NDSS 2026

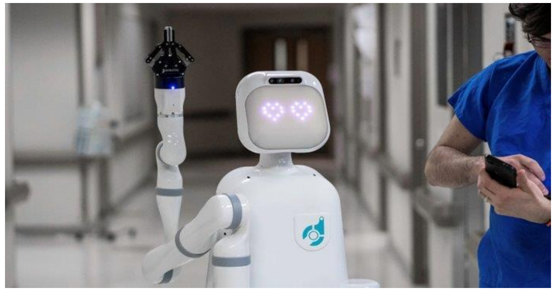
Rethinking Fake Speech Detection: A Generalized Framework Leveraging Spectrogram Magnitude

Zihao Liu, Aobo Chen, Yan Zhang, Wensheng Zhang, Chenglin Miao
Iowa State University

Speaker: Zihao Liu

Speech Synthesis

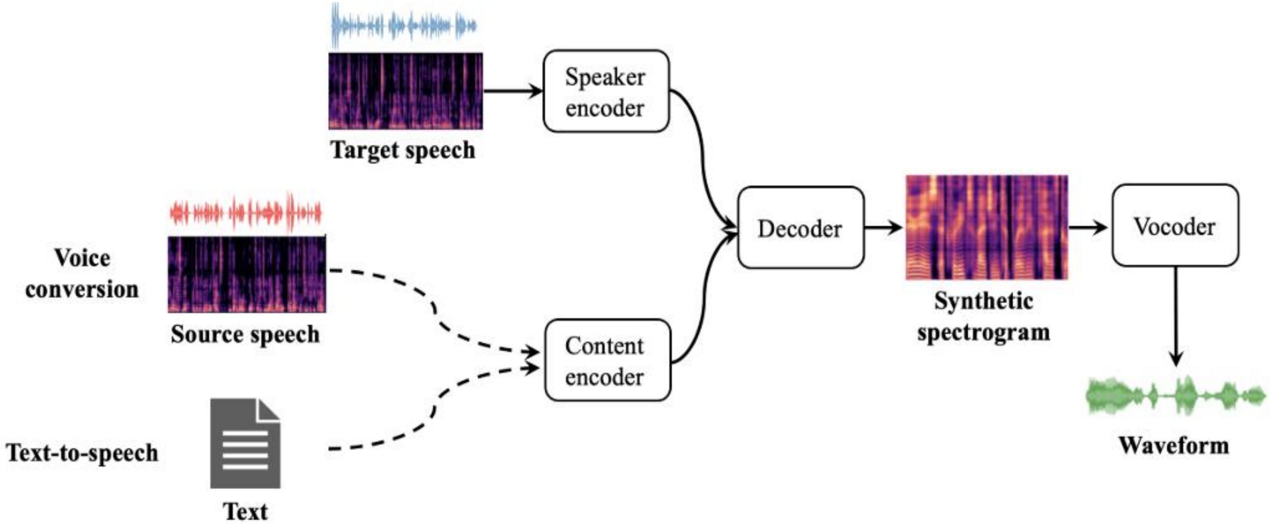
- **Speech synthesis** aims to generate synthetic speech in a voice of a target speaker
- **Applications of speech synthesis**
  - Help people who have lost their voice
  - Language translation
  - Increase human trust to healthcare robots



Speech Synthesis

- **Voice conversion (VC)**
  - Convert a source speaker's voice to sound as if spoken by the target speaker while keeping linguistic contents unchanged

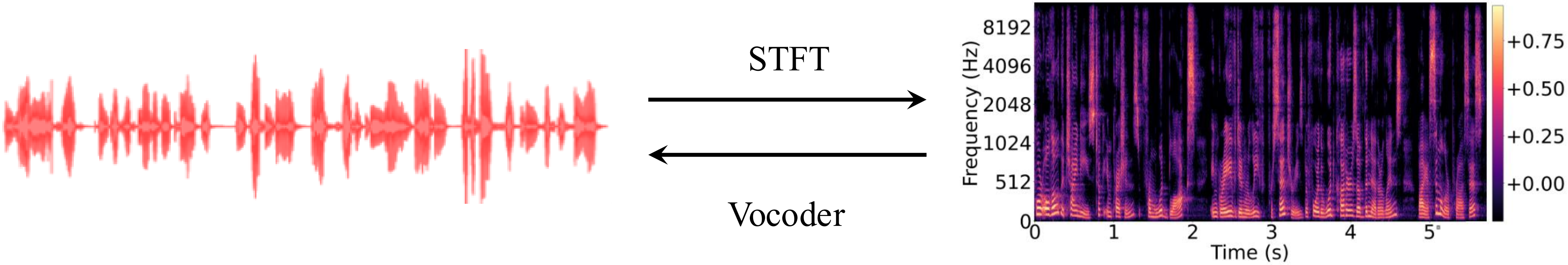
- **Text-to-speech (TTS)**
  - Convert arbitrary texts and a target speech sample that provides voice characteristics as inputs to synthesize a speech



Speech Synthesis

- **Spectrogram transformation and vocoder**
  - The **STFT** provides a joint time–frequency representation of speech, preserving both temporal dynamics and spectral information
  - The **vocoder** reconstructs the waveform from a magnitude-only spectrogram, which contains the absolute values of the STFT results

* The color in the spectrogram represents the signal magnitude at each specific time and frequency

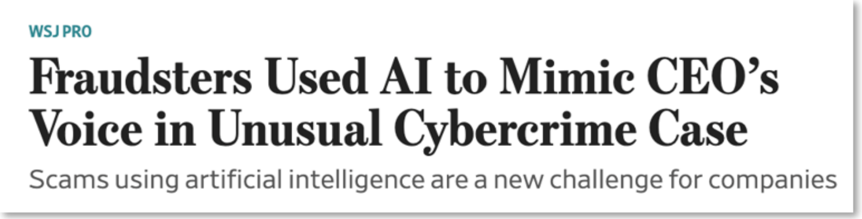


STFT

Vocoder

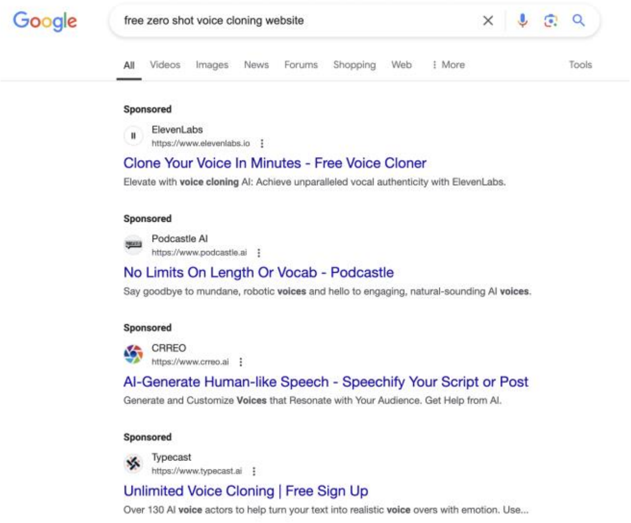
Speech Synthesis Attack

- **Speech synthesis attack:** An attacker aims to *mimic the voice of a target speaker* and transform his chosen text or voice samples into the same content spoken by the target
  - Carrying out a heist
  - Fool voice-based authentication systems built in devices
  - Fool human beings for financial or other malicious purposes


WSJ PRO

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

Scams using artificial intelligence are a new challenge for companies

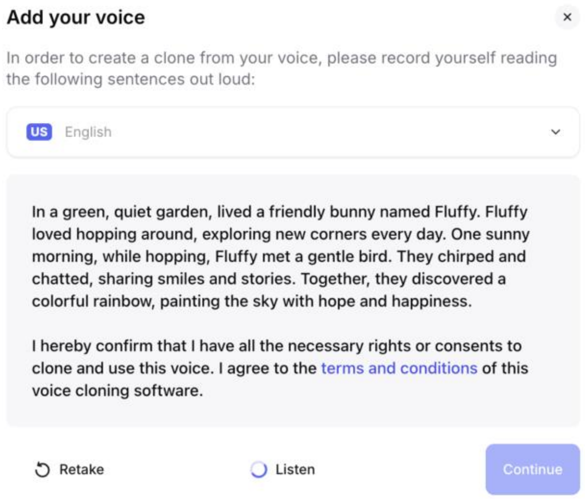


Sponsored
ElevenLabs
https://www.elevenlabs.io
Clone Your Voice In Minutes - Free Voice Cloner
Elevate with **voice cloning** AI: Achieve unparalleled vocal authenticity with ElevenLabs.

Sponsored
Podcastle AI
https://www.podcastle.ai
No Limits On Length Or Vocab - Podcastle
Say goodbye to mundane, robotic **voices** and hello to engaging, natural-sounding AI **voices**.

Sponsored
CRREO
https://www.crreo.ai
AI-Generate Human-like Speech - Speechify Your Script or Post
Generate and Customize **Voices** that Resonate with Your Audience. Get Help from AI.

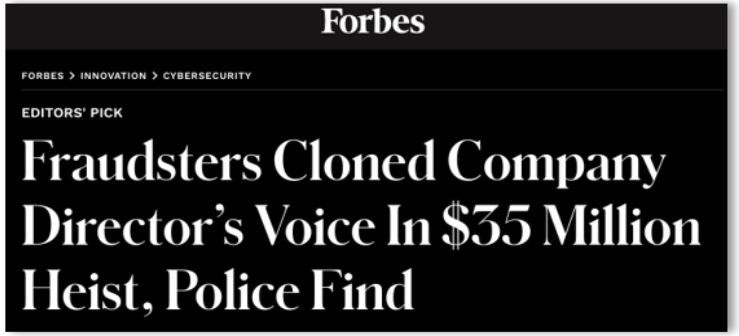
Sponsored
Typecast
https://www.typecast.ai
Unlimited Voice Cloning | Free Sign Up
Over 130 AI **voice** actors to help turn your text into realistic **voice** overs with emotion. Use...


Add your voice

In order to create a clone from your voice, please record yourself reading the following sentences out loud:

In a green, quiet garden, lived a friendly bunny named Fluffy. Fluffy loved hopping around, exploring new corners every day. One sunny morning, while hopping, Fluffy met a gentle bird. They chirped and chatted, sharing smiles and stories. Together, they discovered a colorful rainbow, painting the sky with hope and happiness.

I hereby confirm that I have all the necessary rights or consents to clone and use this voice. I agree to the terms and conditions of this voice cloning software.


Forbes

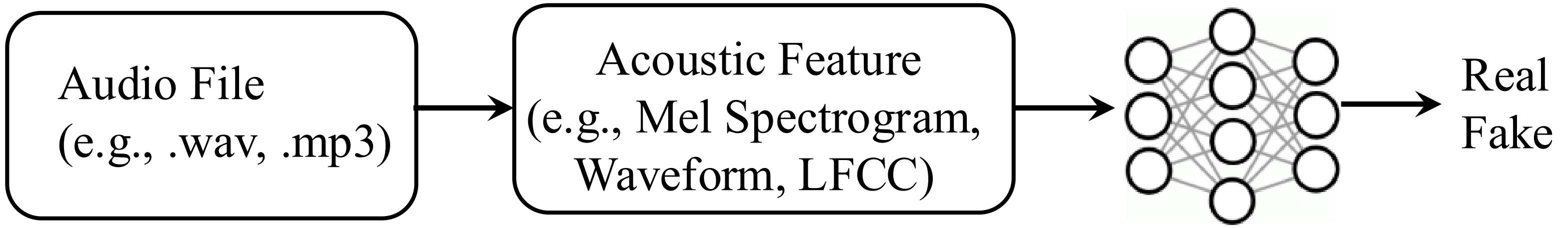
FORBES > INNOVATION > CYBERSECURITY

EDITORS' PICK

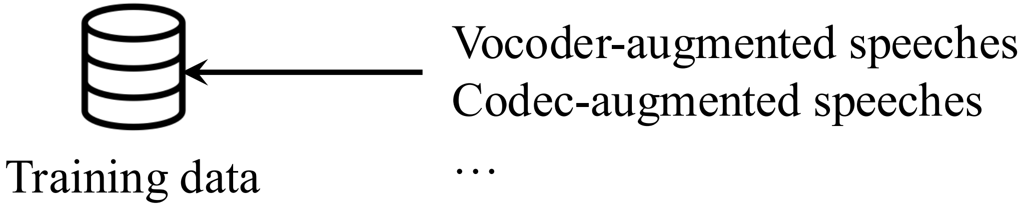
Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find

Fake Speech Detection

- **Fake speech detection**
  - Search for the best combinations of acoustic features and neural architectures



  - Detect liveness cues (e.g., microphone traces, heartbeat, breathing)
  - Improve training robustness with specific augmentations



Fake Speech Detection

- **Limitations of existing fake speech detection**
  - Rely on specific assumptions and recording conditions
  - Limited generalization
  - Lack of artifact explanation

- **Challenges of fake speech detection**
  - Audio signals are complex
  - Artifacts are abstract concepts
  - Data scarcity and distribution gap

Fake Speech Analysis

- Q1: Can current speech synthesis models generate perfect synthetic speech that is indistinguishable from real speech? If not, what factors prevent perfection, and how are artifacts introduced during the synthesis process?

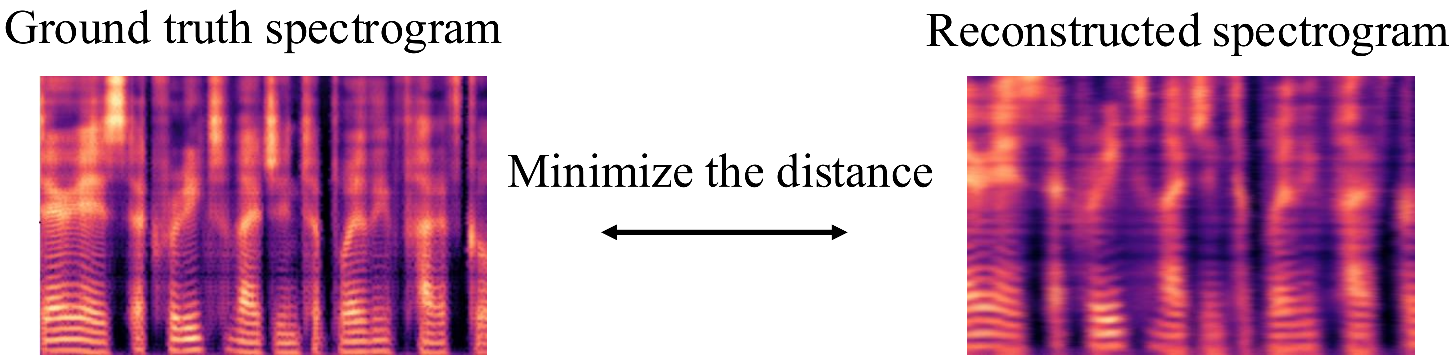
- Q2: What artifacts make fake speech distinguishable, where are they located, and how are they manifested?

- Q3: Do these artifacts share generalizable characteristics?

Fake Speech Analysis

Q1: Can current speech synthesis models generate perfect synthetic speech that is indistinguishable from real speech? If not, what factors prevent perfection, and how are artifacts introduced during the synthesis process?

- Artifacts inevitably arise during vocoding process due to lack of accurate phase information
- "Naturalness" is not explicitly learnable during speech synthesis training process (most vocoders prioritize minimizing the "reconstruction error"


Ground truth spectrogram

Minimize the distance

Reconstructed spectrogram

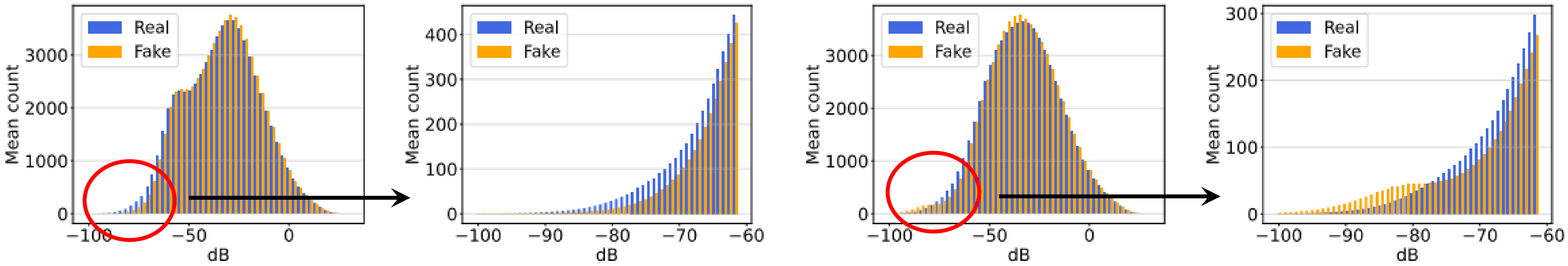
- Dimensional increase typically introduce artifacts (e.g. AutoVC converts low-dim embedding to high-dim spectrogram: $\mathbb{R}^{2\times 256} \rightarrow \mathbb{R}^{80\times T}$)

Fake Speech Analysis

- **Q2: What artifacts make fake speech distinguishable, where are they located, and how are they manifested?**

WaveFake Dataset

LibriSecVoc Dataset



Full magnitude range

Small magnitude range

Full magnitude range

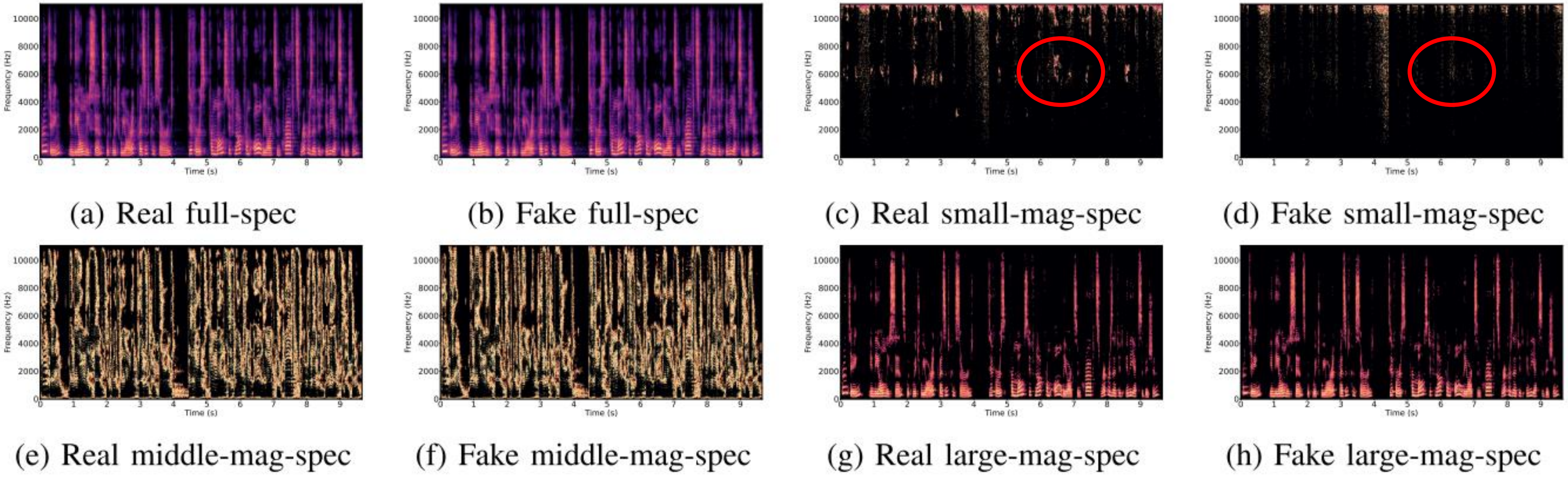
Small magnitude range

Distribution of time-frequency point counts across different magnitude ranges

Observation: The differences become significantly more pronounced in low-magnitude regions

Fake Speech Analysis

- **Q2: What artifacts make fake speech distinguishable, where are they located, and how are they manifested?**
  - Fake speech typically lacks texture details and energy in small-magnitude ranges
  - Real speech typically demonstrates more irregular energy patterns/clusters in small-magnitude ranges

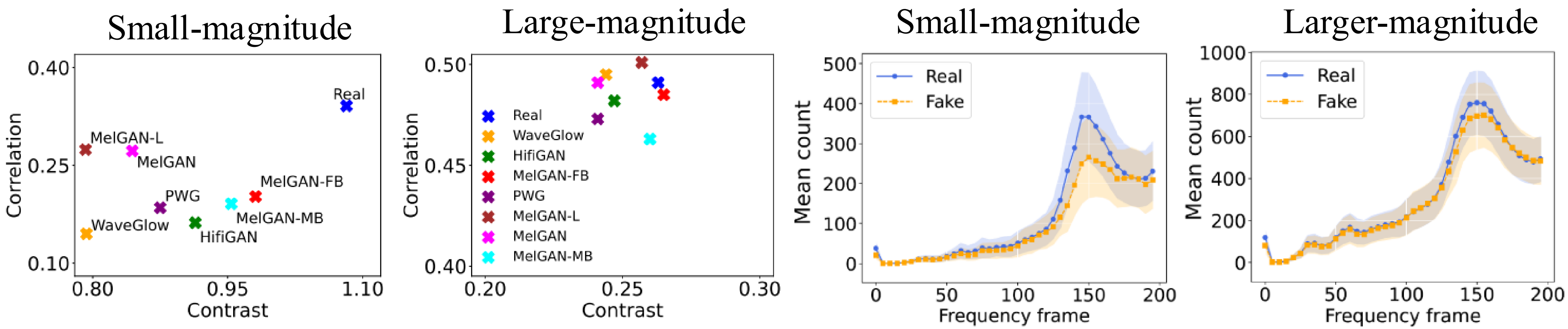


(a) Real full-spec (b) Fake full-spec (c) Real small-mag-spec (d) Fake small-mag-spec

(e) Real middle-mag-spec (f) Fake middle-mag-spec (g) Real large-mag-spec (h) Fake large-mag-spec

Fake Speech Analysis

- **Q2: What artifacts make fake speech distinguishable, where are they located, and how are they manifested?**



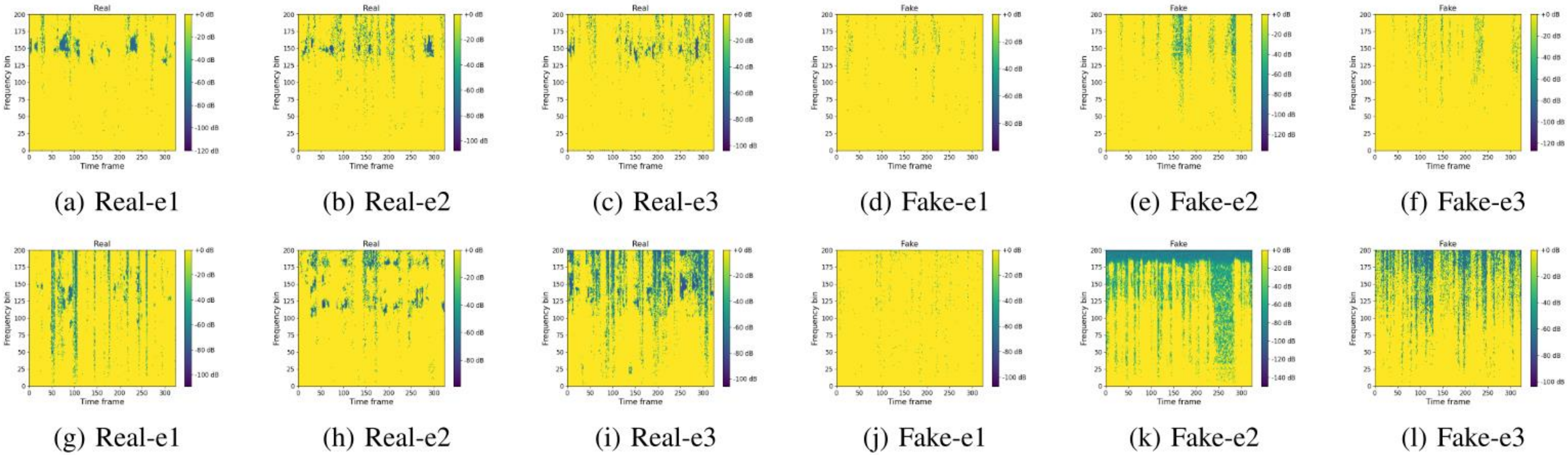
(a) (b) (c) (d)

Spectrogram textures show greater pattern variation and dependency in small-magnitude ranges (measured using the correlation and contrast metrics of GLCM)

Count differences between real and fake speech typically appear in higher frequency ranges (e.g., 5000–7000 Hz)

Fake Speech Analysis

- **Q3: Do these artifacts share generalizable characteristics?**
  - The observation is consistent across different datasets and vocoders

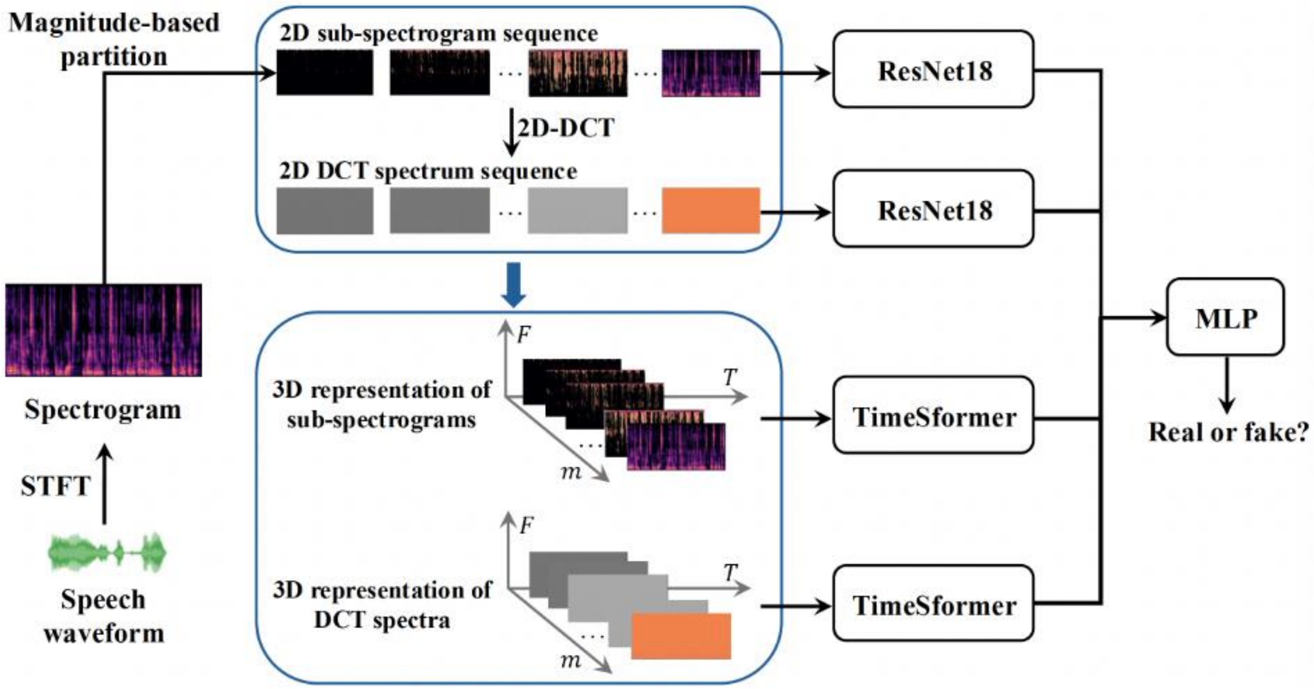


(a) Real-e1 (b) Real-e2 (c) Real-e3 (d) Fake-e1 (e) Fake-e2 (f) Fake-e3

(g) Real-e1 (h) Real-e2 (i) Real-e3 (j) Fake-e1 (k) Fake-e2 (l) Fake-e3

Fake Speech Detection Framework

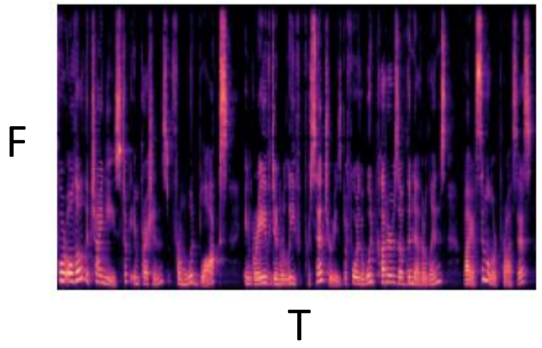
- **Overview**
  - Partition spectrogram into layered 2D sub-spectrograms based on magnitude
  - Detect in both spatial and DCT frequency domain
  - Leverage both 2D and 3D perspectives of a spectrogram to explore its “depth” information and layer-consistency


Magnitude-based partition
2D sub-spectrogram sequence
2D-DCT
2D DCT spectrum sequence
ResNet18
ResNet18
MLP
Real or fake?
Spectrogram
STFT
Speech waveform
3D representation of sub-spectrograms
F
T
m
TimeSformer
3D representation of DCT spectra
F
T
m
TimeSformer

Fake Speech Detection Framework

- **Magnitude-based partition**

Raw spectrogram: $X = [X_{f,t}]_{F\times T}$



$$\downarrow$$

2D representation

$$\{X_{\text{sub}}^{(\tau_0,\tau_1)}, X_{\text{sub}}^{(\tau_0,\tau_2)}, \dots, X_{\text{sub}}^{(\tau_0,\tau_i)}, \dots, X_{\text{sub}}^{(\tau_0,\tau_m)}\}$$

$$\downarrow$$

$$\{X_{\text{sub-dct}}^{(\tau_0,\tau_1)}, X_{\text{sub-dct}}^{(\tau_0,\tau_2)}, \dots, X_{\text{sub-dct}}^{(\tau_0,\tau_i)}, \dots, X_{\text{sub-dct}}^{(\tau_0,\tau_m)}\}$$

3D representation

$$X_v \in \mathbb{R}^{F\times T\times m} \qquad X_v^{\text{dct}} \in \mathbb{R}^{F\times T\times m}$$

Performance Evaluation

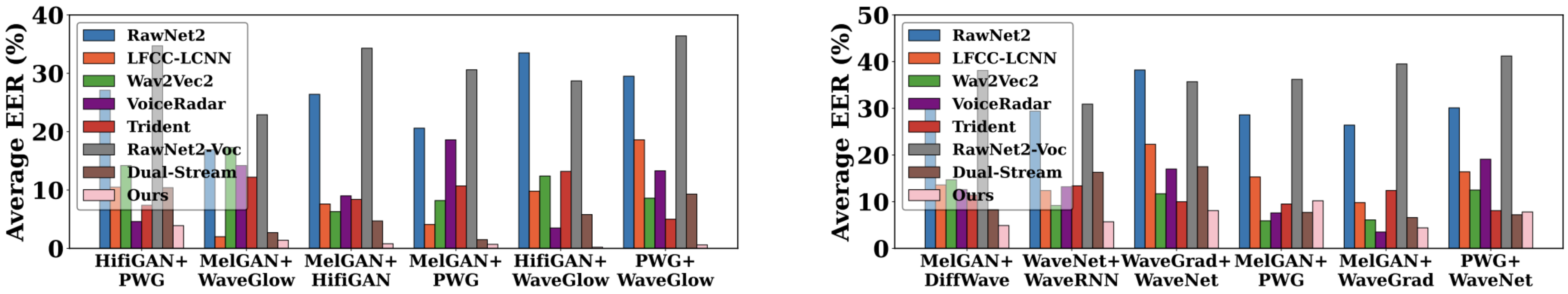
- **Experimental setting**
  - **Two main scenarios:**
    - *Leave-one-out:* detection model is trained on samples generated by all vocoders except one within the dataset and tested on the excluded one
    - *Leave-most-out:* detection model is trained on samples generated by two vocoders within the dataset and tested on remaining unseen vocoders
  - **Evaluation metric: Equal Error Rate (EER):**
    - The point where false acceptance and false rejection rates are equal

Performance Evaluation

- **Detection EER (%) on the WaveFake dataset in the Leave-One-Out scenario**

Method	Unseen vocoder							Average
	MelGAN	MelGAN-FB	MelGAN-MB	MelGAN-L	HiFiGAN	PWG	WaveGlow	
RawNet2	2.3	24.2	18.3	4.8	20.9	7.4	17.3	13.6
LFCC-LCNN	3.6	44.1	12.6	1.1	4.7	0.9	9.8	11.0
Wav2Vec2	6.9	10.2	3.1	3.0	12.6	9.7	2.9	6.9
VoiceRadar	5.3	7.3	17.6	10.2	9.4	4.4	13.8	9.7
Trident	3.2	15.6	2.7	4.0	5.9	5.1	1.9	5.5
RawNet2-Voc	0.6	40.2	9.3	27.4	36.3	22.4	30.0	23.7
Dual-Stream	9.6	1.2	0.6	**0.1**	8.7	4.1	**0.2**	3.5
Ours	**0.1**	**0.3**	**0.1**	**0.1**	**1.0**	**0.2**	0.4	**0.3**

- **Detection EER (%) on the WaveFake and LibriSevoc dataset in the Leave-Many-Out scenario**



(a) The WaveFake dataset

(b) The LibriSeVoc dataset

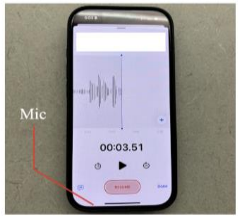
Performance Evaluation

- **Detection EER (%) for Web voice-cloning APIs.**

Method	Voice-cloning API					Average
	ElevenLabs	FishAudio	DupDub	Speechify	PlayHT	
RawNet2	46.4	72.5	54.1	69.7	41.0	56.7
LFCC-LCNN	38.6	24.7	43.1	32.5	35.4	34.9
Wav2Vec2	24.9	56.2	38.9	37.5	44.0	40.3
VoiceRadar	23.5	32.9	19.7	13.4	19.3	21.8
Trident	24.7	25.8	29.4	31.9	18.5	26.1
Ours	**6.8**	**9.4**	**8.3**	**12.6**	**4.5**	**8.3**

* We recruit 13 English-speaking participants (8 male and 5 female) aged between 19 and 45

(a) Laptop Mic



(b) Phone Mic

(c) USB Mic

- **Detection EER (%) under PGD attack and Post-editing attack**

Vocoders for training	PGD-R	PGD-T	Edit-2L	Edit-4L
MelGAN + HifiGAN	1.4	0.8	2.6	1.3
PWG + WaveGlow	1.0	0.9	1.8	0.6

Conclusions

- **We introduce a novel assumption-free and generalized framework for fake speech detection**
- **We conduct a comprehensive analysis to explore why, how, and where artifacts manifest in fake speech**
- **The framework partitions spectrograms into layered magnitude-based representations and detects artifacts in both spatial and DCT domains using 2D and 3D inputs**
- **The desirable performance of the defense schemes is verified on both public datasets and black-box web voice cloning API**

Thank you!