ou}, T_{ou}, F_{ou})
\end{aligned}
$$

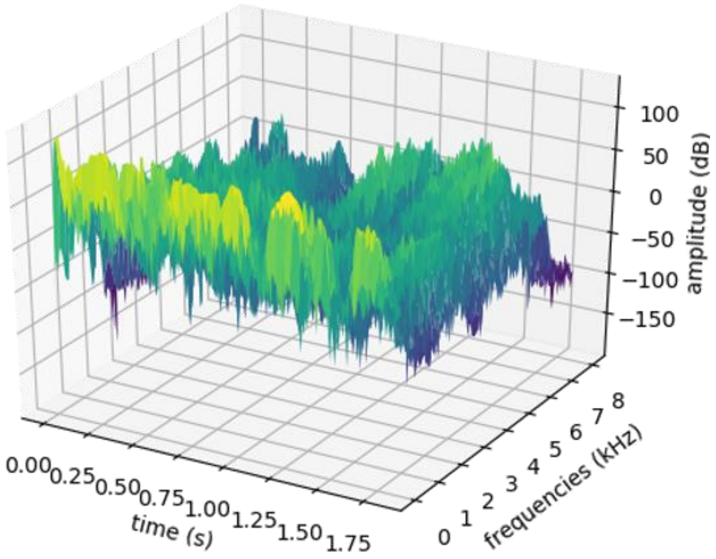
Voice Merger

- For all $a_{ou} \in A_{ou}, t \in T_{ou}, f \in F_{ou}$, we have:

$$a_{ou} = max(v(W_{cr}, t, f), v(W_{cl}, t, f))$$

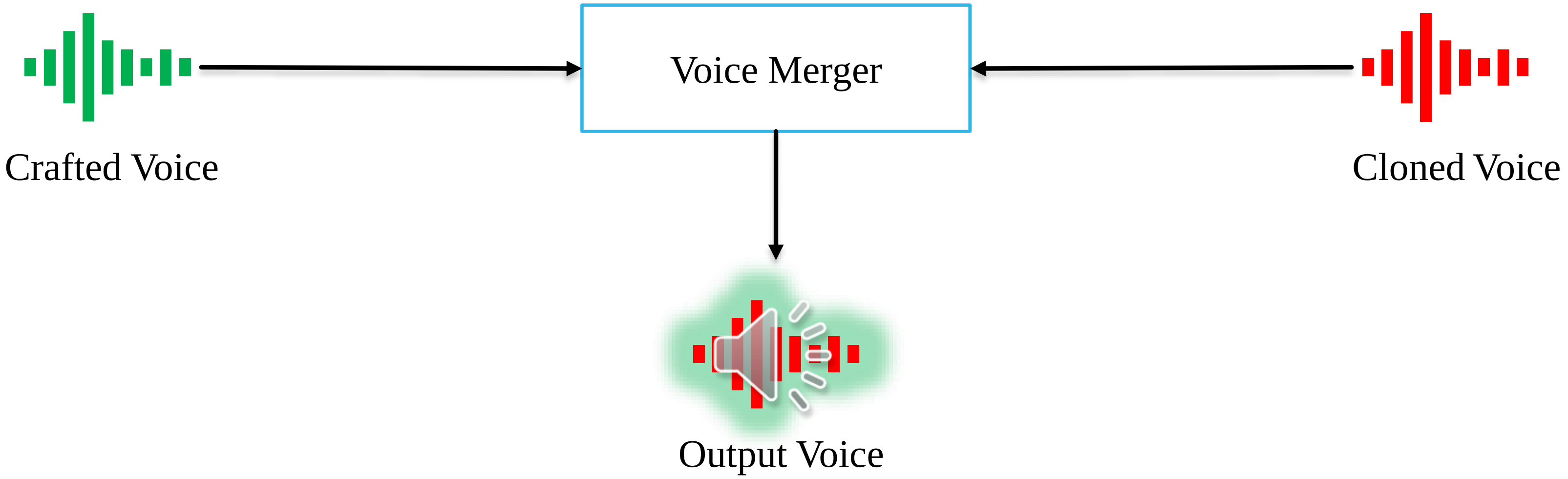
- Time-domain spectrums:
  - Cloned Voice (Left) and Output Voice (Right)
  - Clearly, the differences are very little



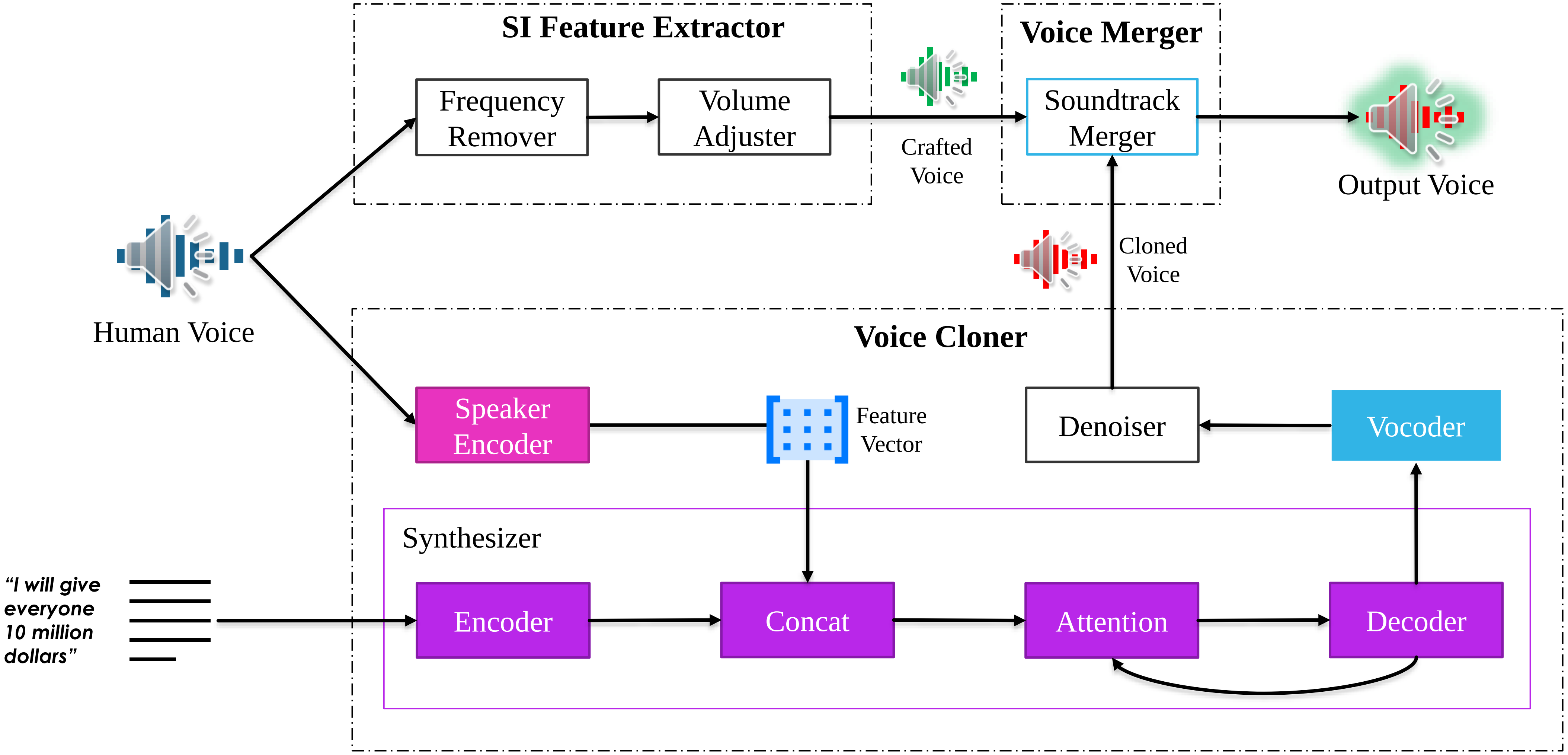


A Demo: Mr. Musk

- Based on a recording of Mr. Elon Reeve Musk
  - Target: “**I will give everyone 10 million dollars!**”
  - Here are the original voice, the cloned voice and the output voice
- The latest detector published on top journal (Deep4SNet) marks it as ***real***



A Demo: Mr. Musk



Evaluation

Evaluation

- Questions for evaluation
  - RQ1: Are existing detection approaches practical in real-world environments
  - RQ2: Do the speaker-irrelative features really affect existing detection approaches
  - RQ3: Can the SiFDeepVC-generated cloned voices bypass existing detection approaches
  - RQ4: Can people understand the speeches generated by SiF-DeepVC

- Baseline Datasets
  - Original human recordings from two open-source datasets as the base datasets
    - FoR Validation: English, 5400 original human recordings, all silence-removed
    - MagicData Test: In Mandarin, 24279 Samples
  - Recordings covering different ages, lengths, and environments, and are well represented
  - **None of these recordings are included in the publications of existing detectors.**

Selected Existing Detectors

- **Deep4SNet**
  - Latest CV-based approach
  - We use the implementation open-sourced by the original authors
- **RawNet2**
  - E2E-based approach as ASVspoof 2021 baseline
  - We use the implementation open-sourced by the ASVspoof 2021
- **Farid et al.**
  - First CV-Based approach on top conference
  - We use the implementation open-sourced in BlackHat USA 2019
- DeepSonar (Not selected)
  - First NNF-based approach, but no open-source implementation available
  - We have tried our best to contact the authors, but have received no response yet

Real-world FPR of Existing Approaches

- This experiment is to answer RQ1
- Real-world Environment
  - Unbalanced samples. Human speeches are **much more than** fake speeches
  - False Alarm Rate (FPR) is very important for real-world deployment
- Results
  - We use the baseline datasets to evaluate the FPR of existing approaches
  - Obviously, **NONE** of their FPRs is acceptable for real world
- **Answer to RQ1: NO**

Approach	Baseline	Positive	FPR
Farid et al.	English	3,656	67.70%
	Mandarin	11,020	45.39%
Deep4SNet	English	3,597	66.61%
	Mandarin	22,081	90.95%
RawNet2	English	5,099	94.43%
	Mandarin	11,580	47.70%

Speaker-irrelative Feature Evaluation

- This experiment is to answer RQ2
- There are two parts in this experiment
  - Slight denoise
    - Not affect the human voice but removes the current sound and the background noise
    - The detection results should not be changed, in theory.
  - Silence removal
    - Only use the Mandarin baseline dataset (English baseline dataset has no silence)
    - Crop the samples to remove the silences before and after the human voice
    - Theoretically, these silences have nothing to do with human speech
    - The detection results should not be changed, in theory.

Speaker-irrelative Feature Evaluation

- Slight denoise
  - **ALL** existing approaches are significantly affected by background noise
  - This means that the noise of human recordings may help fake voices bypass the detection of existing approaches.
- Diff*
  - Compared with original baseline results

Approach	Baseline	DN-FPR	Diff *
Farid et al.	English	75.09%	↑ 10.92%
	Mandarin	84.37%	↑ 85.88%
Deep4SNet	English	59.85%	↓ 10.15%
	Mandarin	99.37%	↑ 9.26%
RawNet2	English	97.22%	↑ 2.95%
	Mandarin	55.74%	↑ 16.86%

Speaker-irrelative Feature Evaluation

- Silence removal
  - **ALL** existing approaches are significantly affected by meaningless silence
  - This means that the silence part of human recordings may help fake voices bypass the detection of existing approaches.
- Diff*
  - Compared with original baseline results

Approach	Baseline	DN-FPR	Diff *
Farid et al.	Mandarin	84.37%	↑ 85.88%
Deep4SNet	Mandarin	99.37%	↑ 9.26%
RawNet2	Mandarin	55.74%	↑ 16.86%

Speaker-irrelative Feature Evaluation

- Conclusion
  - Speaker-irrelative features represented by background noise and silences are indeed accounted by the detection systems as part of the feature vector to determine whether a specific sample is a human speech or not.
  - It lays the foundation for this paper to bypass the AI-synthesized speech detection systems through speaker-irrelative features.

- **Answer to RQ2: YES**

Detection Bypass Evaluation

- This experiment is to answer RQ3
- For cost reasons, we only use the English language for this experiment
- We removed the samples falsely reported as positive in baseline dataset
- For each human recording, we generate five new recordings:
  - *I'm not kidding you, this voice is fake*
  - *The weather is really nice today*
  - *I've sent you the number via WeChat*
  - *Can you please lend me some money*
  - *You need to come to the office tomorrow*

Detection Bypass Evaluation

- Compared to the original human recordings in baseline dataset
  - Fake recordings generated by SiF-DeepVC have much higher negative rates
  - It means that SiF-DeepVC can effectively deceive existing approaches
  - SiF-DeepVC recordings are more “human” than real human
- **Answer to RQ3: YES**

Approach	SiF-DeepVC Recordings			Baseline Original Recordings			Diff
	Negative	Fake	Negative Rate (NR)	Negative	Real	Baseline NR	
Farid et al.	2,823	8,720	32.37%	1,744	5,400	32.30%	↑ 1.00%
Deep4SNet	6,294	9,015	69.82%	1,803	5,400	33.39%	↑ 109.10%
RawNet2	46	1,505	3.06%	301	5,400	5.57%	↓ 45.06%
Average	**9,136**	**19,240**	**47.62%**	**3,848**	**16,200**	**23.57%**	**↑ 102.86%**

Speaker-irrelative Features on VC

- This experiment is to answer RQ4
    - It’s randomized and single-blind
- We recruited a group of 10 participants
    - Listen to selected recordings with headphones
    - Manually verify whether the selected recordings are clear and understandable
- We randomly selected these recordings
    - 100 recordings generated by SiF-DeepVC (Output Voice)
    - 100 recordings generated by Voice Cloner (Cloned Voice)
    - 100 recordings from the English baseline dataset (Human Voice)

Speaker-irrelative Features on VC

- Conclusion
  - We can clearly see that there is no statistical difference between these voices
  - We believe that people can understand the output voices of SiF-DeepVC very well

Type	Understandable	Total	Ratio
Baseline	97	100	97.00%
Cloned Voice	94	100	94.00%
Output Voice	96	100	96.00%

- **Answer to RQ4: YES**

Conclusion

Takeaways

- AI-synthesized speeches generation and detection
  - How to generate AI-synthesized speeches
  - Existing detection approaches and their problems
  - A novel attack framework which can **bypass existing detectors**
- Difficulty of defending against AI-synthesized speeches
  - With SiF-DeepVC, the cloned voice can be more “human” than human
  - Risk warning for blue side: **existing solutions are far from "usable"**
- New datasets for future researches
  - We build and open-source several new datasets with high quality

Demos

- We deeply understand the importance of reproducibility
- All code of this project is available on GitHub
  - Some code included in our repository are from the following projects
    - Deep4SNet: https://github.com/yohannarodriguez/Deep4SNet
    - Farid et al.: https://github.com/cmrfrd/DetectingDeepFakes_BlackHat2019
    - RawNet2: https://github.com/eurecom-asp/rawnet2-antispoofing
    - RTVC: https://github.com/CorentinJ/Real-Time-Voice-Cloning
- All datasets used in this project are also available to the public
  - Please note that the size of this zip file is about **7.8 GB**
  - GitHub: https://github.com/dstsmallbird/SiF-DeepVC_Dataset

Description for Datasets

- The zip file contains the following datasets:
  - RQ1:
    - “for-real-validation”: original human recordings from FoR Validation dataset (also used in RQ3)
    - “zh-real-test”: original human recordings from MagicData Test dataset for (also used in RQ3)
  - RQ2:
    - “for-real-validation-denoised”: slightly denoised “for-real-validation” (RQ2)
    - “zh-real-test-denoised”: slightly denoised “zh-real-test” (RQ2)
    - “zh-real-test-silenced”: silence-removed “zh-real-test” (RQ2)
  - RQ3:
    - “for-bh-madefake-final-r4k”: cloned fake voices by **SiF-DeepVC** for Farid et al.
    - “for-deep4s-madefake-final-r4k”: cloned fake voices by **SiF-DeepVC** for Deep4SNet
    - “for-rawnet-madefake-final-r4k”: cloned fake voices by **SiF-DeepVC** for RawNet2
  - RQ4 :
    - Take any sample you want ☺

Sadly, We Cannot Really Detect the Fake Voices Now

Maybe we can do it in the future

Thanks

Contact: xliu2019 [AT] lzu.edu.cn