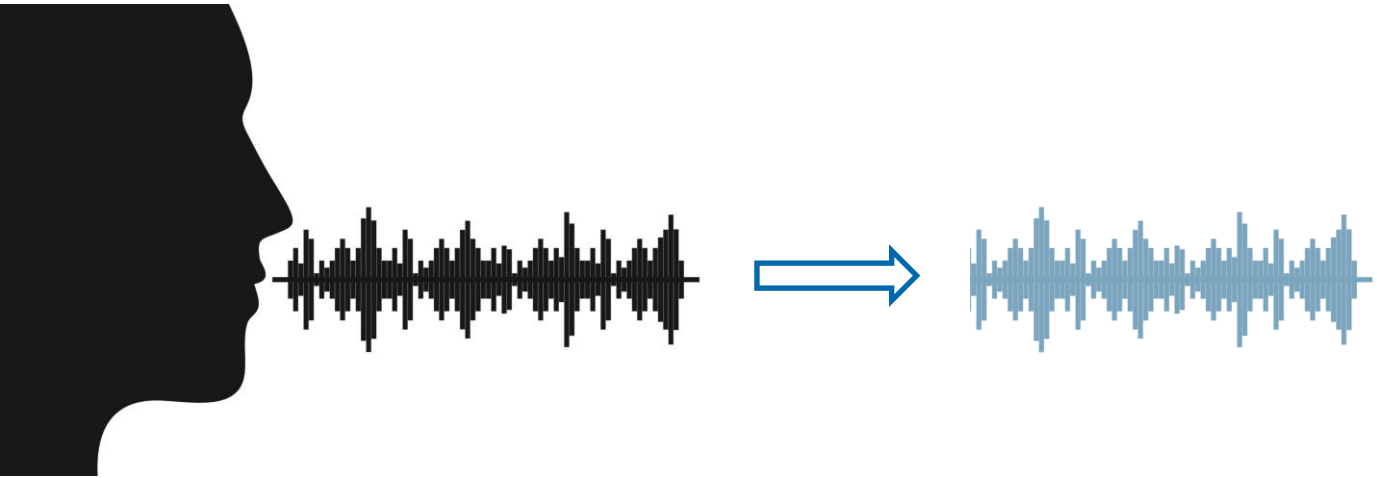
X-vector anonymization using autoencoders and adversarial training for preserving speech privacy

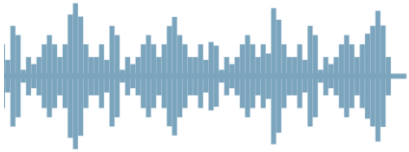
Presenter : Yasin Fakhar

Speaker Anonymization

Speaker anonymization is a method of protecting voice privacy by concealing individual speaker characteristics while preserving linguistic information.



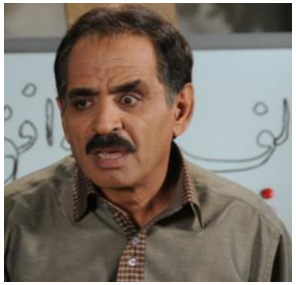
Speech latent features



Speaker identity

gender

accent

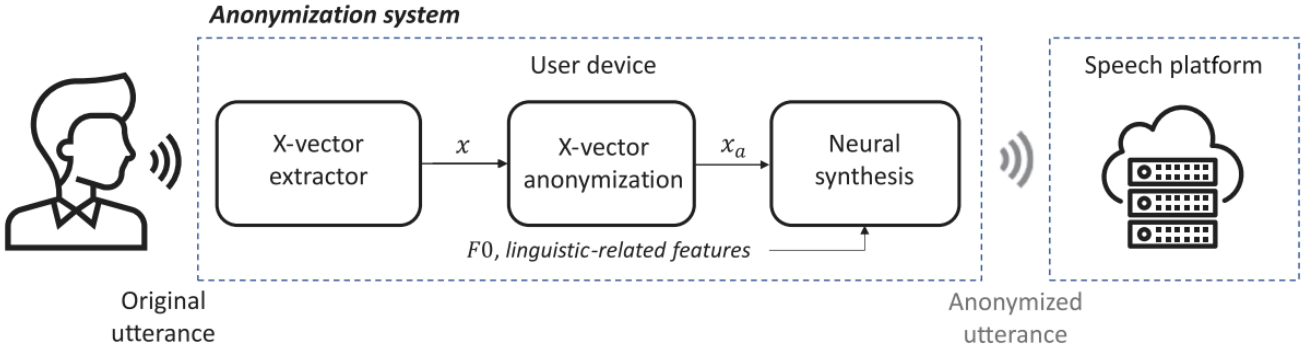


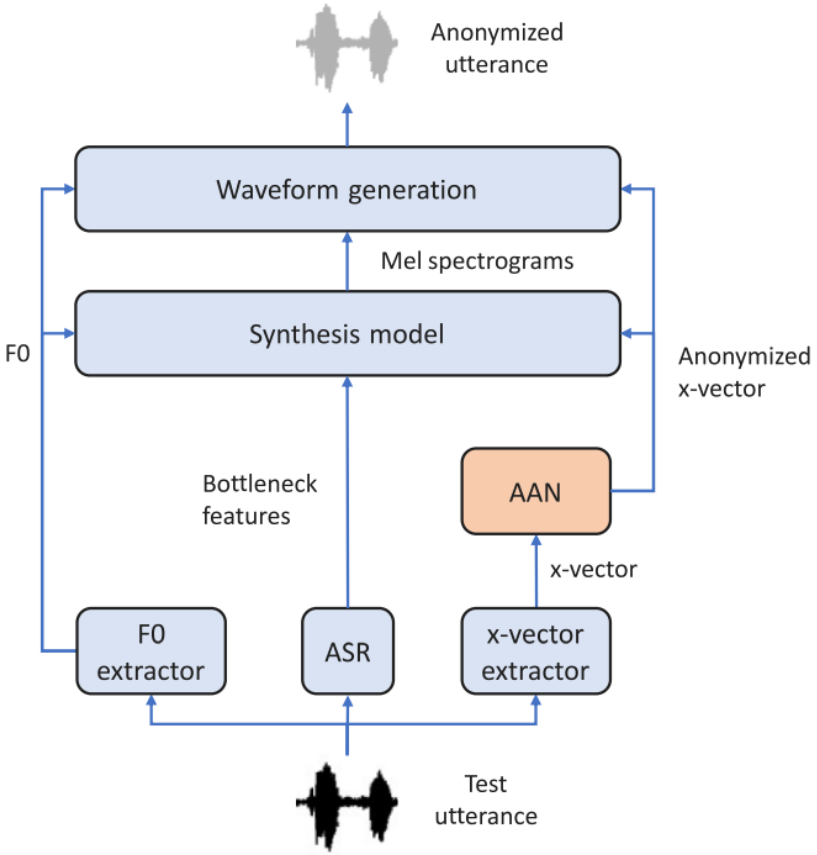
Voice Privacy Initiative

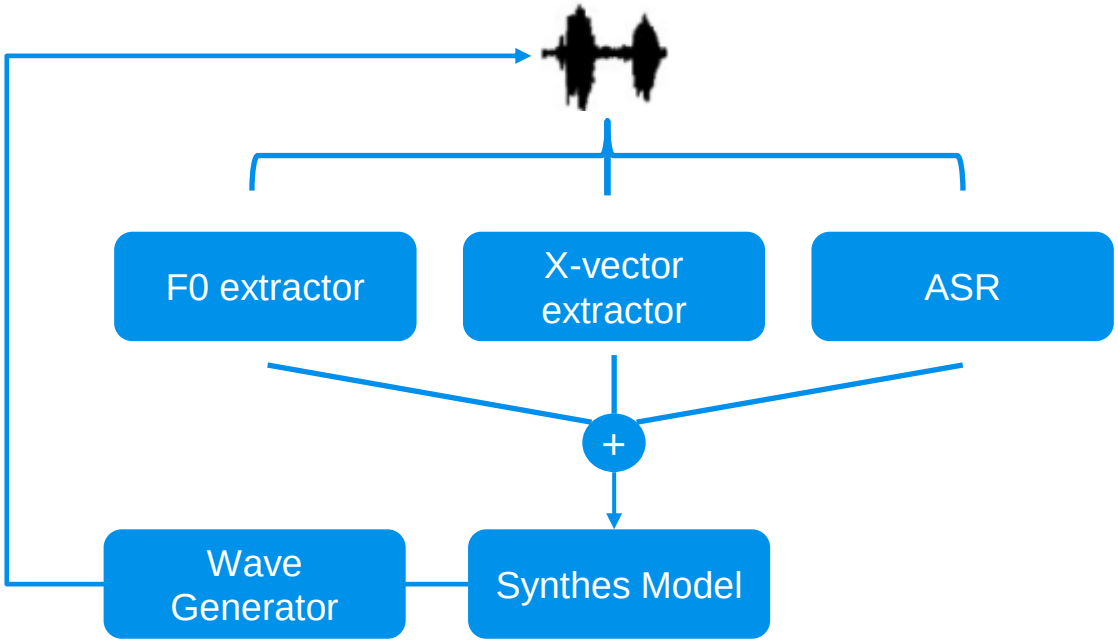
was created with the aim of defining certain protocols, metrics, and an evaluation of the systems used in common datasets provided by the organization.

One of the baseline systems proposed, as will explain further , is based on the extraction and modification of the x-vectors

Over View

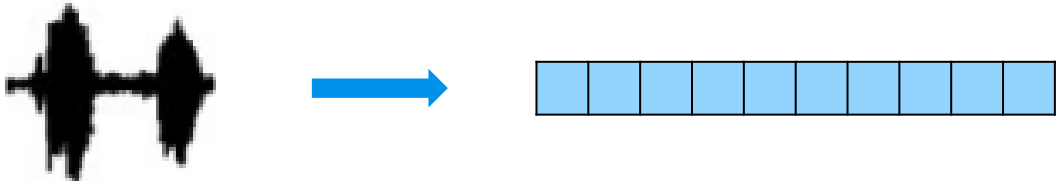

Anonymization system
User device
X-vector extractor
x
X-vector anonymization
x_a
Neural synthesis
$F0$, linguistic-related features
Speech platform
Original utterance
Anonymized utterance


Anonymized utterance
Waveform generation
Mel spectrograms
Synthesis model
F0
Anonymized x-vector
Bottleneck features
AAN
x-vector
F0 extractor
ASR
x-vector extractor
Test utterance



Module	Description	Output features	Training dataset
F0 extractor	F0 estimation based on extractor ensemble (Juvela et al., 2016)	–	–
ASR	Factorized TDNN (Peddinti et al., 2015; Povey et al., 2018)	Bottleneck features (dim 256)	LibriSpeech
x-vector extractor	TDNN model topology (Snyder et al., 2018b)	Speaker x-vectors (dim 512)	VoxCeleb-1,2
Synthesis model	Autoregressive network (Fang et al., 2019)	Mel-filterbanks (dim 80)	LibriTTS
Waveform generator	NSF waveform model (Wang and Yamagishi, 2019)	Speech waveform	LibriTTS

X-vector



A technique for speaker embedding extraction that has demonstrated a significant performance in :

- Speaker verification
- Language recognition
- Speaker diarization

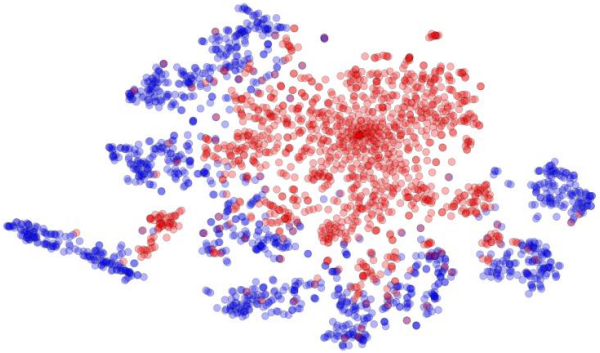
X-vector

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

The features are 24 dimensional filter banks with a frame-length of 25ms, mean-normalized over a sliding window of up to 3 seconds. The same energy Speech activity detection (SAD) as used in the baseline systems filters out nonspeech frames

Gradient Reversal

As can be seen from the image above, A classifier learned on the source distribution (blue) won't perform well on the classifier learned on the target distribution (red) [2]



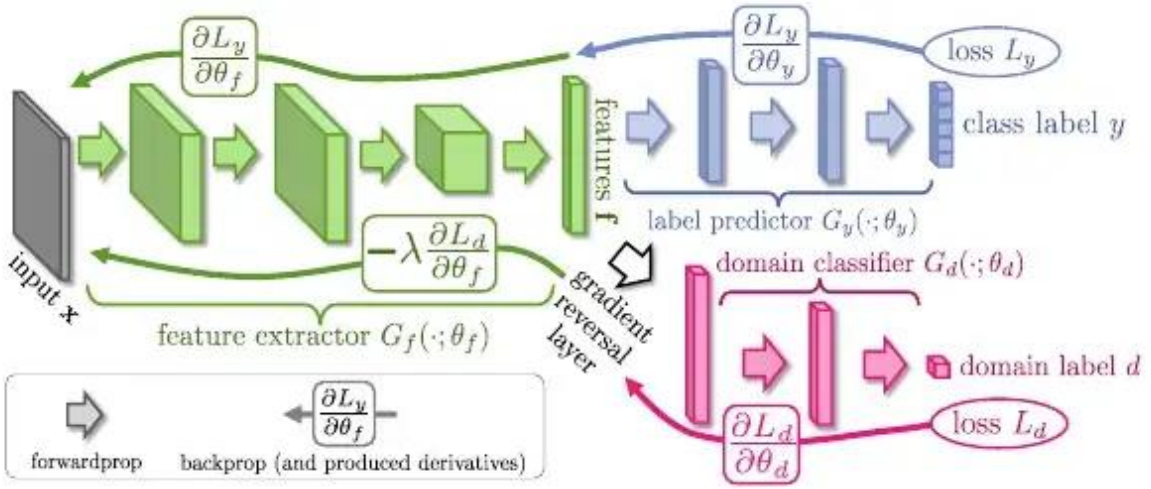
Solution : we need make both this distributions indistinguishable.

Gradient Reversal

Feature Extractor : transformation on the source and target distribution

Label Classifier : Since, source domain is labelled, will learn to perform classification on the transformed source distribution

Domain Classifier : This is a neural network that will be predicting whether the output of the Feature Extractor is from the source distribution is from source distribution or the target distribution.

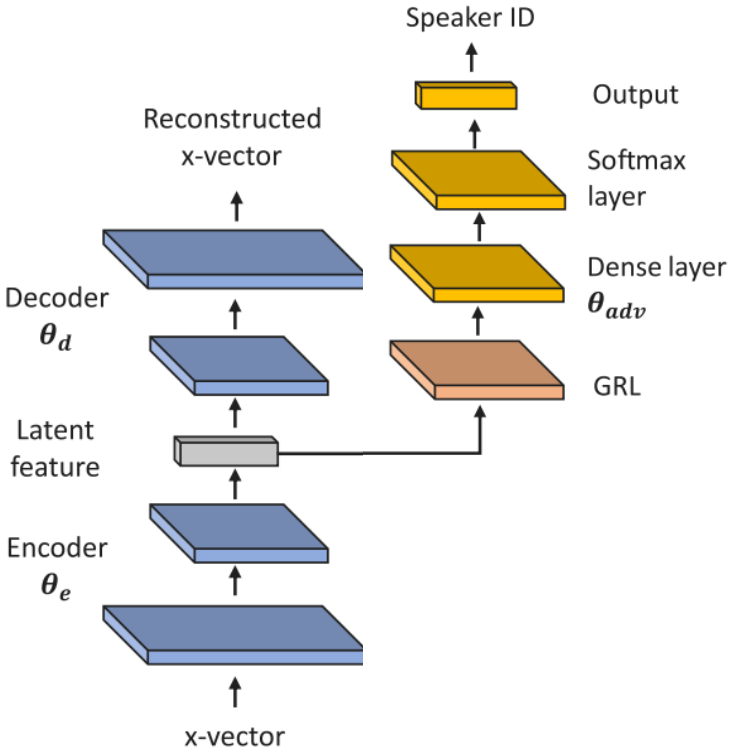


Proposed method

$$L_{au}(\theta_e, \theta_d) = \frac{1}{N} \sum_k (x_k - y_k; \theta_e, \theta_d)^2$$

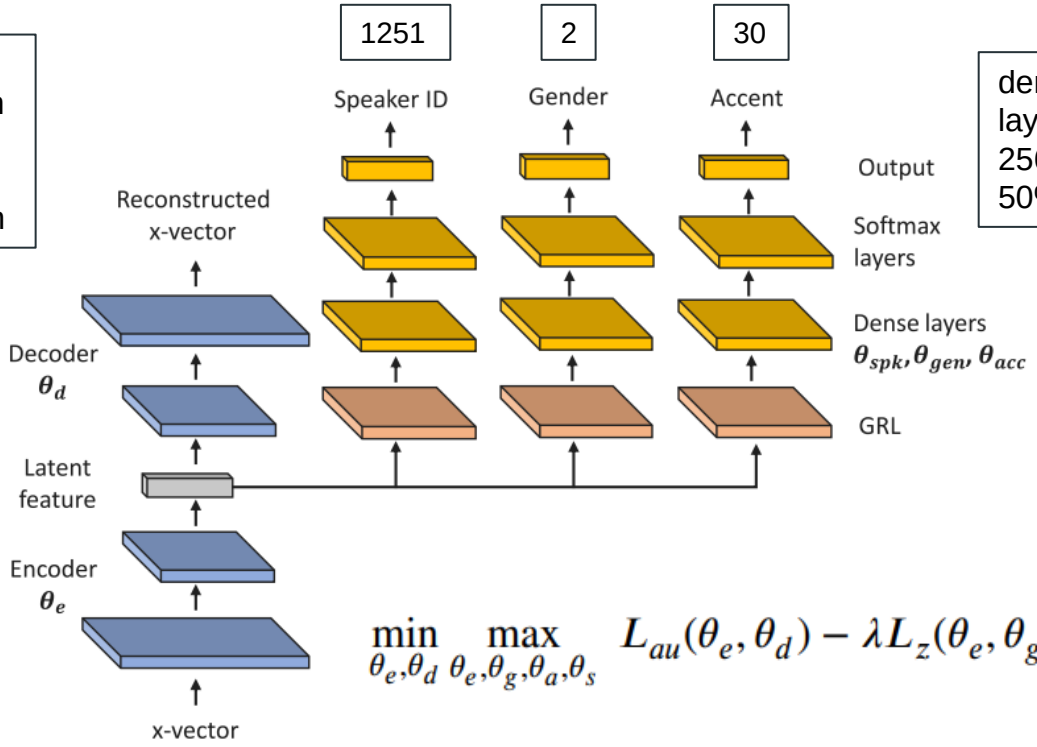
$$L_z(\theta_e, \theta_{adv}) = -\sum_k \log P(z_k | x_k; \theta_e, \theta_{adv})$$

$$\min_{\theta_e, \theta_d} \max_{\theta_e, \theta_{adv}} L_{au}(\theta_e, \theta_d) - \lambda L_z(\theta_e, \theta_{adv}),$$



Multiple Domain Adaptation

4 dense layer with 512 units and $tanh$ activation



dense-ReLU layer with 256 units, a 50% dropout

$$\min_{\theta_e,\theta_d} \max_{\theta_e,\theta_g,\theta_a,\theta_s} L_{au}(\theta_e,\theta_d) - \lambda L_z(\theta_e,\theta_g,\theta_a,\theta_s),$$

Dataset

Training

Vox-Celeb-1 dataset : 330 h of recordings from 1251 speakers, including labels for speaker identity, gender, and nationality

Evaluation (According to the VoicePrivacy Challenge)

using a subset of VCTK corpus

Common : utterance #1-24 identical for all speakers

Different : distinct utterances for all speakers

In librispeech : speakers in enrollment set is subset of trial

Statistics of the VCTK-dev, VCTK-test, and LibriSpeech test clean dataset. Adopted from VoicePrivacy Initiative (Tomashenko et al., 2020b).

Dataset	Subset	# Female spk	# Male spk	# Total spk	# Utt
VCTK-dev	Enrollment Trial (different)	15	15	30	600 10677
VCTK-test	Enrollment Trial (common) Trial (different)	15	15	30	600 70 10748
LibriSpeech test clean	Enrollment Trial	16 20	13 20	29 40	438 1496

Evaluation

- automatic speaker verification (ASV) system → EER, LLR metric
- automatic speech recognition (ASR) model → WER metric

ASV

VCTK-test different and LibriSpeech test-clean datasets were used to evaluate the anonymization system in terms of privacy (EER in Speaker Verification), whereas the whole VCTK-test (common+different) and LibriSpeech test-clean datasets were used to evaluate the intelligibility (WER in Speech Recognition)

Enrollment : known speaker used to train the ASV model (enrollment utterance)
Test : some speakers to check if utterances are correctly verified or not (trial utterances)

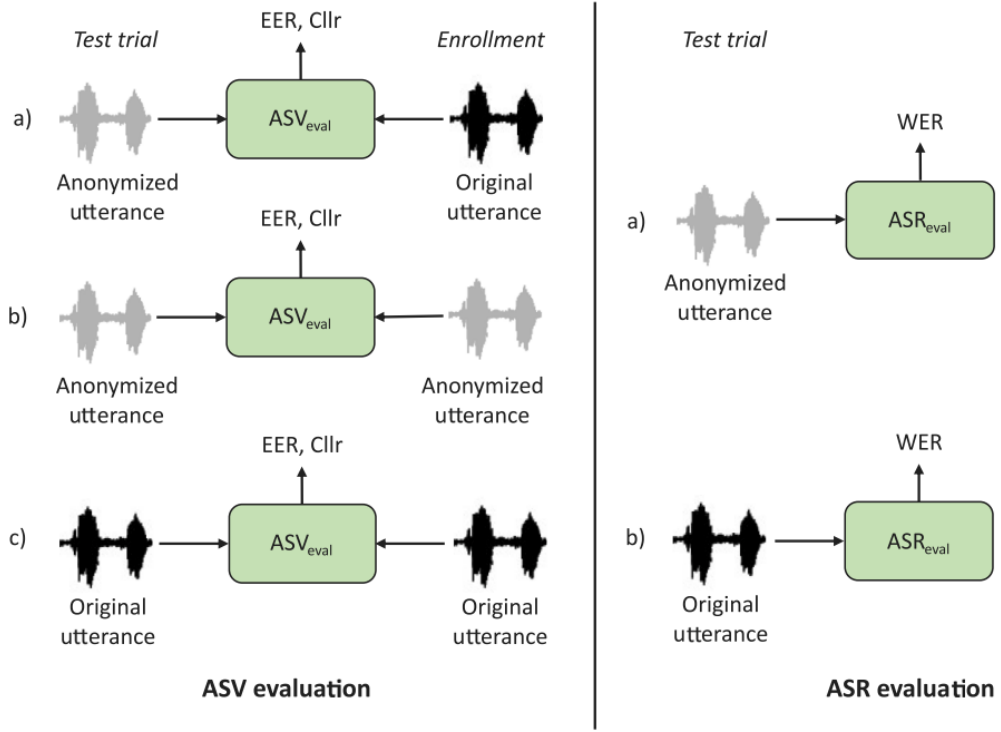
Attacker : try to know the speaker identity with a single trial utterance and several enrollment utterances

Attack Scenario :
Trial utterance : always anonymized
Enrollment utterance :

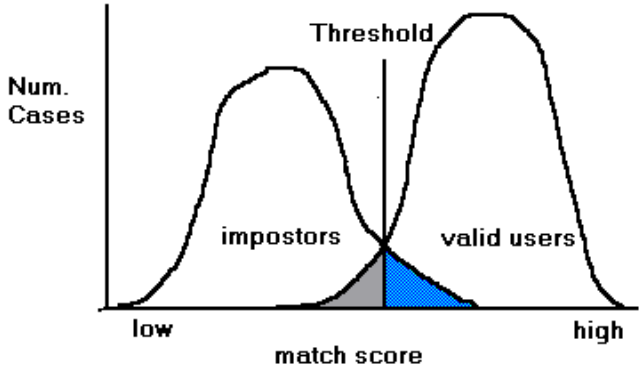
- Original
- Anonymized

Original enrollment – anonymized trial
Anonymized enrollment – anonymized trial
Original enrollment – original trial (for comparing performance of the model with SOTA models

Evaluation


Test trial
EER, Cllr
Enrollment
a)
ASV$_{eval}$
Anonymized utterance
Original utterance
b)
Anonymized utterance
Anonymized utterance
c)
Original utterance
Original utterance
ASV evaluation
Test trial
WER
ASR$_{eval}$
Anonymized utterance
Original utterance
ASR evaluation

EER



Blue = false accept
Grey = false reject
Set the threshold so that the false accept rate (FAR) equals the false reject rate (FRR)

LLR

It can be shown that each computed LLR can be expressed as the sum of two contributions. The first, *intrinsic information*, is available at the channel output before any decoding stage; the second, *extrinsic information*, is provided by exploiting the dependencies (due to convolution, parity, …) existing between the symbol being processed and the other symbols processed by the decoder [3]

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{tar}} \sum_{i \in targets} \log_2(1 + e^{-LLR_i}) + \frac{1}{N_{imp}} \sum_{j \in impostors} \log_2(1 + e^{LLR_j}) \right),$$

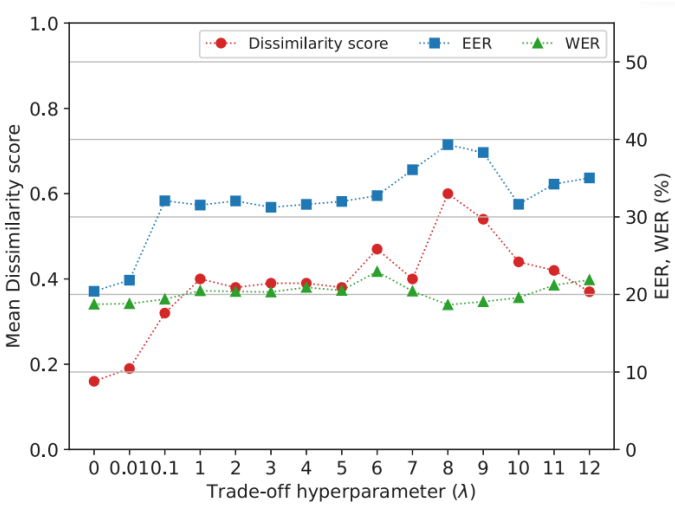
WER

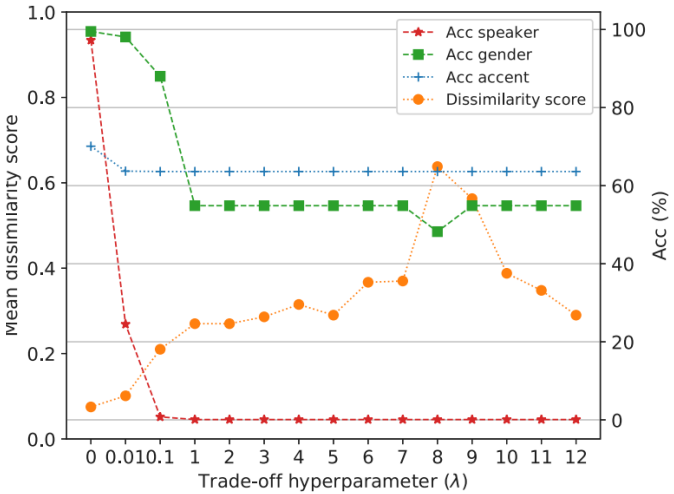
Put simply, WER is the ratio of errors in a transcript to the total words spoken. A lower WER in speech-to-text means better accuracy in recognizing speech. For example, a 20% WER means the transcript is 80% accurate.

Word Error Rate = (Substitutions + Insertions + Deletions) / Number of Words Spoken

- You read a paragraph out loud. It has X number of words.
- An ASR system hears you and outputs a string of text.
- There's some misspellings that are substituted — S.
- Sometimes the system inserts words that weren't said — I.
- And some words are deleted, as in not picked up at all — D.

Lambda trade-off





Results

Privacy evaluation results for baseline and AAN anonymization obtained through an evaluation of the pretrained ASV_{eval} model for the three attack scenarios: (1) original enrollment-anonymized trial (o-enroll, a-trial); (2) anonymized enrollment-anonymized trial (a-enroll, a-trial); and (3) original enrollment-original trial scenario, i.e., no anonymization, (o-enroll, o-trial).

Dataset	Attack scenario	Gender	Baseline			AAN		
			EER %	C_{llr}^{min}	C_{llr}	EER %	C_{llr}^{min}	C_{llr}
VCTK-test *diff*	o-enroll, o-trial	Female	4.88	0.169	1.495	–	–	–
	o-enroll, a-trial		48.05	0.998	146.929	48.35	0.997	155.964
	a-enroll, a-trial		31.74	0.847	11.527	34	0.881	21.306
	o-enroll, o-trial	Male	2.06	0.072	1.817	–	–	–
	o-enroll, a-trial		53.85	1	167.824	48.16	0.996	157.427
	a-enroll, a-trial		30.94	0.834	23.842	39.04	0.947	33.550
LibriSpeech test clean	o-enroll, o-trial	Female	7.664	0.183	26.793	–	–	–
	o-enroll, a-trial		47.260	0.995	151.822	43.980	0.972	168.557
	a-enroll, a-trial		32.120	0.839	16.270	34.850	0.886	27.144
	o-enroll, o-trial	Male	1.114	0.041	15.303	–	–	–
	o-enroll, a-trial		52.120	0.999	166.658	45.430	0.980	155.451
	a-enroll, a-trial		36.750	0.903	33.928	46.100	0.979	47.663

Results

Performance of the x-vector anonymization systems, in terms of mean EER (over all VoicePrivacy development and test datasets) and WER, submitted to the VoicePrivacy 2020 Challenge. Those results are extracted from the VoicePrivacy Challenge setup and results presented during the VoicePrivacy 2020 Virtual Workshop at Odyssey 2020 (Tomashenko et al., 2020c). The mean EER values are shown for the two attack scenarios: original enrollment-anonymized trial (o-enroll, a-trial) and anonymized enrollment-anonymized trial (a-enroll, a-trial)

System	Mean EER %		WER %
	o-enroll, a-trial	a-enroll, a-trial	
Baseline	51.6	31.9	10.9
Our approach	47.5	37.8	11.0
Mawalin (Mawalim et al., 2020)	53.1	27.4	11.0
Turner (Turner et al., 2020)	46	38.3	11.2
Champion (Champion et al., 2020)	52.7	32.1	13.7

references

[1]	Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5329-5333). IEEE.
[2]	Ganin, Y., & Lempitsky, V. (2015, June). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180-1189). PMLR.
[3]	Boutillon, E., Masera, G., Declercq, D., Fossorier, M., & Biglieri, E. (2014). Hardware design and realization for iteratively decodable codes. In *Channel coding: Theory, algorithms, and applications* (pp. 583-642). Elsevier.
[4]	Perero-Codosero, J. M., Espinoza-Cuadros, F. M., & Hernández-Gómez, L. A. (2022). X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. *Computer Speech & Language*, *74*, 101351.

Thanks for your attention!