ACSAC 2023

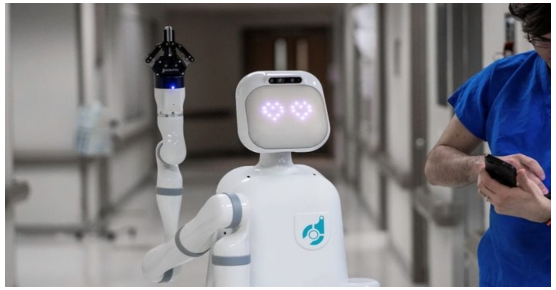
Protecting Your Voice from Speech Synthesis Attacks

Zihao Liu, Yan Zhang, Chenglin Miao

Iowa State University

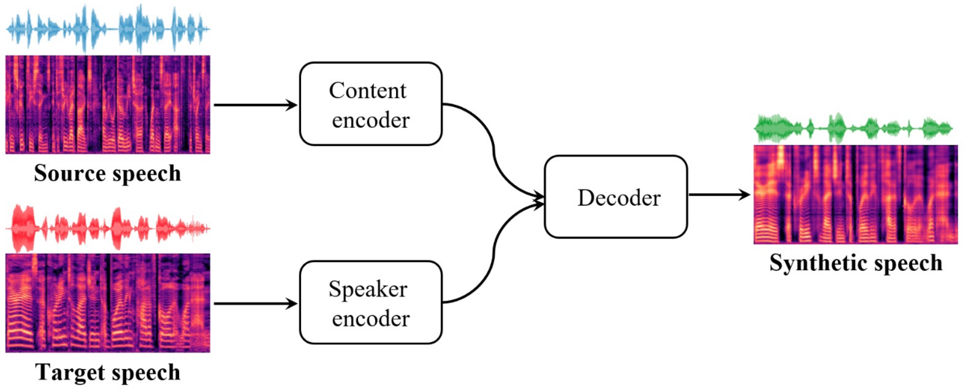
Speech Synthesis

- **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

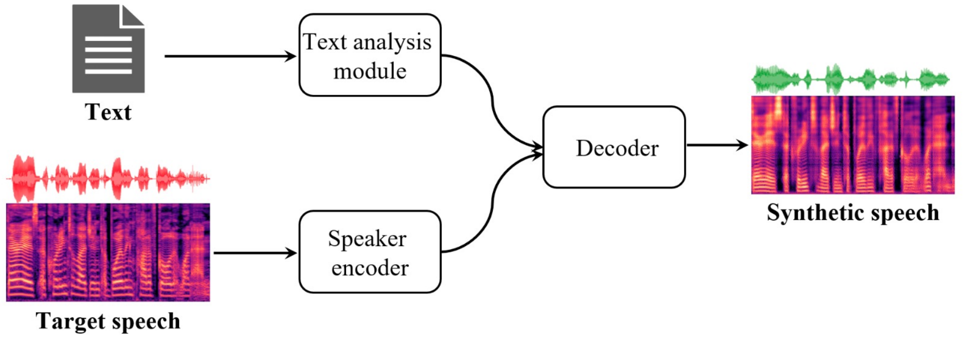


Speech Synthesis

- **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 the target utterance that provides voice characteristics as inputs to synthesize a speech.



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

Forbes

FORBES > INNOVATION > CYBERSECURITY

EDITORS' PICK

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

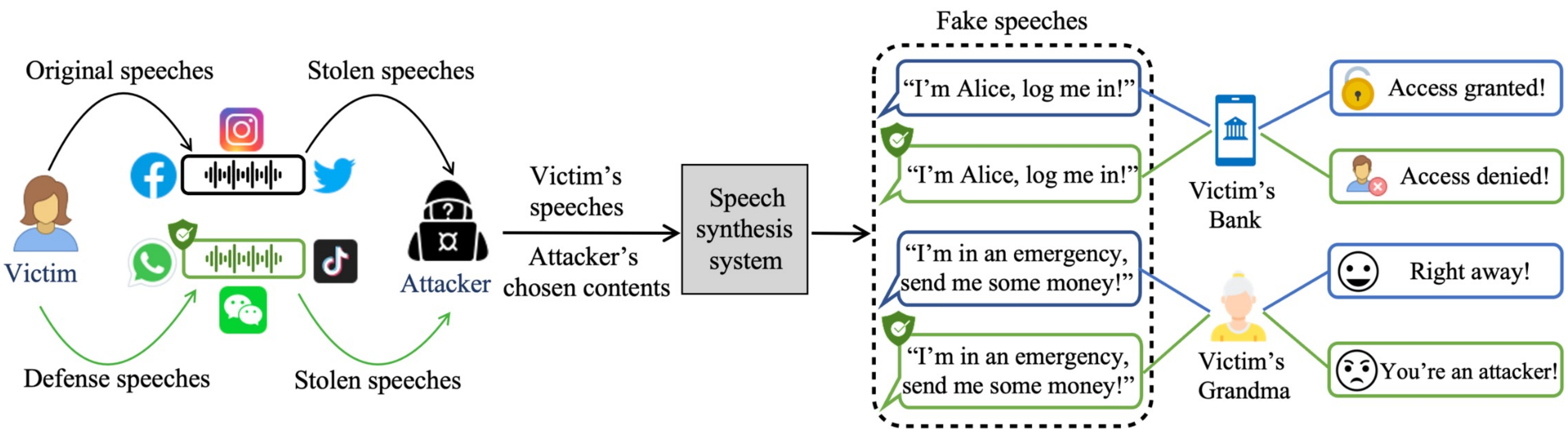
Defense Schemes

- **Fake speech detection:** By discovering artifacts of fake speeches or identifying unique evidence of real speeches
  - Specific assumptions and recording conditions
  - Severe consequences have already occurred

- **Fake speech prevention:** By adding carefully-designed perturbations to the target speaker's speeches before the attacker obtains them
  - Large perturbations
  - White-box setting
  - Low efficiency

Our Goal

- Develop a fake speech **prevention** scheme.
- The **target speaker** can use the scheme to **process his or her speeches before publishing them**.
- The attacker **cannot generate desirable synthetic speeches**.
- The scheme **has little impact on the sound of the target speaker's voice**.



Problem Setting

- **Metrics for defense goal**
  - Quality change of raw speech

$$\Delta Q_d(x) = 1 - S(E_s(x_d), E_s(x))$$

The **similarity score** between the speech embeddings **before and after defense**

  - Quality change of synthetic speech

The generated synthetic speech

$$\Delta Q_I(x) = S(E_s(\mathcal{W}(x)), e_s) - S(E_s(\mathcal{W}(x_d)), e_s)$$

- **Overall defense goal**

$$\max_{\mathcal{D}} \ \Delta Q_I(x)$$
$$\text{s.t.} \ \Delta Q_d(x) < \tau_d,$$

$\mathcal{D}$ denotes the defense strategy

Limit the sample distortion after defense

Defense via Frequency Modification

- **Three modification methods**
  - Zero Mask

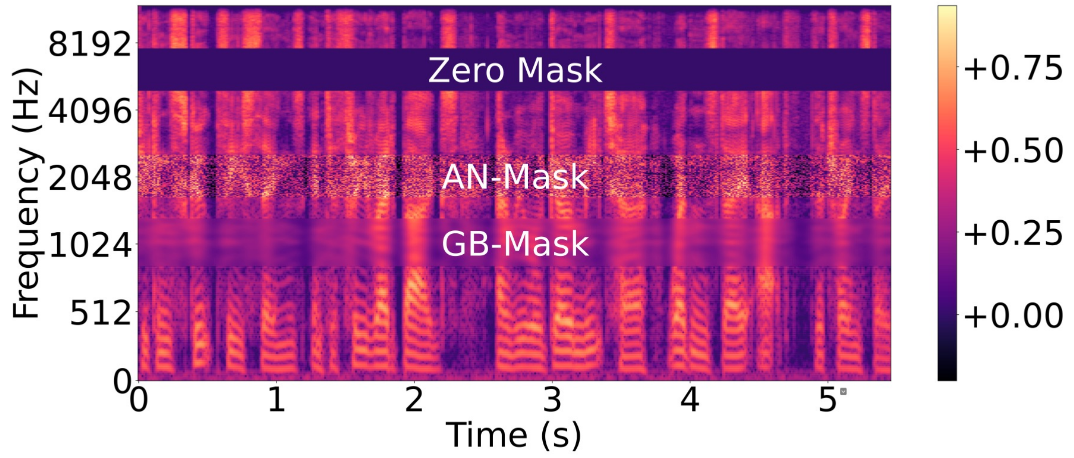
    $$\mathcal{M}_Z(\boldsymbol{x}, \mathbb{F}) = \{\boldsymbol{x} | a_f^t = 0, \forall f \in \mathbb{F} \text{ and } \forall t \in [0, T]\}$$

  - Adaptive Noise Mask (AN-Mask)

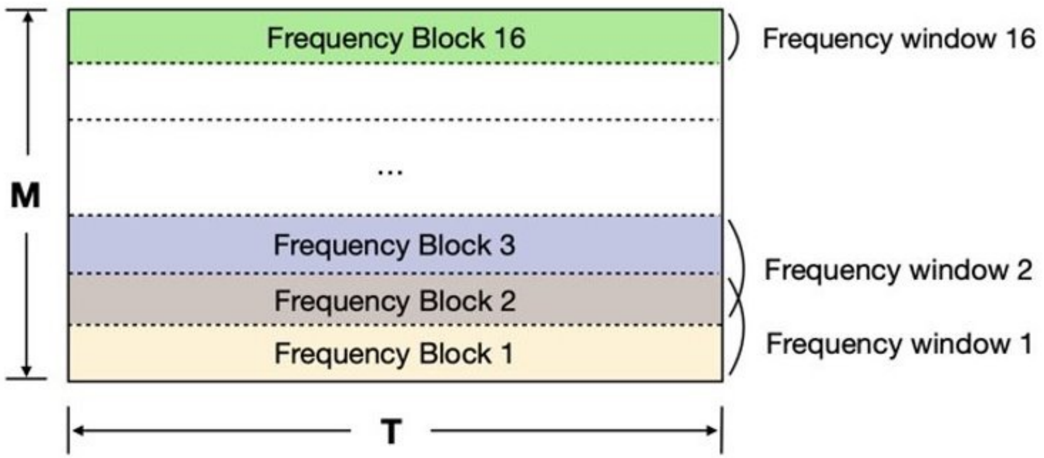
    $$\mathcal{M}_{AN}(\boldsymbol{x}, \mathbb{F}) = \{\boldsymbol{x} | a_f^t = a_f^t + C(\eta(\cdot)), \forall f \in \mathbb{F} \text{ and } \forall t \in [0, T]\}$$

  - Gaussian Blur Mask (GB-Mask)

    $$\mathcal{M}_{GB}(\boldsymbol{x}, \mathbb{F}) = \{\boldsymbol{x} | a_f^t = \phi(a_f^t, G), \forall f \in \mathbb{F} \text{ and } \forall t \in [0, T]\}$$



- **Frequency partition**
  - Split Mel Spectrogram into many blocks
  - Two continuous frequency blocks are called frequency window



Optimal Defense Strategy

- **Find the best frequency-modification method pairs**

$$\max_{\mathcal{D}} \ \Delta Q_I(\boldsymbol{x}) \quad \text{s.t.} \ \Delta Q_d(\boldsymbol{x}) < \tau_d,$$

$$\boxed{\mathcal{D} = \{(b_i, \mathcal{M}_i)\}_{i=1}^{P}}$$

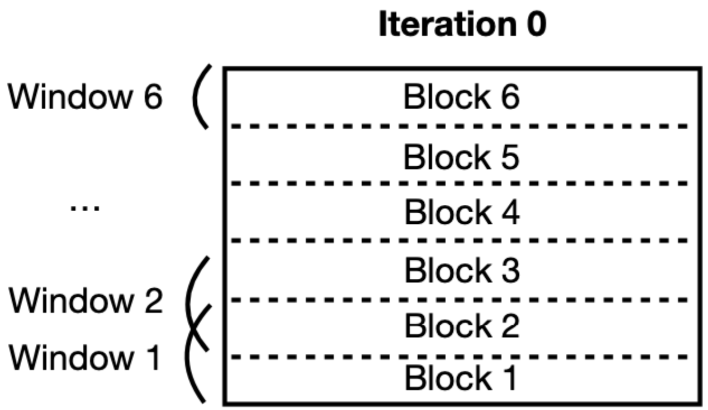
- **Challenges**
  - The defender does not know the model details (black-box setting)
  - The frequency-modification pair selection is not continuous process
- **Solution**
  - Iteratively search with our defined metric, ***frequency sensitivity:***

$$s_j^t = \frac{\Delta Q_I^t(\boldsymbol{x}) - \Delta Q_I^{t-1}(\boldsymbol{x})}{\Delta Q_d^t(\boldsymbol{x}) - \Delta Q_d^{t-1}(\boldsymbol{x})}$$

→ The larger the better

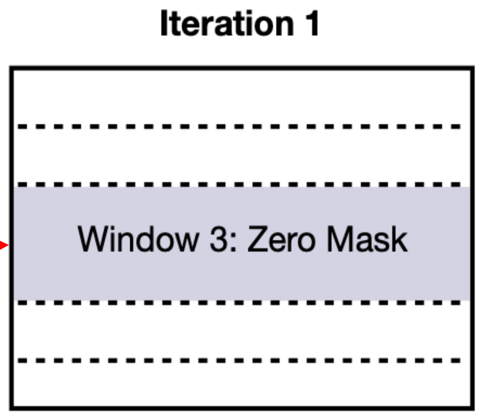
→ The smaller the better

An Example of Iteration Search



$$s_j^t = \frac{\Delta Q_I^t(\boldsymbol{x}) - \Delta Q_I^{t-1}(\boldsymbol{x})}{\Delta Q_d^t(\boldsymbol{x}) - \Delta Q_d^{t-1}(\boldsymbol{x})}$$

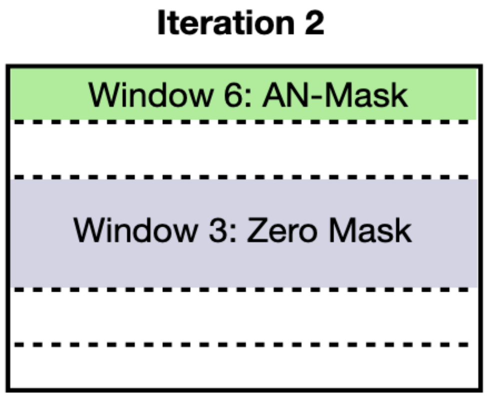
$$s_3^1 = 10$$

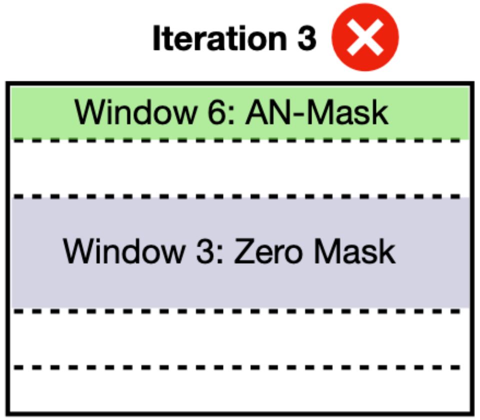


Zero	AN	GB
2	6	3
2	1	4
5	4	3
10	2	5
-	3	1
-	1	-

Iteration 0: Initialize $\Delta Q_I^0(x)$ and $\Delta Q_d^0(x)$ (Both are set to 0)

Iteration 1: (1) Iterate all frequency windows with different modification methods; (2) Select the largest sensitivity among all combinations; (3) perform the corresponding modification





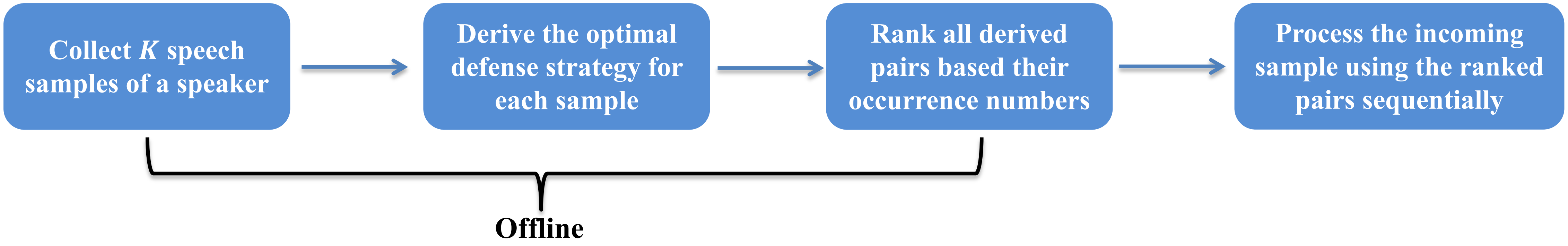
$$\mathcal{D} = \{(b_3, \mathcal{M}_Z), (b_4, \mathcal{M}_Z), (b_6, \mathcal{M}_{AN})\}$$

Iteration 2: repeat the process of Iteration 1

Iteration 3: The sample distortion is beyond the threshold; the search terminates.

Speaker-level Defense

- In some cases, a speaker needs to send instant audio messages to others.
- It is necessary to derive a defense strategy that is **general enough to be directly applied to any speech of a speaker**.



Performance Evaluation

- **Experimental setting**
  - **Dataset:** VCTK
  - **Speech synthesis models:** Chou's[1], AutoVC[2], SV2TTS[3]
  - **Baselines:** Raw (without defense); Attack-VC[4] (a fake speech prevention method)
  - **Speaker recognition (SR) systems:** Resembylzer, Microsoft Azure, Amazon Alexa, WeChat
  - **Metrics**
    - *Attack success rate (ASR):* the percentage of synthetic speeches that successfully fool a specific SR system (the lower the better)
    - *Accept rate (ACR):* the percentage of the modified speeches that are successfully recognized by the SR system (the higher the better)

[1] Chou, et al. "One-shot voice conversion by separating speaker and content representations with instance normalization." arXiv preprint arXiv:1904.05742 (2019).
[2] Kaizhi Qian, et al. "Autovc: Zero-shot voice style transfer with only autoencoder loss." In International Conference on Machine Learning. PMLR, 5210–5219.
[3] Ye Jia, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." Advances in Neural Information Processing Systems 31 (2018).
[4] Chien-yu Huang, et al. "Defending your voice: Adversarial attack on voice conversion." In 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 552–559.

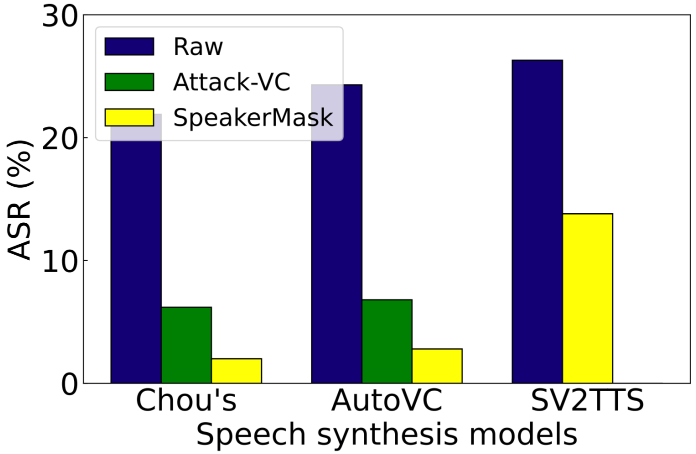
Performance Evaluation

- **Attack success rate (ASR) on Resemblyzer (%)**

	Chou's			AutoVC			SV2TTS	
	Attack-VC	**SampleMask**	**SpeakerMask**	Attack-VC	**SampleMask**	**SpeakerMask**	**SampleMask**	**SpeakerMask**
$\tau_d = 0.06$	69.7	**18.2**	**38.8**	34.3	**19.1**	**24.8**	**19.4**	**49.0**
$\tau_d = 0.12$	46.3	**9.2**	**17.1**	29.3	**13.0**	**15.1**	**8.3**	**29.9**
$\tau_d = 0.18$	30.3	**0.9**	**9.4**	17.2	**6.5**	**10.9**	**3.5**	**13.5**

Raw: Chou's (**84.1%**), AutoVC (**52.4%**), and SV2TTS (**57.1%**)

- **Attack success rate (ASR) on Microsoft Azure**



Acceptance Rate (ACR) of modified speeches (%)

	Chou's	AutoVC	SV2TTS
Resemblyzer	100	100	100
Azure	89.9	84.7	90.1

Performance Evaluation

- **Attack success rate (ASR) on Amazon Alexa (%)**

Commands	Chou's			AutoVC			SV2TTS	
	Raw	Attack-VC	**SpeakerMask**	Raw	Attack-VC	**SpeakerMask**	Raw	**SpeakerMask**
Hey Alexa add an event to my calendar for tomorrow at 5.	50.0	16.7	**8.3**	16.7	0.0	**0.0**	83.3	**75.0**
Hey Alexa check my email	41.7	25.0	**0.0**	25.0	33.3	**0.0**	41.7	**41.7**
Alexa say who is talking with you now	50.0	33.3	**16.7**	16.7	16.7	**0.0**	50.0	**33.3**
Alexa tell me what is on my calendar	66.7	75.0	**16.7**	33.3	41.7	**8.3**	91.7	**66.7**
Tell me what is on my calendar for this week	58.3	66.7	**8.3**	25.0	41.7	**0.0**	75.0	**58.3**
Alexa make an appointment with my doctor	50.0	41.7	**8.3**	33.3	25.0	**0.0**	83.3	**50.0**
Hey Alexa make a donation to the American Cancer Institute	25.0	0.0	**0.0**	0.0	8.3	**0.0**	58.3	**41.7**
< Average across the above 7 commands >	48.8	36.9	**8.3**	21.4	23.8	**1.2**	69.0	**52.4**

- **Attack success rate (ASR) on WeChat**
  - Test on 12 English speakers (7 males/5 females)
  - The ASR is decreased from 41.6% to 8.3% for SV2TTS

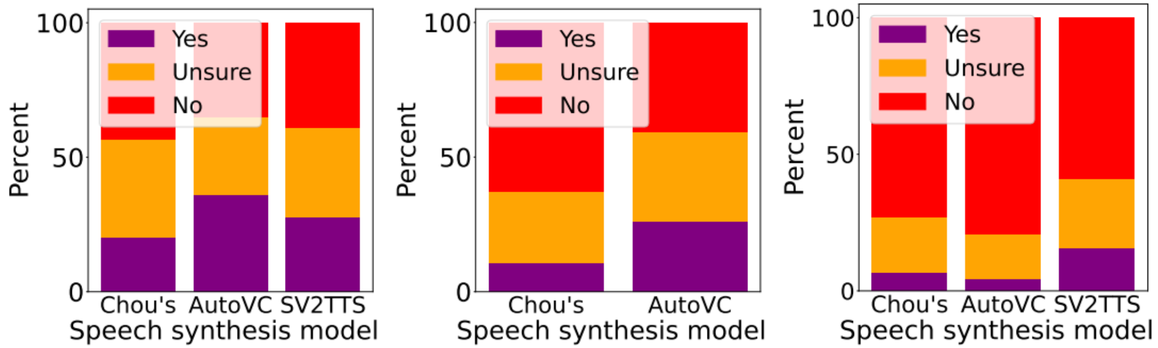
Performance Evaluation

- **User study** (80 participants from Amazon Mechanical Turk)
  - Each participant is asked to listen to some audio pairs and answer the question: **Are the two audio samples from the same speaker?**
  - **Real A/Defense A** (one real speech sample and its corresponding defense sample)
  - **Real A/Fake A** (one real speech sample and its corresponding synthetic speech sample)

Real A/Defense A →

	Chou's		AutoVC		SV2TTS
	Attack-VC	SpeakerMask	Attack-VC	SpeakerMask	SpeakerMask
Yes (%)	70.9	71.5	70.4	69.9	73.7
Unsure (%)	13.1	12.8	16.6	13.5	15.4
No (%)	16.0	15.7	13.0	16.6	10.9

Real A/Fake A →



Raw   Attack-VC   Ours

Conclusions

- We study how to protect a speaker's voice from speech synthesis attacks.
- We propose a novel defense scheme that can significantly degrade the performance of existing speech synthesis models.
- The proposed defense scheme has little impact on the quality of speeches, and the modified speeches can still be used for their normal purposes.
- The desirable performance of the proposed defense schemes is verified on several real-world speaker recognition systems and a user study on a public crowdsourcing platform.

Thank You !