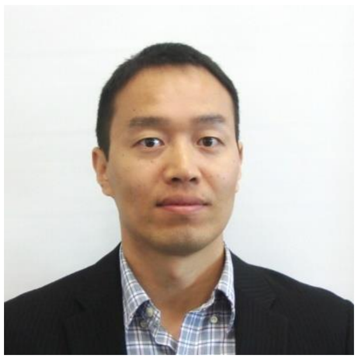
VoicePM: A Robust Privacy Measurement on Voice Anonymity

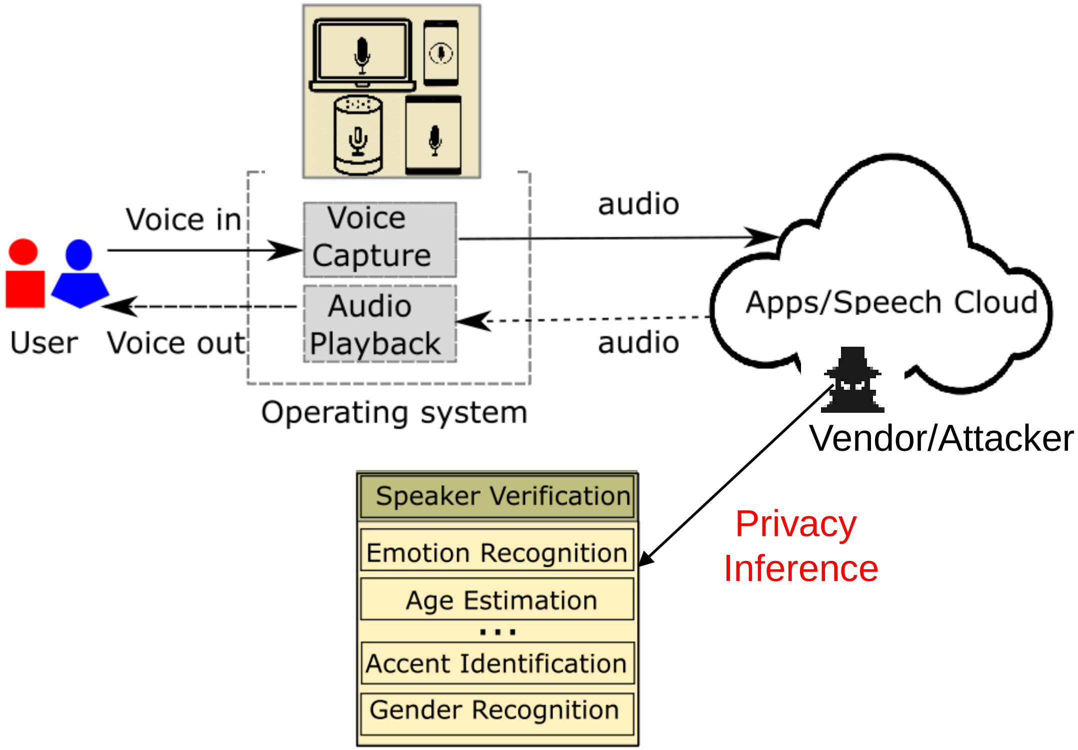

Shaohu Zhang


Zhouyu Li


Anupam Das

Department of Computer Science
North Carolina State University

Data Harvesting of Sensitive Voice Data



Typical data flow in a voice assistant

Examples of Sensitive Data Harvesting


ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

DR. ALEXA —

Amazon patents Alexa tech to tell if you're sick, depressed and sell you meds

Echo could analyze your voice to detect a "physical or emotional abnormality."

JON BRODKIN - 10/11/2018, 1:13 PM

Source: https://shorturl.at/dKOZ2


How Is McDonald's AI System Really Faring At The Drive-Thru Window?

Gargantiopa/Shutterstock

BY HOPE NGO / JUNE 26, 2022 4:34 AM EST

McDonald's had big goals when it first rolled out its voice-ordering technology on a trial basis at 10 Chicago outlets, per Restaurant Dive. The chain had been hoping that AI would help it find a way out of its labor shortage, an issue that has plagued many fast food companies as far back as 2018, per Business Insider. The company said it even offered the hope that – at least where U.S. corporate locations were concerned – AI would allow it to raise wages for entry-level workers from $11 to $17, per Restaurant Dive.

Source: https://shorturl.at/aBEN2

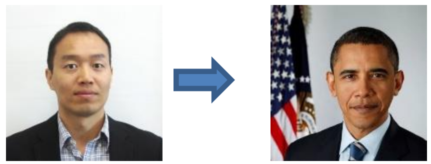
Voice Privacy Challenge in 2020 and 2022

- hide the speaker's identity (maximize speaker verification equal error rate)
- preserve the speech utility (minimize word error rate)

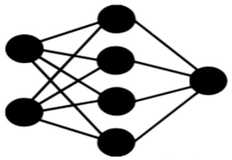
State-of-the-art Anonymization Models

- **Signal processing**
(McAdams Interspeech'21
& VoiceMask/VTLN Sensys'18)

- **Voice synthesis**
(HiFi-GAN NIPS'20)



- **Voice conversion**
(MaskCycleGAN ICASSP'21)



- **Voice adversarial examples**
(V-CLOAK Security'23)

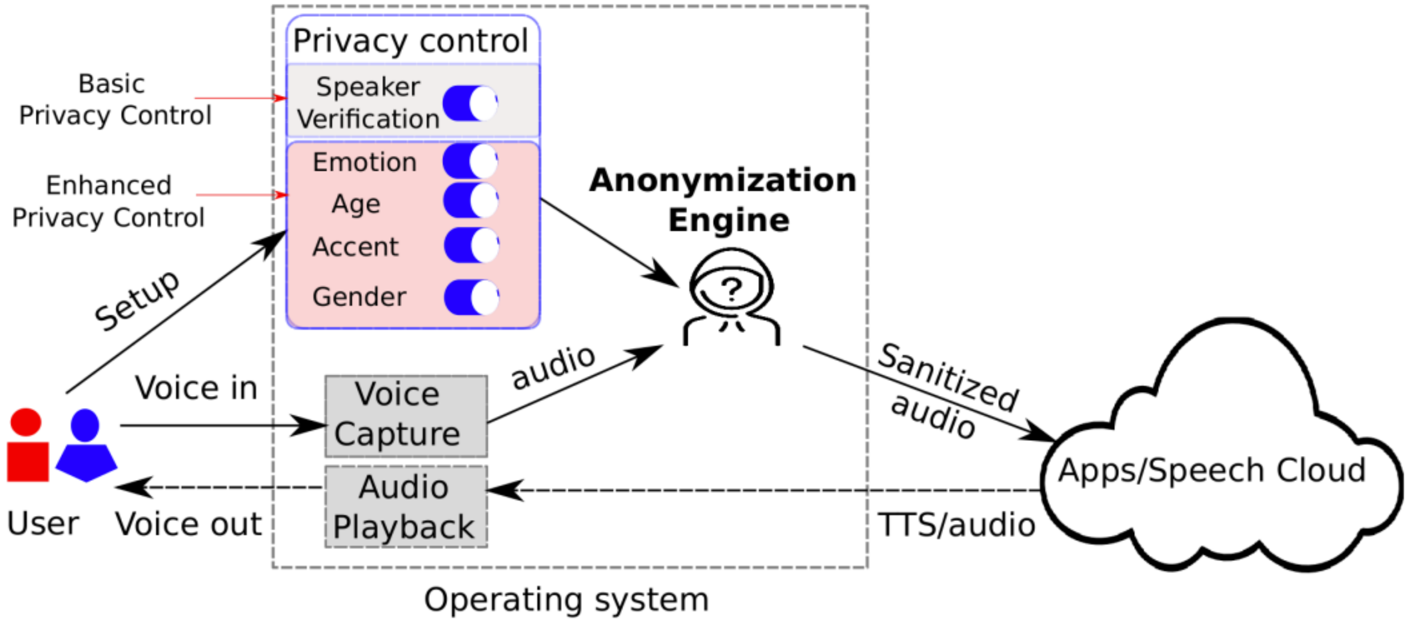
Limitations of Existing Voice Anonymization Approaches

- Limited to analyzing only one or two voice-based attributes
  - Lack a systematic framework to analyze multiple attributes (Aloufi et al., 2020 and Zhu et al., 2021)

- Do not consider the overall tradeoff between speech utility, speaker verification, and inference of voice attributes. (Tomashenko et al., 2022)

Aloufi et al., 2020, Privacy-preserving voice analysis via disentangled representations
Zhu et al., 2021, Anti Leakage: Protecting Privacy Hidden in Our Speech
Tomashenko et al., 2022, The VoicePrivacy 2020 Challenge: Results and findings

Our Design: VoicePM



VoicePM, a robust **Voice P**rivacy **M**easurement on the state-of-the-art of voice anonymization solutions

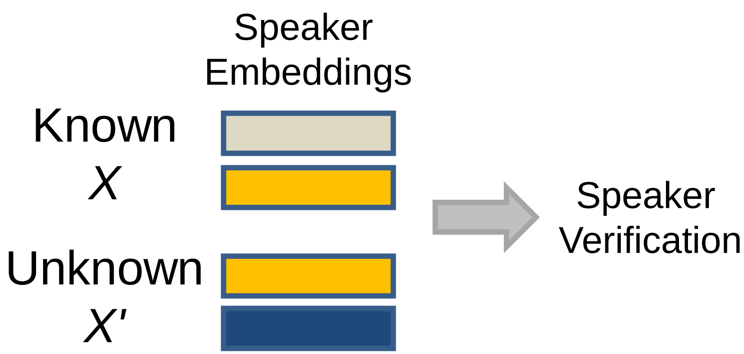
- Incorporate into the operating system
- Provide flexibility to configure the privacy level
- **Preserve transcription utility, hide speaker verification**, and **thwart voice attribute inference**.

Research Objectives

- **R1**: Can we formulate privacy-utility tradeoff to consider speech utility, speaker verification, and inference of physical attributes?

- **R2**: Can we obtain practical level of privacy-utility tradeoff for different voice anonymization techniques?

Threat Model



- **Linkage Attack**: similarity score to decide utterances are from the same speaker
- **Attributes Inference Attack**: identify the speaker's accent, emotion, age, gender, etc.

Speech Utility Metric

- **Word Error Rate (WER)**

$$\text{WER} = \frac{N_{sub} + N_{del} + N_{ins}}{N_{ref}}$$

N_{sub}: # of substitution
N_{del} : # of deletion
N_{ins} : # of insertion
N_{ref}: # of ground truth

$$\text{U} = \frac{1 - WER_{model}}{1 - WER_{baseline}}$$

$WER_{baseline}$ is the WER for the original speech in a database
WER_{model} is for the anonymized speech
Speech Utility $U \in [0, 1]$

Speech Privacy Metric

- **Speaker Verification: *S***

$$S = \frac{EER_{model} - EER_{baseline}}{EER_{model}}$$

Equal Error Rate (EER)
$EER_{baseline}$: EER for the original database
EER_{model}: EER between clean speech and sanitized speech generated by the anonymization model

- **Jaccard Similarity**

$$J(A, A') = \frac{A \cap A'}{A \cup A'}$$

$$J = \frac{J_{model}(A, A')}{J_{baseline}(A, A')}$$

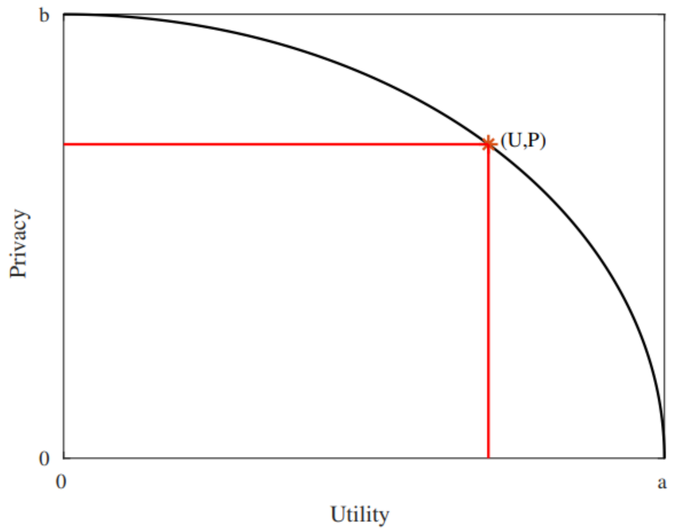
A: set of voice attributes of the original speaker
A': set of inferred voice attributes from the recorded audio

- **Formalizing Privacy Metric: *P***

$$P = \gamma S + (1 - \gamma)(1 - J) \qquad P \in [0, 1]$$

where $\gamma \in (0, 1]$ and prioritizes the individual components within P

Privacy vs. Utility Tradeoff



The typical relationship between privacy and utility

- **Maximize the Tradeoff : *T***

$$U = \frac{1 - WER_{model}}{1 - WER_{baseline}} \qquad U \in [0, 1]$$

$$P = \gamma S + (1 - \gamma)(1 - J) \qquad P \in [0, 1]$$

$$T(S, J, U) = P \times U \qquad T \in [0, 1]$$

$\gamma \in (0, 1]$ is the weight of S and J

Theorem:

Privacy increases while the utility decreases. There exists a point (U, P) where the P and U form a rectangle with the highest area/tradeoff T.

Evaluation Setup

- **Dataset**
  - Mozilla Common Voice (English, 83,242 samples, 7,499 speakers)
  - IEMOCAP (Emotion, English, 5,531 samples, 10 speakers)
  - AISHELL-1 (Mandarin Chinese, 7,176 samples, 400 speakers)

Accents	Alias	# of samples	# of speakers	Length (hrs)
United States	US	10000	2683	13.78
England	EN	10000	1343	13.17
India and South Asia	INSA	10000	1450	13.26
Canadian	CA	10000	649	13.28
Australian	AU	10000	534	12.98
New Zealand	NZ	8514	138	10.80
Scottish	SC	7995	141	11.13
Ireland	IE	6052	164	7.93
Southern African	SA	5794	112	3.26
Chinese	CN	4887	285	10.74

Mozilla Common Voice English Dataset Summary

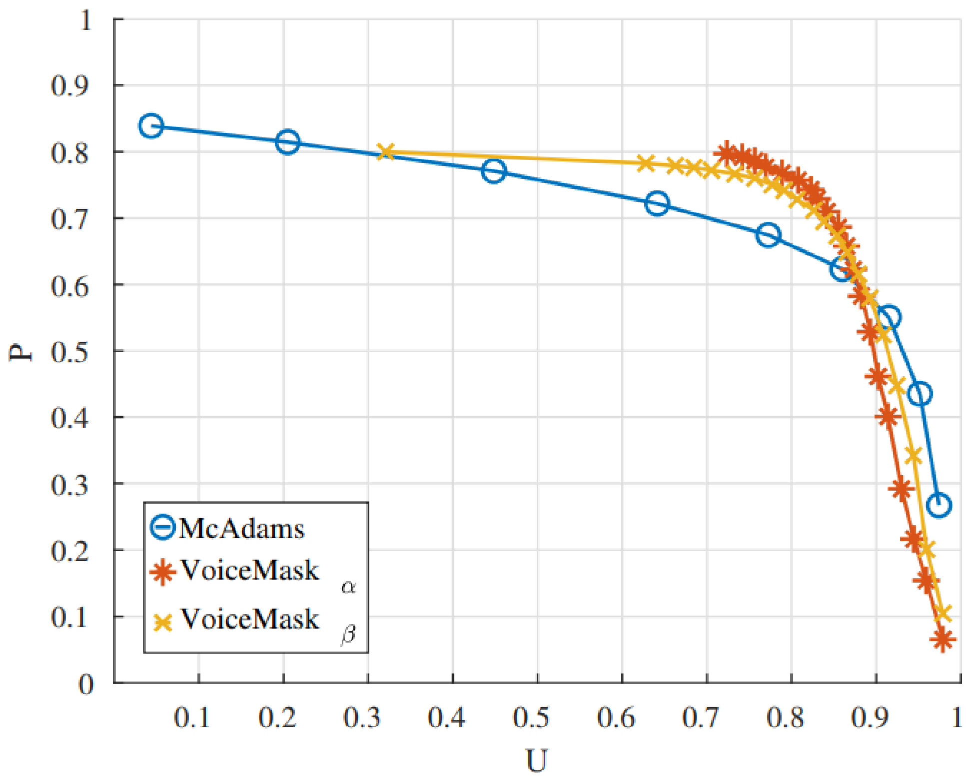
Baseline of Attribute Inference Models

Attributes	Test set (# of utterances)	wav2vec2 Base	ECAPA-TDNN
Emotion	happiness (167), anger (122) sadness (113), neutral (149)	77.31%	65.15%
Age	teens (876), twenties (2,799) thirties (1,703), forties (1,601) fifties (783), senior (563)	85.36%	80.95%
Accent	AU (969), NZ (872), CN (480), SA (609) INSA (1,006), CA (1,005), EN (1,013) IE (630), SC (797), US (944)	87.72%	82.10%
Gender	male (6,562), female (1,763)	99.06%	97.87%

- wav2vec2: 90.2 million parameters (emotion and accent model)
- ECAPA-TDNN: 5.5 million parameters (gender and age model)

ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network

Privacy vs. Utility Tradeoff Relationship

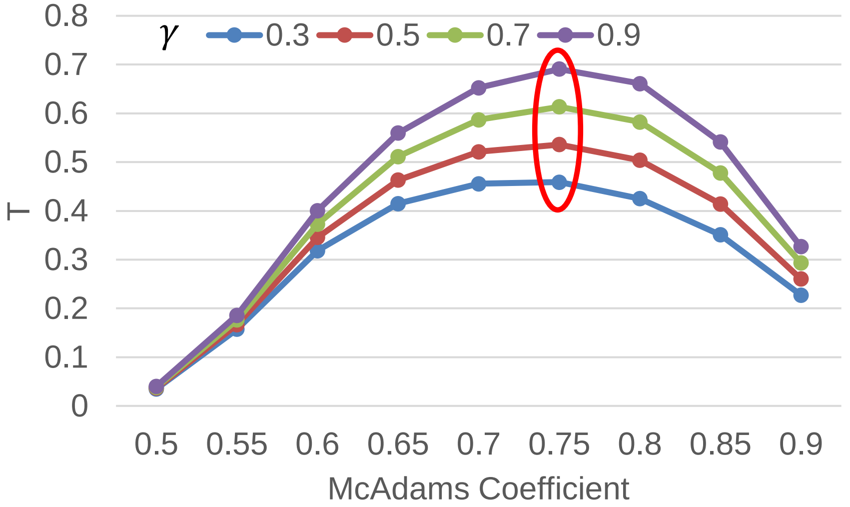


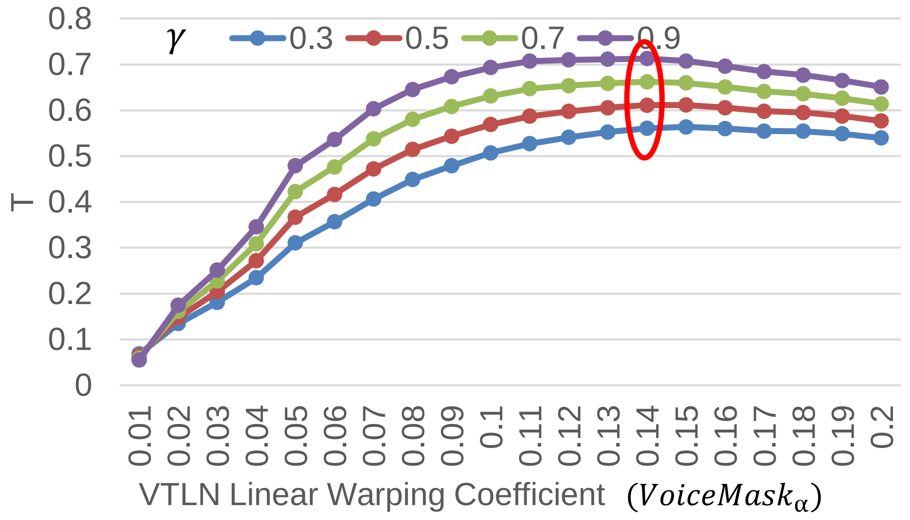
Privacy and utility form a non-linear pattern like an arc

- McAdams: 0.5 ~ 0.9
- $VoiceMask_{\alpha}$: VTLN Linear Warping Coefficient $|\alpha| \in [0.01, 0.2]$
- $VoiceMask_{\beta}$: VTLN Quadratic function Warping Coefficient $|\beta| \in [0, 1]$

Determining the Impact of γ on Tradeoff

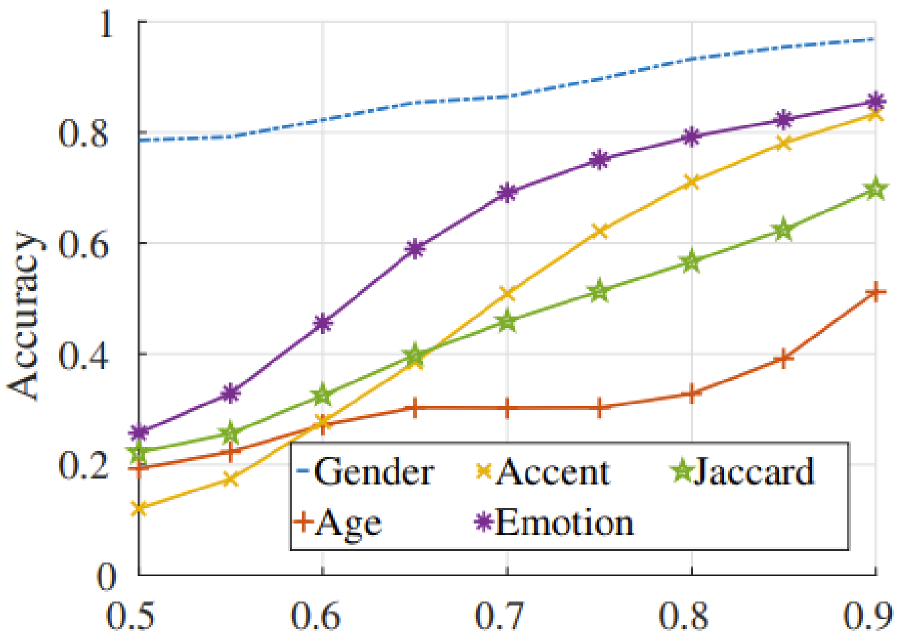
$$T\ (S, J, U\ ) = \mathrm{P} \times \mathrm{U} = [\gamma S + (1-\gamma)\ (1 - J)] \times \mathrm{U}$$





γ changes the tradeoff *T* and the optimum coefficient

Controlling Voice Attribute: McAdams Coefficient



McAdams Coefficient

Sentence:
What are you talking about?

Attributes of an original speech:
[thirties, male, Chinese, anger]

Attributes of anonymized speech (coefficient of 0.75):
[thirties, male, New Zealand, anger]

- Gender inference changes slightly
- Accent varies significantly

Measurement of Different Voice Anonymity Systems

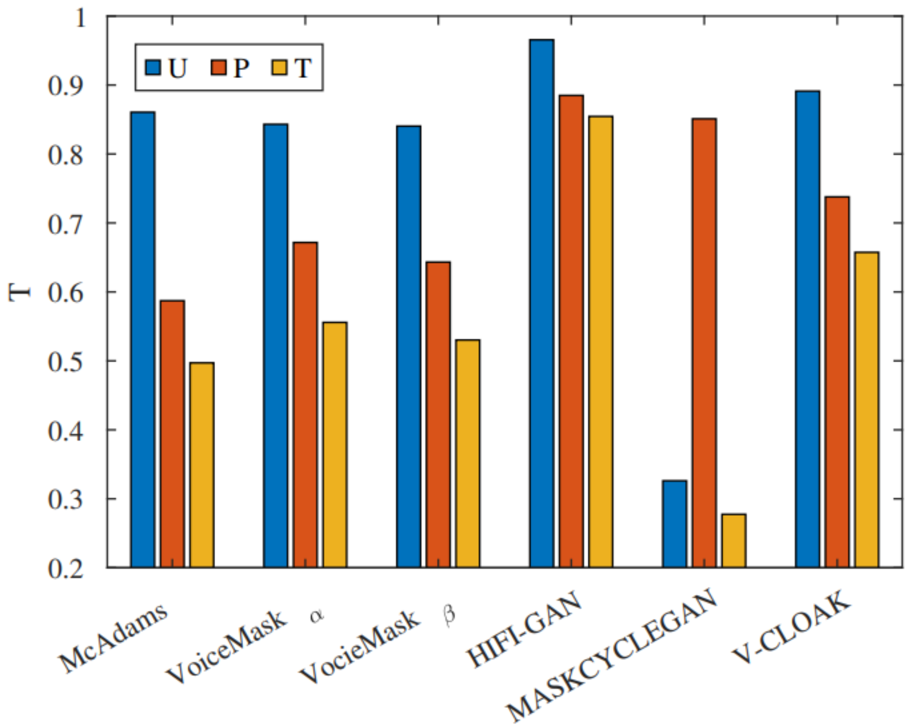
Model	Emotion	Age	Accent	Gender	Jaccard	EER	WER	U	P	T
Baseline	100	80.95	87.94	97.87	0.8534	2.28	14.5	1	0	0
McAdams	76.07	35.24	62.96	90.14	0.5386	18.39	26.42	0.8606	0.5872	0.4971
$VoiceMask_{\alpha}$	71.97	37.25	49.89	50.67	0.4038	20.58	27.92	0.8431	0.6717	0.5558
$VoiceMask_{\beta}$	71.7	36.34	54.2	67.45	0.4534	20.85	28.15	0.8404	0.6432	0.5303
HiFi-GAN	40.5	19.39	12.13	24.28	0.1561	48.32	17.44	0.9656	0.8849	0.8545
MaskCycleGAN	36.18	24.21	19.32	40.25	0.2056	39.95	72.12	0.3261	0.8510	0.2775
V-CLOAK	60.54	25.13	51.08	81.26	0.4107	52.79	23.81	0.8911	0.7378	0.6574

Different anonymity systems hide the attributes in different levels

Optimum tradeoff from high to low:
HiFi-GAN → V-CLOAK → VoiceMask → McAdams → MaskCycleGAN

Privacy-Utility Tradeoff for Different Voice Anonymity Systems

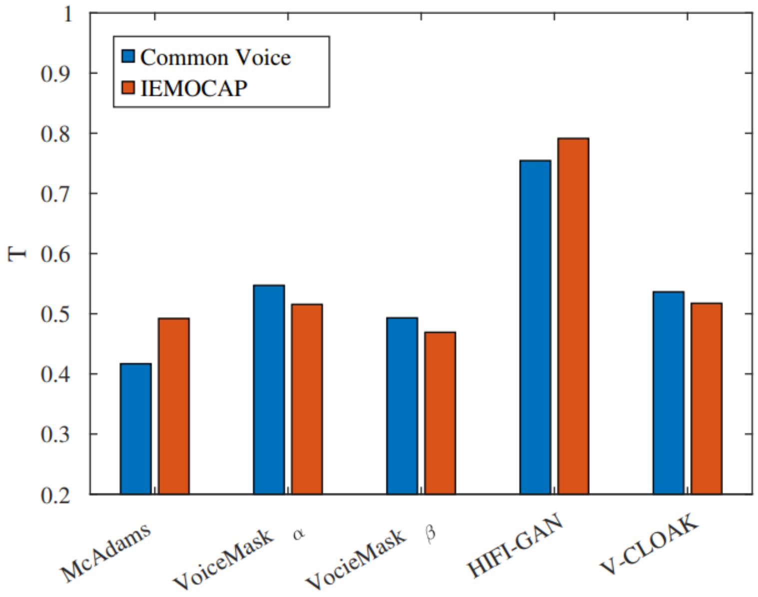
“Right, it shouldn’t be too difficult to rework them”



Attributes of an original speech:
[forties, male, New Zealand, neutral]

HiFi-GAN:
[thirties, female, Canadian, anger]

Generalizability Across Different English Dataset



γ = 0.5

- Include gender and emotion attributes
- Overall relative ranking does not change with English dataset

Key Takeaway

***VoicePM* provides a voice privacy measurement framework:**

- effectively measure the tradeoff of different anonymization models
- anonymization models with varying privacy levels can be pre-defined
- showcases the feasibility for attributes configuration

Limitations

- Accuracy of the emotion (77.31%) and age (80.95%) inference model is relatively low
- Lacks human perception verification of the altered audio

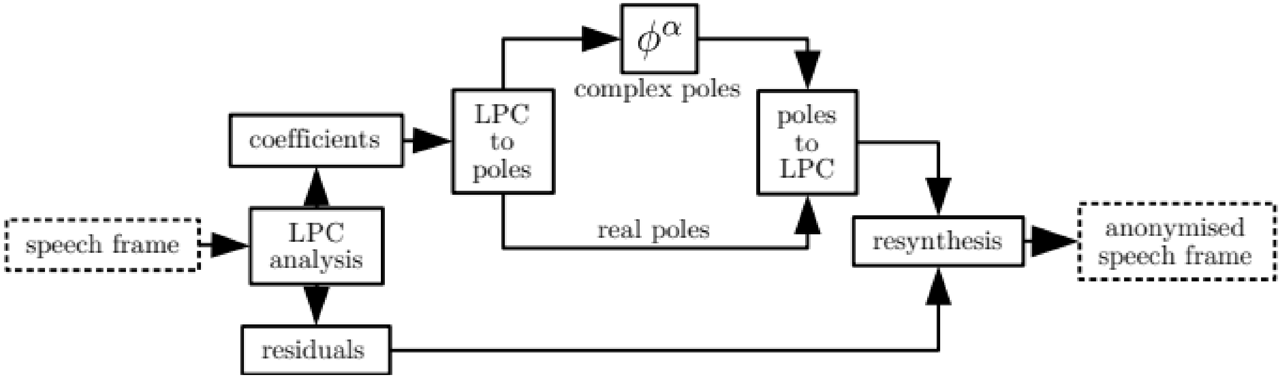
VoicePM: A Robust Privacy Measurement on Voice Anonymity.
Shaohu Zhang, Zhouyu Li, Anupam Das. 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks. ACM **WiSec'23**

Thank you!

VoicePM Project Website
(Code will be released soon)
https://github.com/zhangshaohu/VoicePM

Backup Slides

McAdams

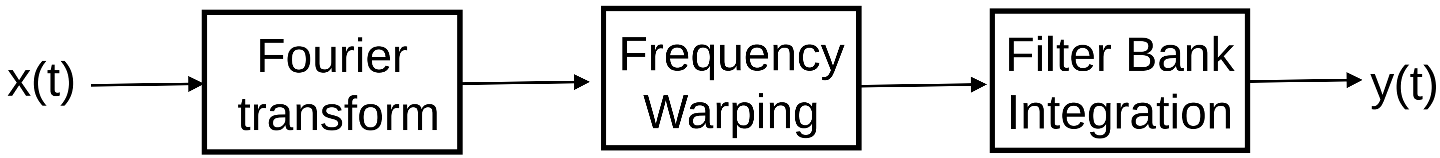


$$X(t) = \sum_{k=1}^{K} A_k(t) \cos\left(2\pi (k f_0)^\alpha t + \phi_k\right)$$

where k is the harmonic index, $A_k\,(t)$ is signal amplitude, ϕ_k is the phase, and α is the McAdams coefficient, which is usually in the range of [0.5, 1].

Patino et al., 2020. Speaker anonymization using the McAdams coefficient
linear predictive coding (LPC)

Vocal Tract Length Normalization (VTLN)



- Bilinear warping function ($\boldsymbol{VoiceMask_{\alpha}}$)

$$\varphi_{\alpha} = \omega + 2\arctan^{-1}\left(\frac{(1-\alpha)\sin\omega}{1-(1-\alpha)\cos\omega}\right)$$

where $\omega \in [0, \pi]$ is the normalized frequency, and $\alpha \in (-1, 1)$ is a warping factor used to tune the strength of voice conversion.

- Quadratic function ($\boldsymbol{VoiceMask_{\beta}}$)

$$\varphi_{\beta} = \omega + \beta\left(\frac{\omega}{\pi} - \left(\frac{\omega}{\pi}\right)^2\right)$$

where $\beta \in (-1, 1)$ is the warping factor.

HiFi-GAN

- synthesize high-fidelity waveforms from Mel-spectrograms
- convert the voice to a pre-defined speaker

MaskCycleGAN-VC

- non-parallel VC technique
- apply a temporal mask to the input Mel-spectrogram

V-CLOAK

- add imperceptible noises to audio
- generate adversarial examples to fool speaker verification system

Selection of Automatic Speech Recognition Systems

Model	Source	Language	Dataset	WER(%)
wav2vec2+CTC	SpeechBrain	English	CV	14.50
CRDNN + CTC/Attention	SpeechBrain	English	CV	25.90
DeepSpeech	DeepSpeech	English	CV	27.09
Google Speech2Text	Google Cloud	English	CV	28.19
wav2vec2+CTC	SpeechBrain	English	IEMOCAP	24.57
CRDNN + CTC/Attention	SpeechBrain	English	IEMOCAP	37.15
Google Speech2Text	Google Cloud	English	IEMOCAP	37.76
wav2vec2+CTC	SpeechBrain	Mandarin Chinese	AISHELL1-test	5.04
Transformer	SpeechBrain	Mandarin Chinese	AISHELL1-test	6.04
Google speech2text	Google Cloud	Mandarin Chinese	AISHELL1-test	7.69

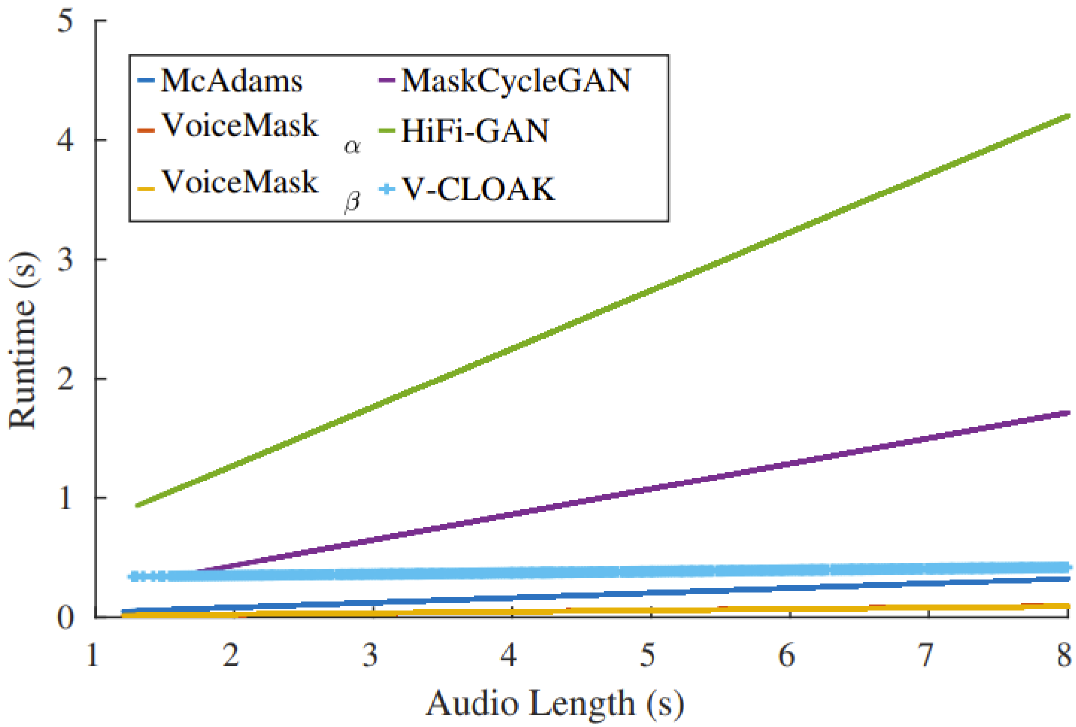
Performance of Different ASR Systems

Feasibility for Attributes Configuration

Attributes	McAdams (U=0.8466) P	McAdams T	VoiceMask$_\alpha$ (U=0.8274) P	VoiceMask$_\alpha$ T	VocieMask$_\beta$ (U=0.8245) P	VocieMask$_\beta$ T	HiFi-GAN (U=0.9130) P	HiFi-GAN T	MaskCycleGAN (U=0.3261) P	MaskCycleGAN T	V-CLOAK (U=0.8911) P	V-CLOAK T
basic privacy	0.4431	0.3752	0.4488	0.3714	0.4493	0.3704	0.4764	0.4350	0.4715	0.1538	0.4784	0.4263
emotion	0.5066	0.4289	0.5420	0.4484	0.5399	0.4451	0.7391	0.6748	0.7595	0.2477	0.6237	0.5557
age	0.7634	0.6463	0.7562	0.6257	0.7627	0.6288	0.8628	0.7878	0.8296	0.2706	0.8312	0.7406
accent	0.5878	0.4976	0.6798	0.5625	0.6529	0.5383	0.9053	0.8266	0.8583	0.2799	0.6791	0.6052
gender	0.4189	0.3547	0.6803	0.5628	0.5729	0.4723	0.8342	0.7616	0.7356	0.2399	0.5023	0.4476
emotion+accent	0.5900	0.4996	0.6609	0.5468	0.6436	0.5306	0.8638	0.7886	0.8445	0.2754	0.7015	0.6251
emotion+age	0.6936	0.5872	0.7040	0.5825	0.7081	0.5838	0.8438	0.7704	0.8373	0.2731	0.7835	0.6981
emotion+gender	0.4924	0.4169	0.6612	0.5471	0.5981	0.4931	0.8264	0.7545	0.7861	0.2564	0.6018	0.5363
age+accent	0.7265	0.6151	0.7651	0.6331	0.7566	0.6237	0.9105	0.8313	0.8744	0.2852	0.8024	0.7149
gender+accent	0.5428	0.4596	0.7286	0.6029	0.6579	0.5424	0.8985	0.8203	0.8363	0.2727	0.6386	0.5690
gender+age	0.6548	0.5544	0.7643	0.6324	0.7196	0.5933	0.8823	0.8056	0.8264	0.2695	0.7312	0.6515
emotion+age+accent	0.6889	0.5832	0.7275	0.6020	0.7208	0.5943	0.8825	0.8058	0.8624	0.2812	0.7797	0.6948
emotion+accent+gender	0.5577	0.4722	0.7013	0.5803	0.6503	0.5361	0.8731	0.7971	0.8337	0.2719	0.6655	0.5930
emotion+age+gender	0.6347	0.5373	0.7274	0.6019	0.6939	0.5721	0.8620	0.7871	0.8296	0.2706	0.7257	0.6466
gender+age+accent	0.6611	0.5597	0.7679	0.6354	0.7283	0.6005	0.9047	0.8260	0.8569	0.2795	0.7431	0.6622
emotion+age+accent+gender	0.6454	0.5464	0.7393	0.6117	0.7072	0.5830	0.8849	0.8080	0.8510	0.2775	0.7378	0.6574

Emotion Ranking: HiFi-GAN → V-CLOAK → VoiceMask → McAdams → MaskCycleGAN

Runtime



From low to high:
VoiceMask→ McAdams → V-CLOAK → MaskCycleGAN → HiFi-GAN