Privacy-preserving Voice Analysis via Disentangled Representations

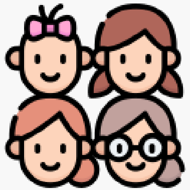
Ranya Aloufi, Hamed Haddadi and David Boyle

CCSW'20: Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, November 2020

Voice Privacy Issues

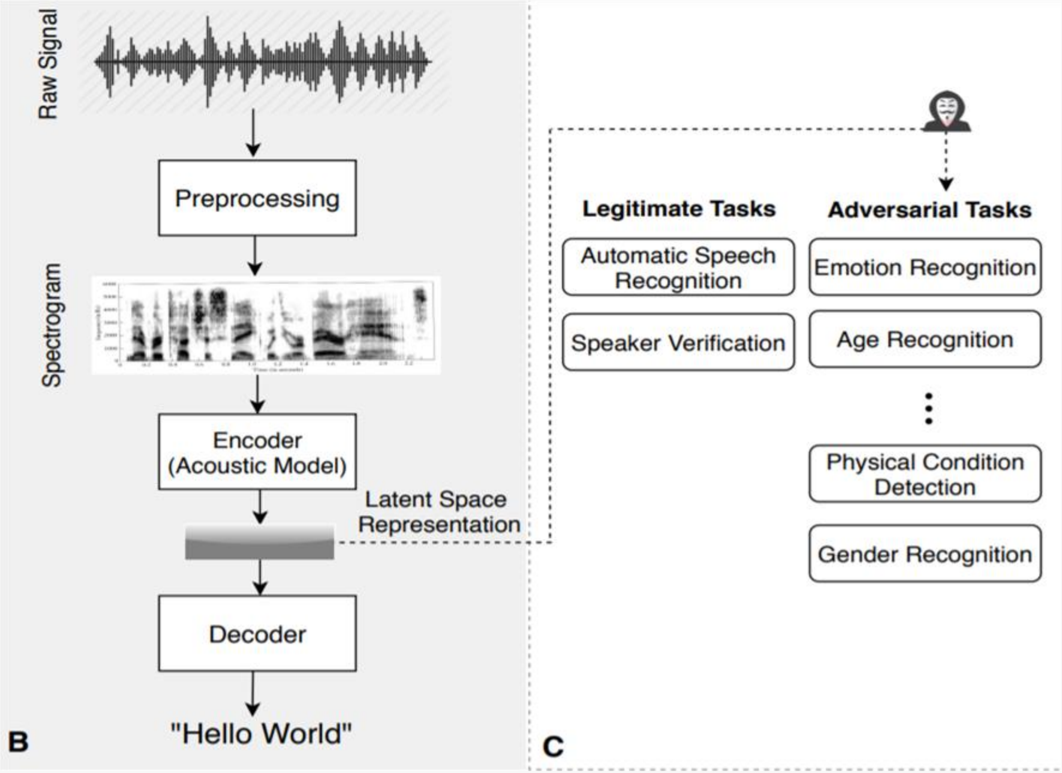
Emotion

Health conditions



Age, gender, accent

Deep Learning Latent Representations



(Aloufi et al., 2020, Figure 1)

Paper's Contributions

1. Show inference attacks of **emotion**, **identity** and **gender** on commonly used acoustic models

1. Propose a privacy-aware framework with different levels of privacy. The framework is based on a quantized variational autoencoder model and disentanglement learning.

1. Evaluate their framework on 5 different datasets

Scenario and Threat Model

1. User shares voice recordings with cloud service providers to accomplish a certain task but do not wish to share additional attributes..

1. Attacker (service provider, surveillance agency, advertiser) wants to infer sensitive attributes to track the user, advertise to them or sell their data.

Research Questions

1. To what extent can an attacker infer sensitive attributes?

1. Can we build a effective defense?

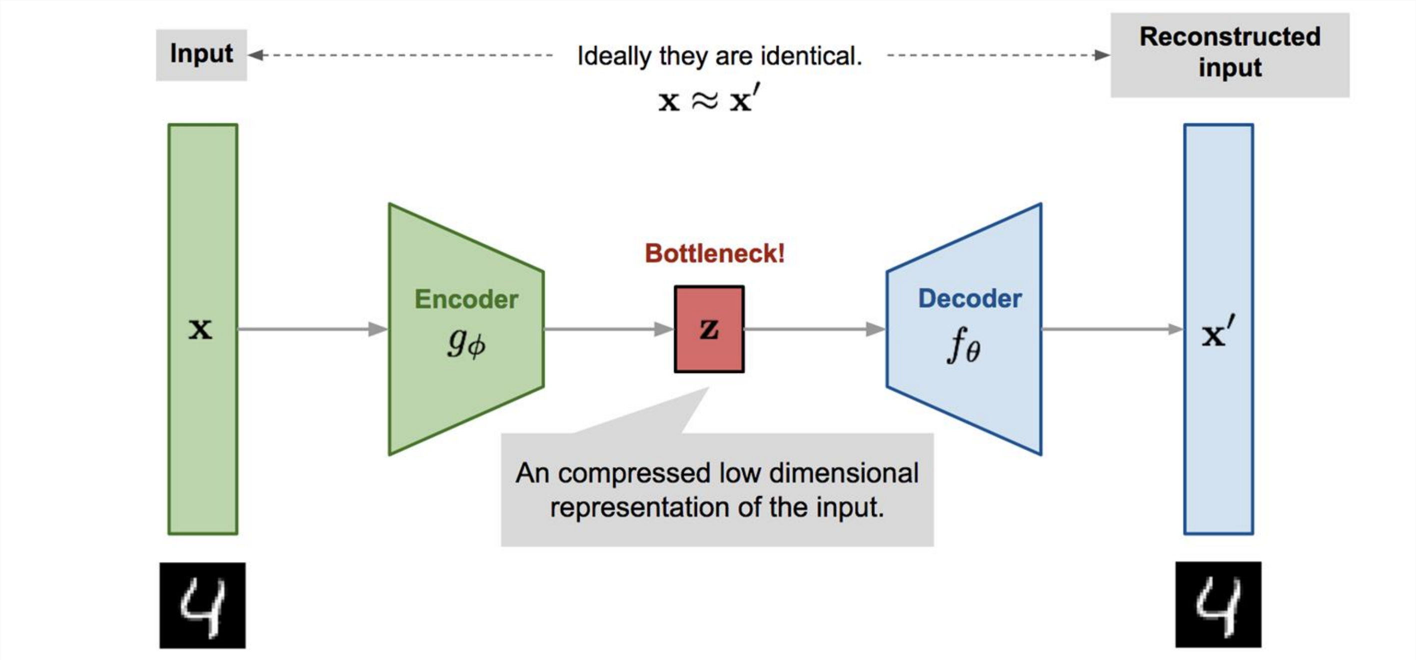
Background : Disentanglement

Learning technique to separate representations

- Computer vision: body pose or shape, face shape, make up
- NLP: syntax and semantics
- Speech : content, accent, prosody, emotion, language, environment

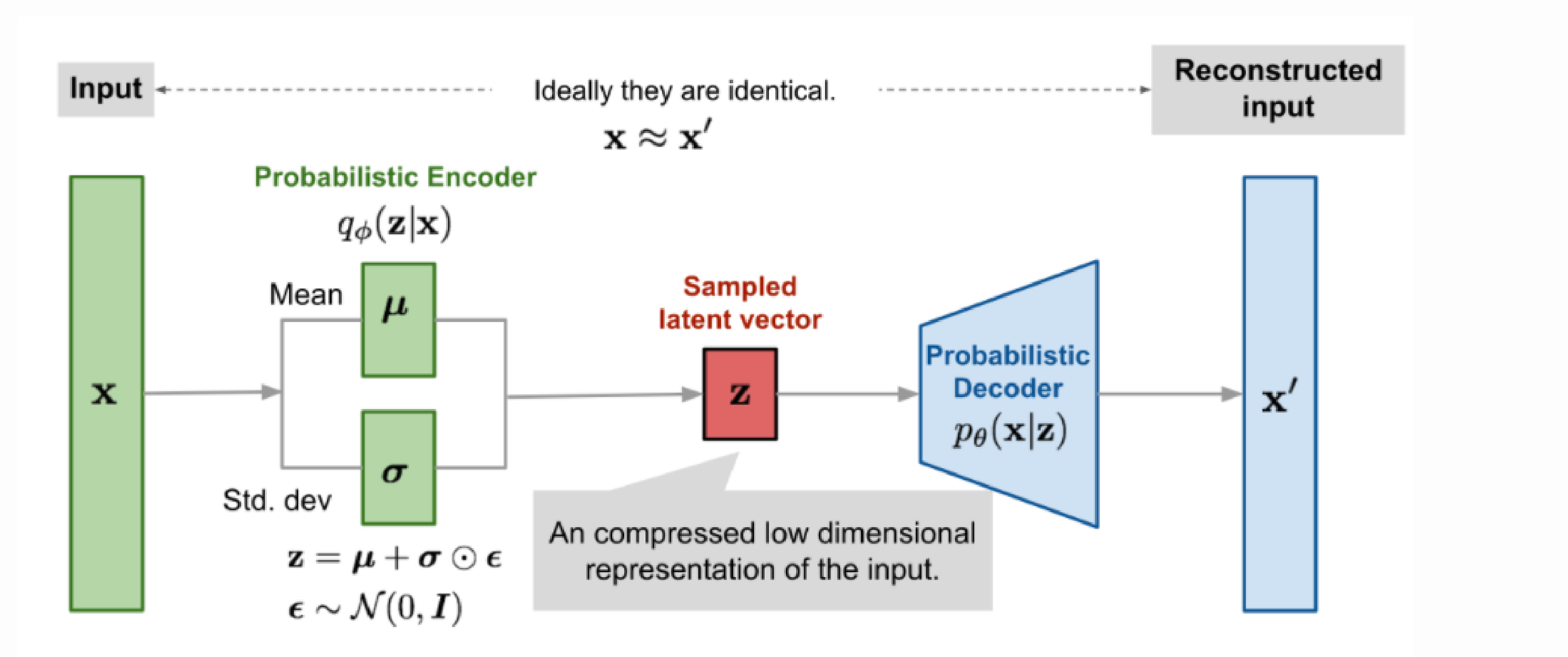
7

Background : Auto Encoder



https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

Background : Variational Auto Encoder



https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

Background : Variational Autoencoder

$$z = latent\ vector$$
$$x = input\ data$$

$$Encoder = q_\theta(z \mid x)$$
$$Decoder = p_\varphi(x \mid z)$$
$$p(z) = Normal(0, 1)$$

$$Loss = Reconstruction\ Loss + Regularizer$$
$$Loss = E_{q_\theta(z|x)}[log\ p_\varphi(x \mid z)] - KL(q_\theta(z|x) \mid\mid p(z))$$

https://medium.com/ai-deep-learning-tutorials/variational-autoencoders-an-intuitive-explanation-some-keras-code-bc0362bc8113

Framework

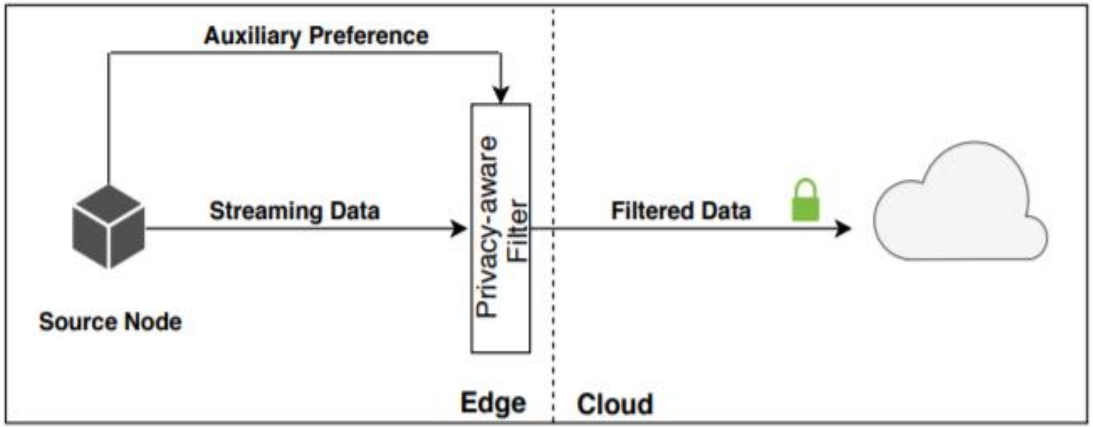


Figure 2: The workflow of the proposed framework: it serves as a filter between the edge and the cloud to purify data from a source node based on an auxiliary user preference

(Aloufi et al., 2020, Figure 2)

General Framework

1. Given user preference, map it into into *n* tasks
2. Models
   a. For each task, build a specific encoder branch
   b. Decoder: vocoder to concatenate features from each branch and reconstructs speech.

Proposed Framework

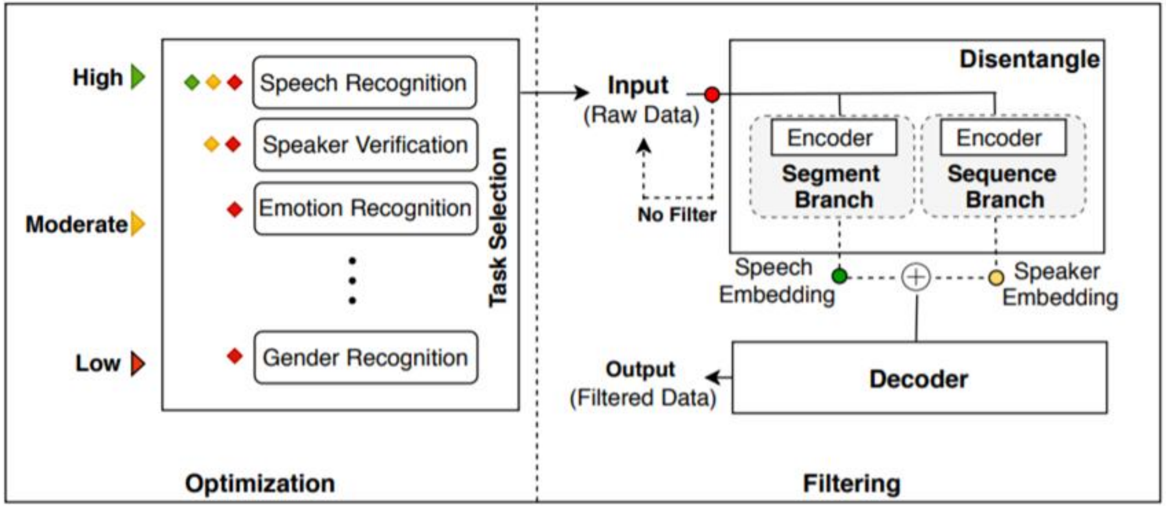
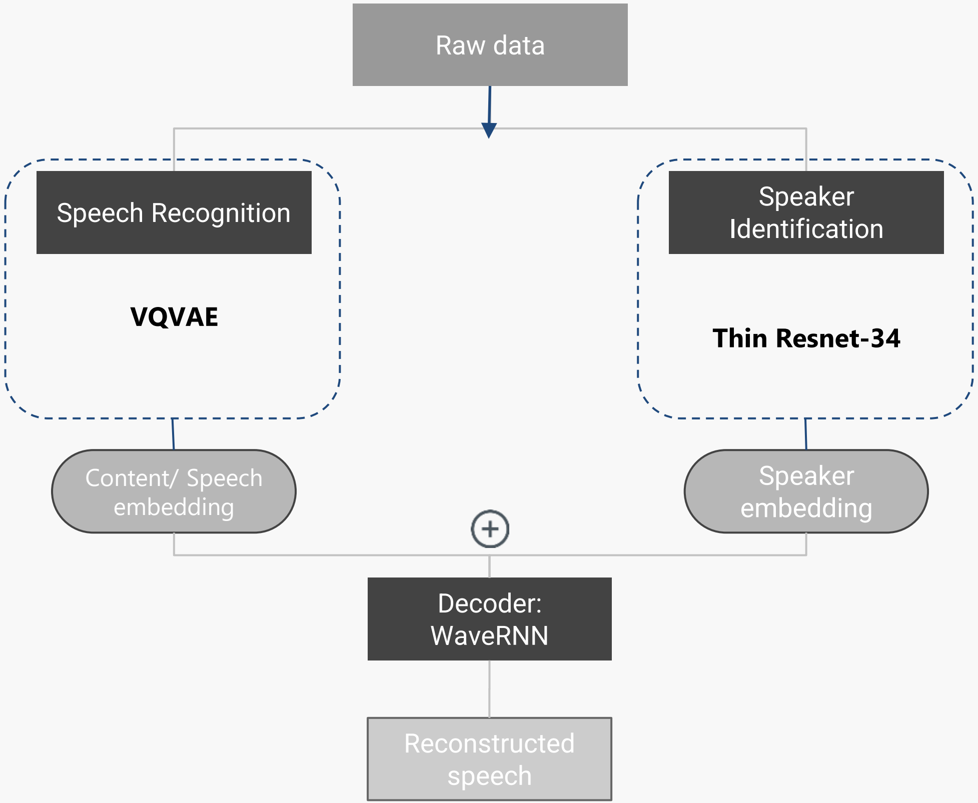


Figure 3: The proposed framework begins by adjusting the privacy preferences (high, moderate, and low; left) that are used as a control signal to extract the corresponding representations and reconstruct the output (right)

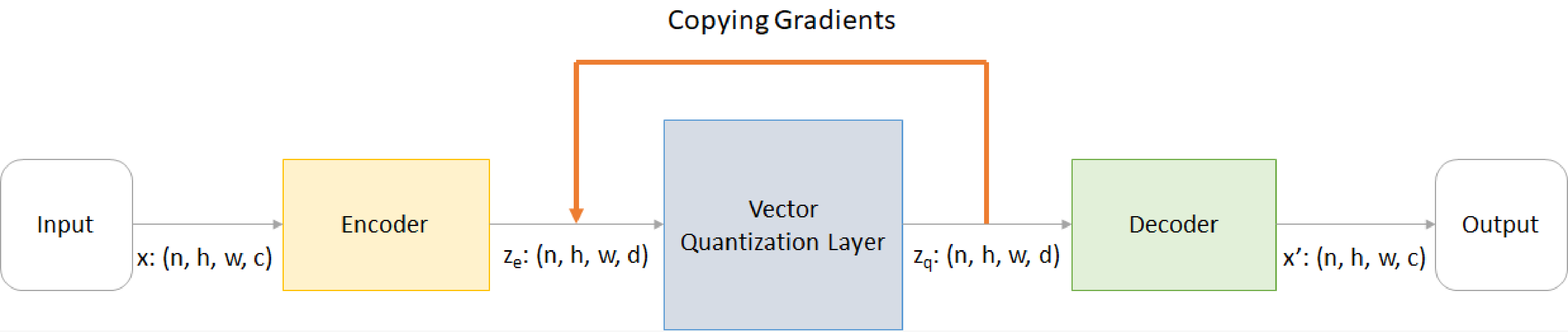
(Aloufi et al., 2020, Figure 3)

Proposed Framework: Disentanglement



Speech Recognition Task: Vector Quantized Variational Autoencoder

Learns discrete latent representations by mapping the output of the encoder to the closest vector from a codebook of K vectors.



https://blog.usejournal.com/understanding-vector-quantized-variational-autoencoders-vq-vae-323d710a888a

Speech Recognition Task: Vector Quantized Variational Autoencoder

Learns discrete latent representations by mapping the output of the encoder to the closest vector from a codebook of K code vectors.

$$L = \underbrace{\|\mathbf{x} - D(\mathbf{e}_k)\|_2^2}_{\text{reconstruction loss}} + \underbrace{\|\text{sg}[E(\mathbf{x})] - \mathbf{e}_k\|_2^2}_{\text{VQ loss}} + \underbrace{\beta\|E(\mathbf{x}) - \text{sg}[\mathbf{e}_k]\|_2^2}_{\text{commitment loss}}$$

- ***VQ loss*****:** The L2 error between the embedding space and the encoder outputs.
- ***Commitment loss*****:** A measure to encourage the encoder output to stay close to the embedding space and to prevent it from fluctuating too frequently from one code vector to another.

https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

Speaker Verification Task: Thin Resnet-34

1. CNN model trained to learn speaker embeddings
2. Trained on Voxceleb, speaker identification task

Experiments

1. Attribute inference attack on representations extracted from pretrained acoustic models

1. Defense efficiency of the framework

18

Attributes Per Dataset

1. **Emotion and gender:** IEMOCAP (12h, 4 emotions), RAVDESS (1,440 recordings, 7 emotions)
2. **Emotion:** SAVEE (480 recordings, 7 emotions)
3. **Gender:** Librispeech (100 hours, audiobooks), VoxCeleb (1,251 celebrities, 1,200 recordings)

Experiment 1: Inference attacks

1. Extract representations from pre-trained wav2vec model and DeepSpeech2 as input features

1. Train Logistic regression, SVM, Random Forest, Multilayer perceptron to infer gender, emotion on both on all 5 datasets

Results

Table 1: Accuracy of attribute inference attack using different acoustic models to extract the representation (G=gender (binary); E=emotion)

	wav2vec Model							DeepSpeech2 Model						
Attacker Model	LibriSpeech	VoxCeleb	SAVEE	IEMOCAP		RAVDESS		LibriSpeech	VoxCeleb	SAVEE	IEMOCAP		RAVDESS	
	G(%)	G(%)	E(%)	G(%)	E(%)	G(%)	E(%)	G(%)	G(%)	E(%)	G(%)	E(%)	G(%)	E(%)
LR	85.8	90.4	62.2	82.9	56.4	99.4	74.4	60	78.3	53.1	58.8	47.7	93	57.2
RF	86.7	80.8	43.2	86.4	55	95.6	61.9	50.7	63.5	42.2	62	50.1	86	53.5
MLP	75.8	78.8	39	76.4	51.2	93.8	64.4	56.7	57.8	40.5	58.4	45.3	95.3	63.2
SVM	76.7	85.6	55.7	85	57.9	94.4	60.2	66.7	73.9	46.2	54.3	55.6	88.4	61

(Aloufi et al., 2020, Table 1)

Experiment 2: Framework evaluation

1. Train the model branches and decoder for speech recognition and speaker identification, using Librispeech dataset

1. Evaluate inference accuracy on reconstructed speech

Results: Gender

Table 3: Success accuracy in inferring the sex attribute after implementing the DDF framework with different privacy preference options (W2V: wav2vector model, DS: DeepSpeech2 model, Mod.:moderate, Rec_m: reconstructed speech with moderate option, Rec_h: reconstructed speech with high option)

	LibriSpeech (%)				VoxCeleb (%)				IEMOCAP (%)				RAVDESS (%)			
	Low		Mod.	High	Low		Mod.	High	Low		Mod.	High	Low		Mod.	High
Attack Model	Raw (w2v)	Raw (DS)	Rec_m	Rec_h	Raw (w2v)	Raw (DS)	Rec_m	Rec_h	Raw (w2v)	Raw (DS)	Rec_m	Rec_h	Raw (w2v)	Raw (DS)	Rec_m	Rec_h
LR	85.8	60	53.8	43.8	90.4	78.3	57.1	54.0	82.9	58.8	55.7	41.5	99.4	93	69.1	48.2
RF	86.7	50.7	55.0	46.6	80.8	63.5	64.2	52.3	86.4	62.2	57.4	48.7	95.6	86	53.4	49.2
MLP	75.8	56.7	52.7	46.9	78.8	57.8	51.1	42.2	76.4	58.4	60.0	44.9	93.8	95.3	67.4	41.7
SVM	76.7	66.7	60.2	54.3	85.6	73.9	62.2	49.7	85	54.3	66.2	47.1	94.4	88.4	55.9	45.6

(Aloufi et al., 2020, Table 3)

Results: Emotion

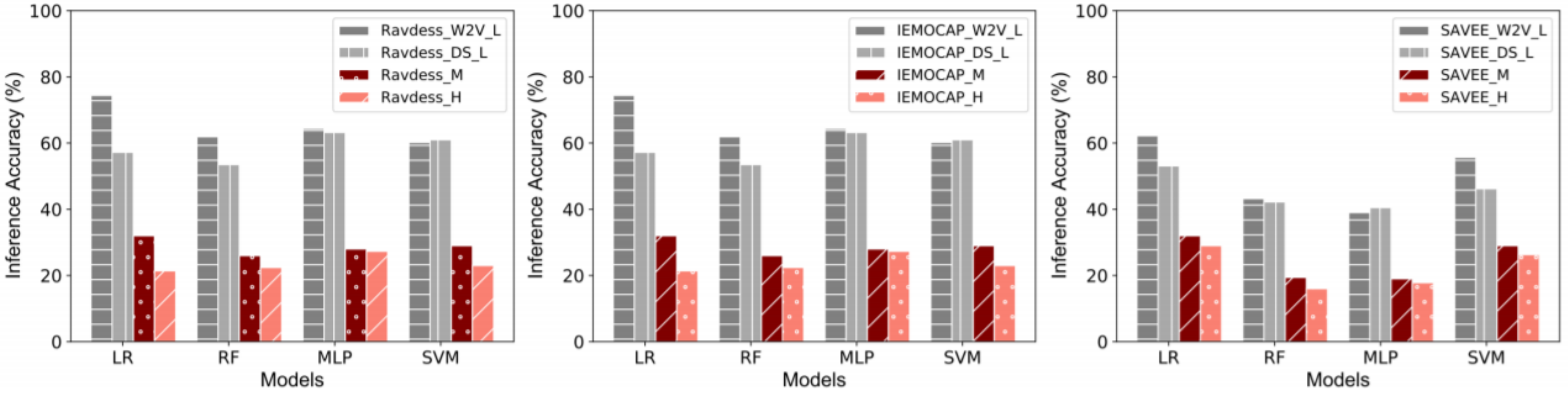


Figure 6: Accuracy in inferring the emotion attribute after implementing the DDF framework with diffrent privacy preference options (W2V: wav2vector model, DS: DeepSpeech2 model, L: low option, M: moderate option, and H: high option)

(Aloufi et al., 2020, Figure 6)

Results

Table 2: Speech recognition and speaker verification measurements for voices generated by the proposed framework with different privacy settings

	Generated (Hide Identity)		Generated (Preserve Identity)	
Dataset	WER (%)	EER (%)	WER (%)	EER (%)
LibriSpeech	1.16	N/A	0.32	0.03
VoxCeleb	0.80	N/A	0.13	0.0
IEMOCAP	0.86	N/A	0.29	0.07
RAVDESS	0.63	N/A	0.14	0.0
SAVEE	0.66	N/A	0.20	0.01

(Aloufi et al., 2020, Table 2)

Examples

https://www.youtube.com/watch?v=hQAcK_Ynun0

Discussion

1. Shows plausibility of inference attacks for gender and emotion using models trained on other tasks

1. Low and medium reduces inference attacks to nearly random

1. Speech recognition performance suffers slightly

Limitations

1. Only acoustic features are considered

1. Requires a new model for each feature type

1. Does not evaluate their approach on other personal attributes (mental/physical abilities, age)

28

1. Quality of audio is altered

Related Work

1. Preech: A system for privacy-preserving speech transcription. *S Ahmed, AR Chowdhury, K Fawaz, P Ramanathan - 29th USENIX Security ..., 2020*
    - *Considers linguistic and acoustic feature.*

1. *Voice-Indistinguishability: Protecting Voiceprint In Privacy-Preserving Speech Data Release.Y. Han, S. Li, Y. Cao, Q. Ma and M. Yoshikawa -IEEE International Conference on Multimedia and Expo (ICME), London, United Kingdom, 2020.*

1. *Paralinguistic Privacy Protection at the Edge. Ranya Aloufi, Hamed Haddadi, David Boyle*
    - *Extension for edge*