Enhancing Speaker Anonymization Using Disentanglement Learning

Nikita Kuzmin

Phd Student NTU/A*STAR year 3

Supervisors:
Prof. Chng Eng Siong
Dr. Sailor Hardik Bhupendra

Jan. 2025

About me

1. Name: Nikita Kuzmin
2. Status:
   a. Matriculated on 08-Aug-2022
   b. 1, 2, 3 TAC appraisal passed
   c. All school requirement fulfilled for QE (GAP hours, TA courses)
3. CGPA: 4.67

4. Publications:
   a. **N. Kuzmin,** Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S. (2024) NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 72-79, doi: 10.21437/SPSC.2024-13

   b. **N. Kuzmin*,** A. Sholokhov*, K. A. Lee and E. S. Chng, ”Probabilistic Back-ends for Online Speaker Recognition and Clustering,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10097032.

Outline

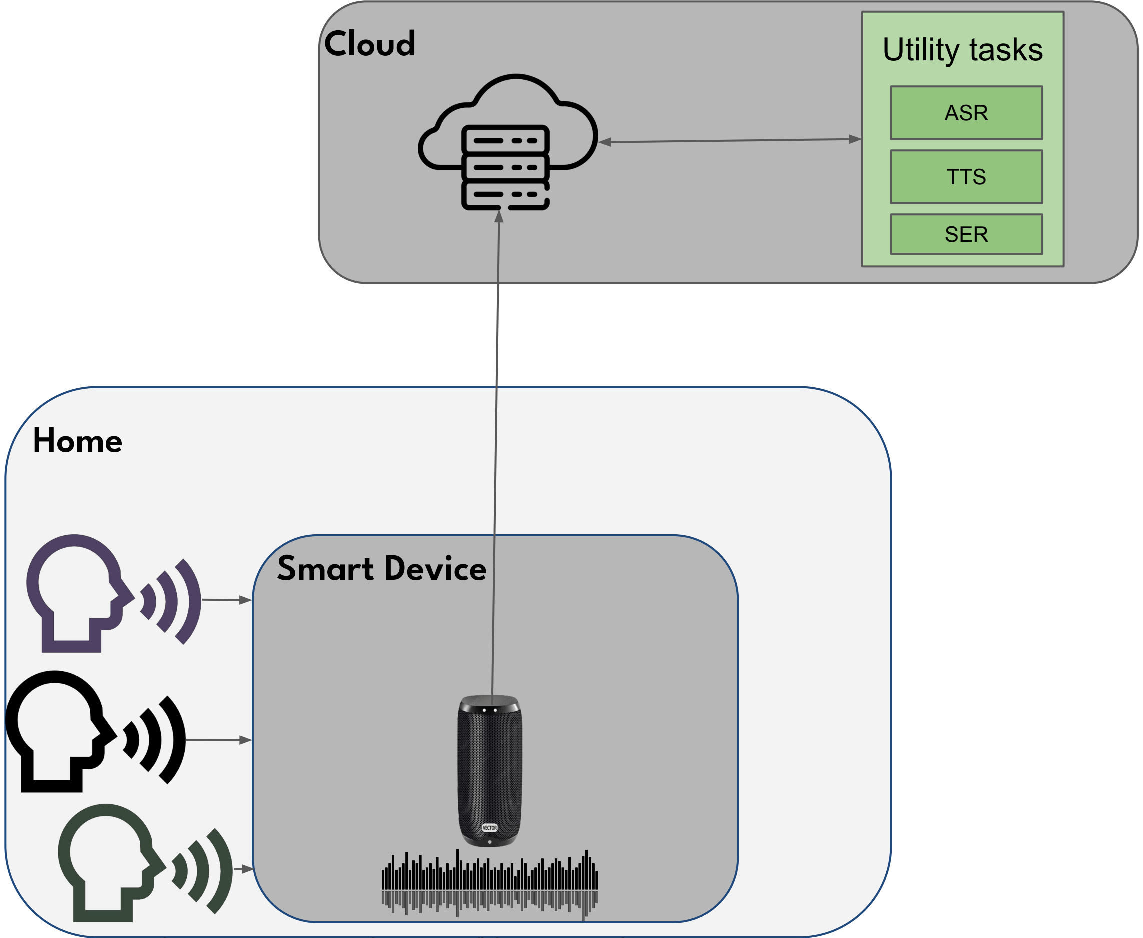
1. **Introduction**

2. Literature Review
    a. Speaker Anonymization
    b. Disentanglement Learning

3. Disentanglement-based Approaches for Anonymization
    a. Problem 1: Anonymization models do not preserve emotions well
        i. Contribution 1.1: NS3 FACodec
        ii. Contribution 1.2: Emotion embeddings
    b. Problem 2: Identity Leakage in B5
        i. Contribution 2: Mean-reversion + Noise

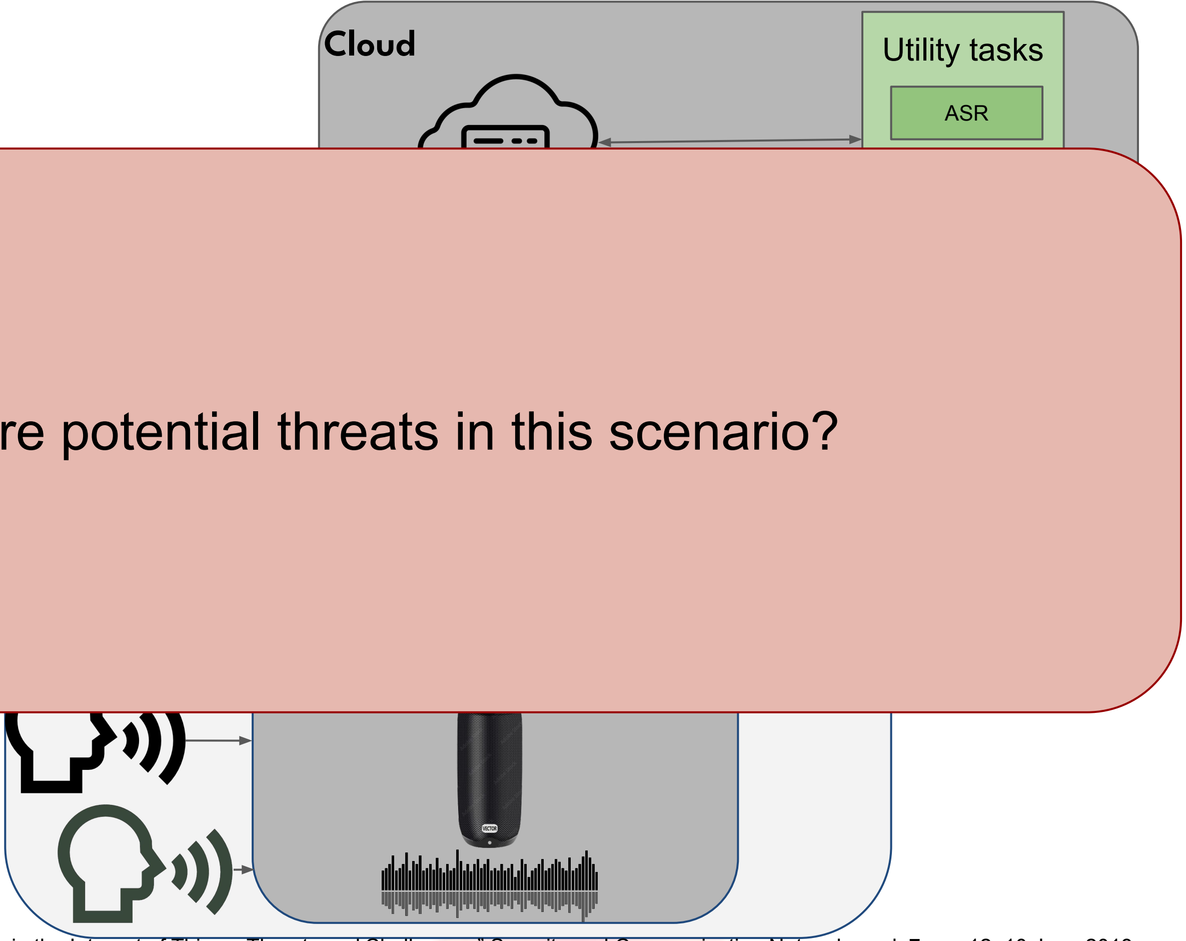
4. Conclusions and Future Work

1.1. Intro to Speaker Anonymization



Ziegeldorf, Jan Henrik, et al. “Privacy in the Internet of Things: Threats and Challenges.” Security and Communication Networks, vol. 7, no. 12, 10 June 2013, pp. 2728–2742, https://doi.org/10.1002/sec.795.

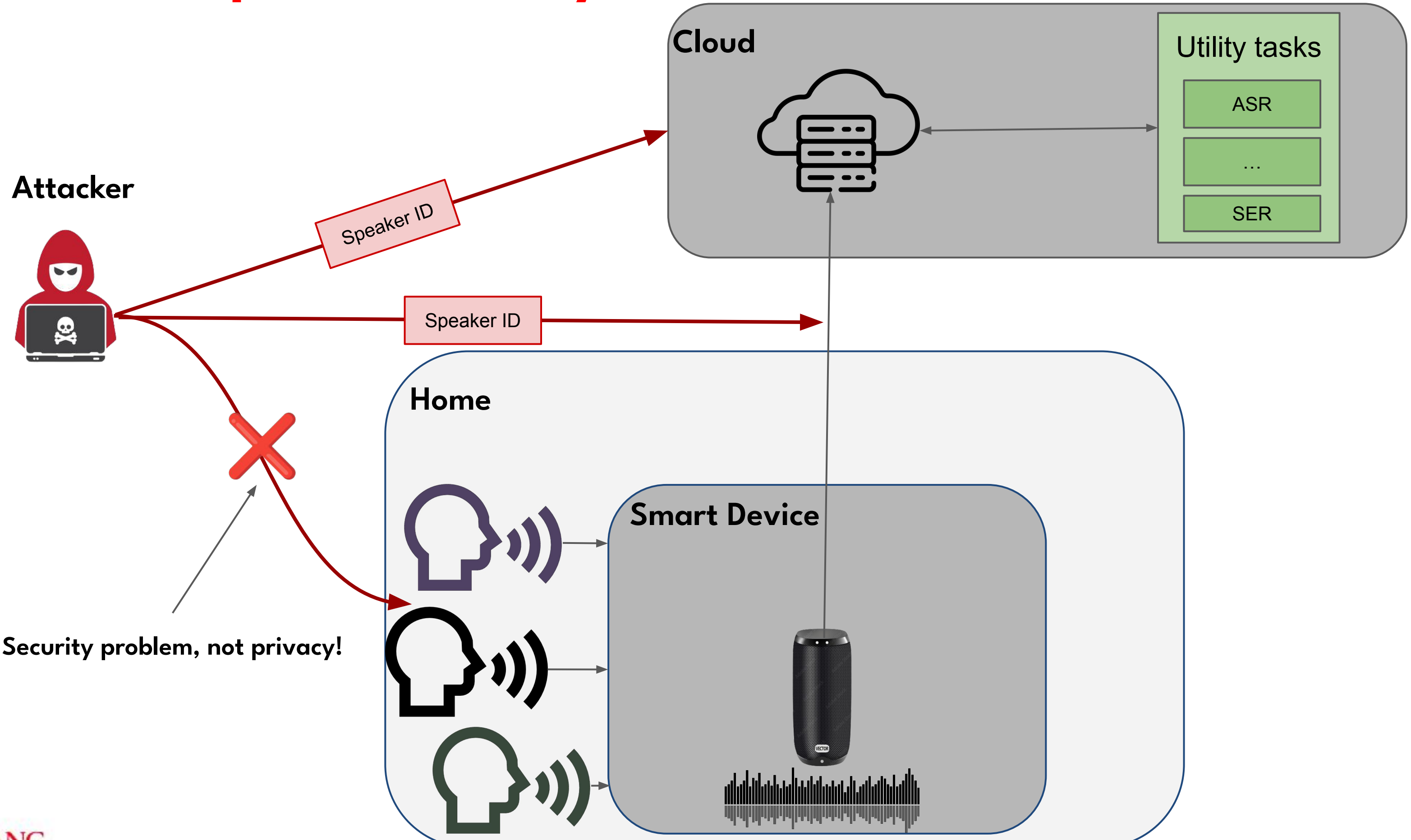
1.1. Intro to Speaker Anonymization



What are potential threats in this scenario?

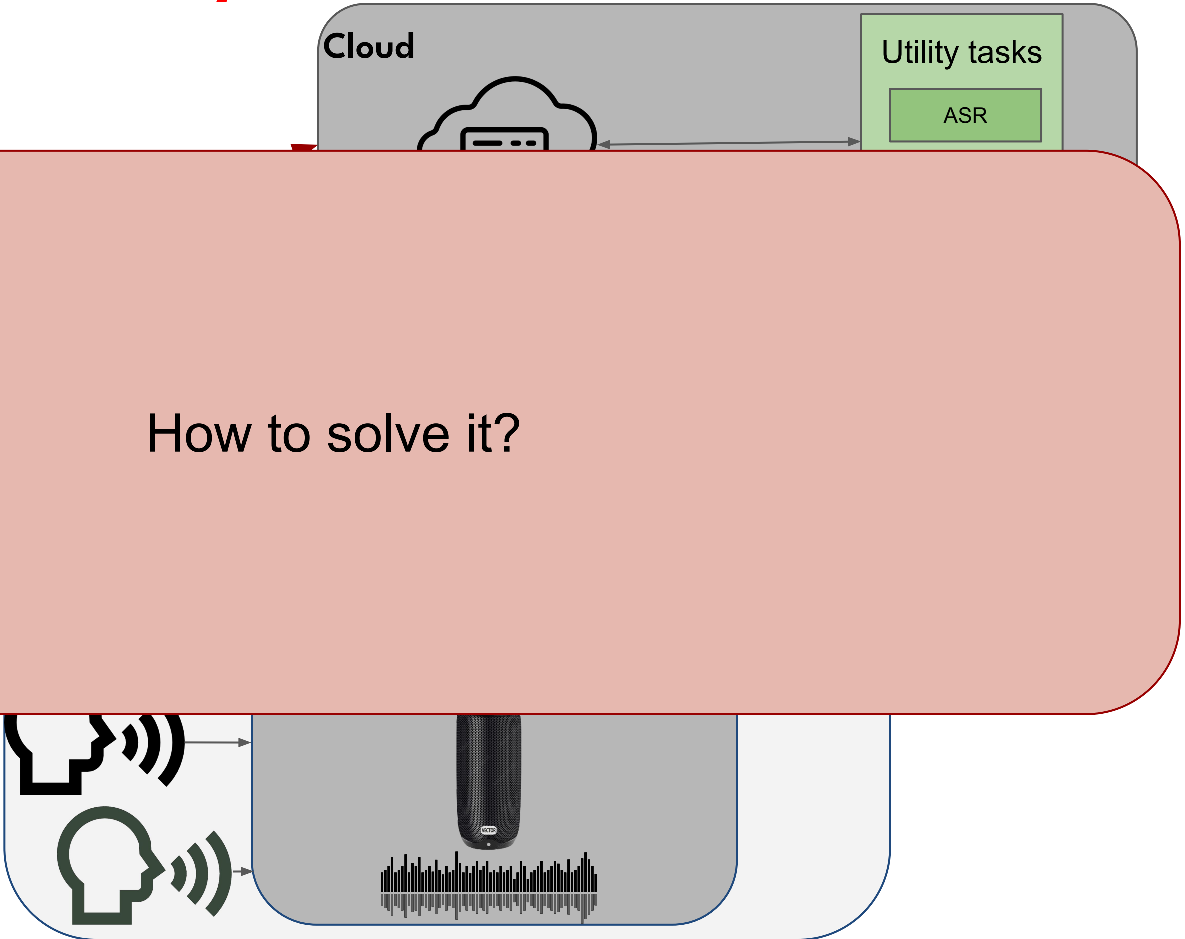
Ziegeldorf, Jan Henrik, et al. “Privacy in the Internet of Things: Threats and Challenges.” Security and Communication Networks, vol. 7, no. 12, 10 June 2013, pp. 2728–2742, https://doi.org/10.1002/sec.795.

1.1. Intro to Speaker Anonymization



Ziegeldorf, Jan Henrik, et al. “Privacy in the Internet of Things: Threats and Challenges.” Security and Communication Networks, vol. 7, no. 12, 10 June 2013, pp. 2728–2742, https://doi.org/10.1002/sec.795.

1.1. Intro to Speaker Anonymization

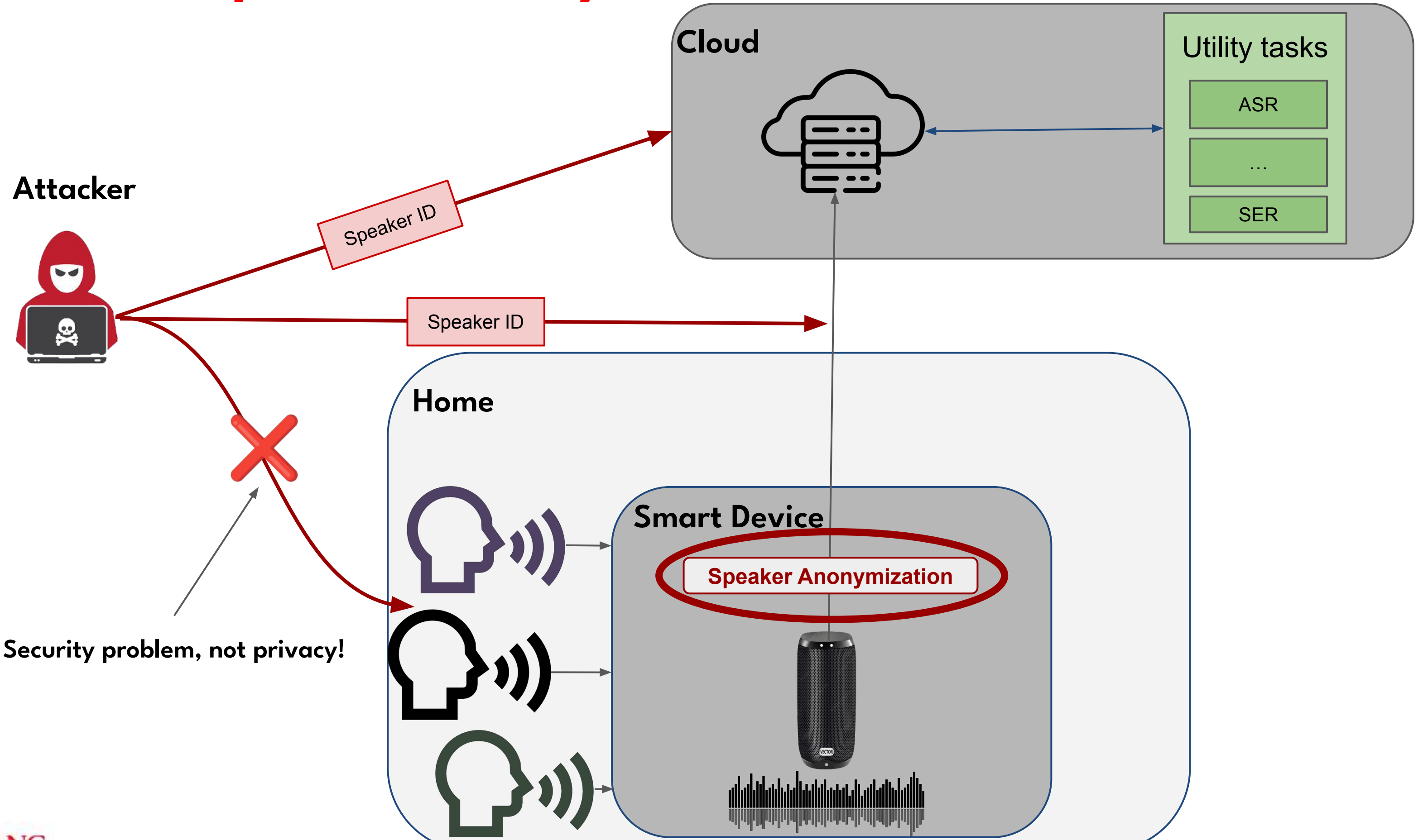


How to solve it?

Security problem, not privacy!

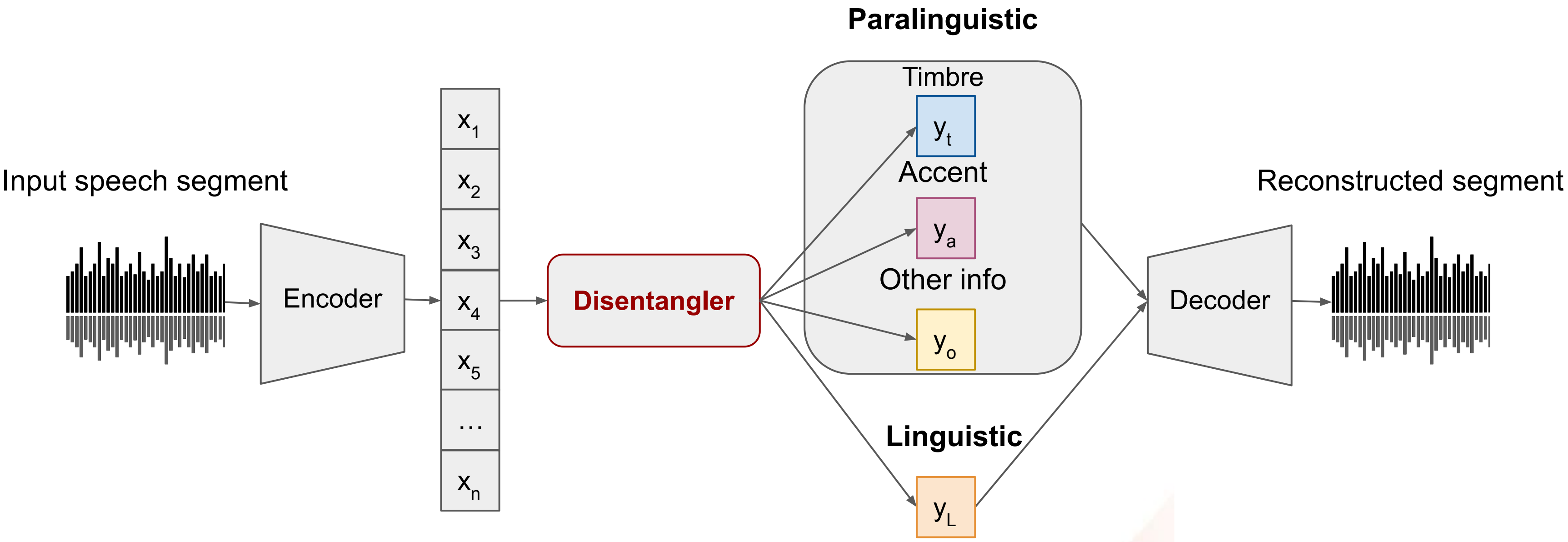
Ziegeldorf, Jan Henrik, et al. “Privacy in the Internet of Things: Threats and Challenges.” Security and Communication Networks, vol. 7, no. 12, 10 June 2013, pp. 2728–2742, https://doi.org/10.1002/sec.795.

1.1. Intro to Speaker Anonymization



Ziegeldorf, Jan Henrik, et al. “Privacy in the Internet of Things: Threats and Challenges.” Security and Communication Networks, vol. 7, no. 12, 10 June 2013, pp. 2728–2742, https://doi.org/10.1002/sec.795.

1.2. Intro to Disentanglement Learning



J. Williams, “Learning disentangled speech representations,” Ph.D. dissertation, University of Edinburgh, 2022.

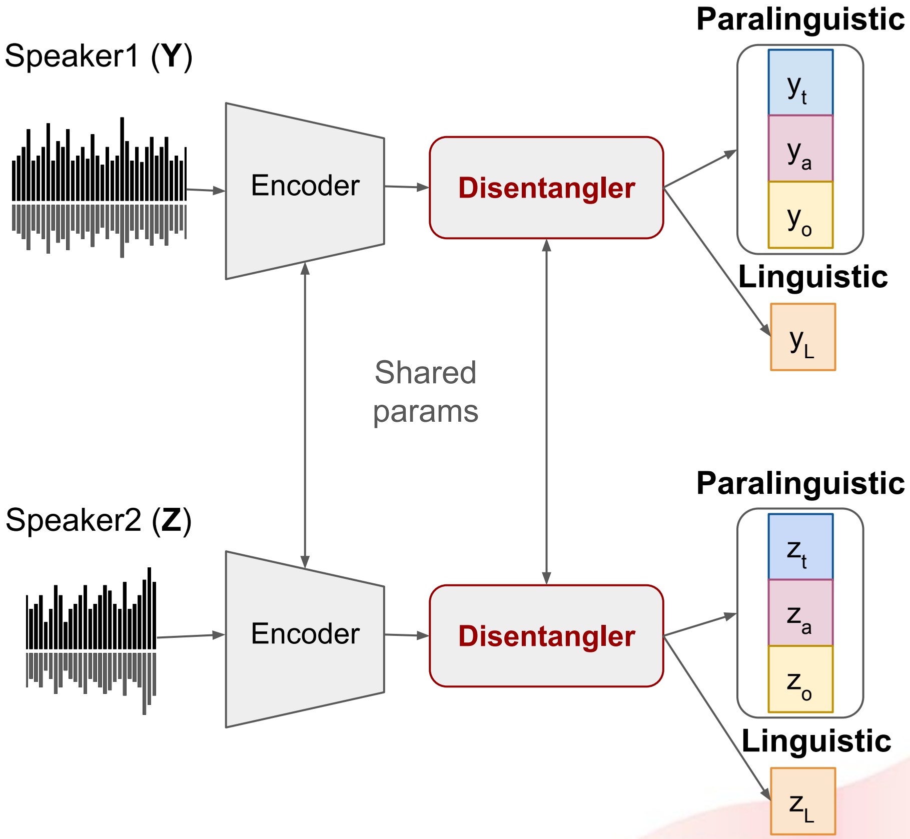
1.2. Intro to Disentanglement Learning

Why do we need Disentanglement?

ent

J. Williams, “Learning disentangled speech representations,” Ph.D. dissertation, University of Edinburgh, 2022.

1.2. Intro to Disentanglement Learning



Speaker Verification

Emotion Recognition

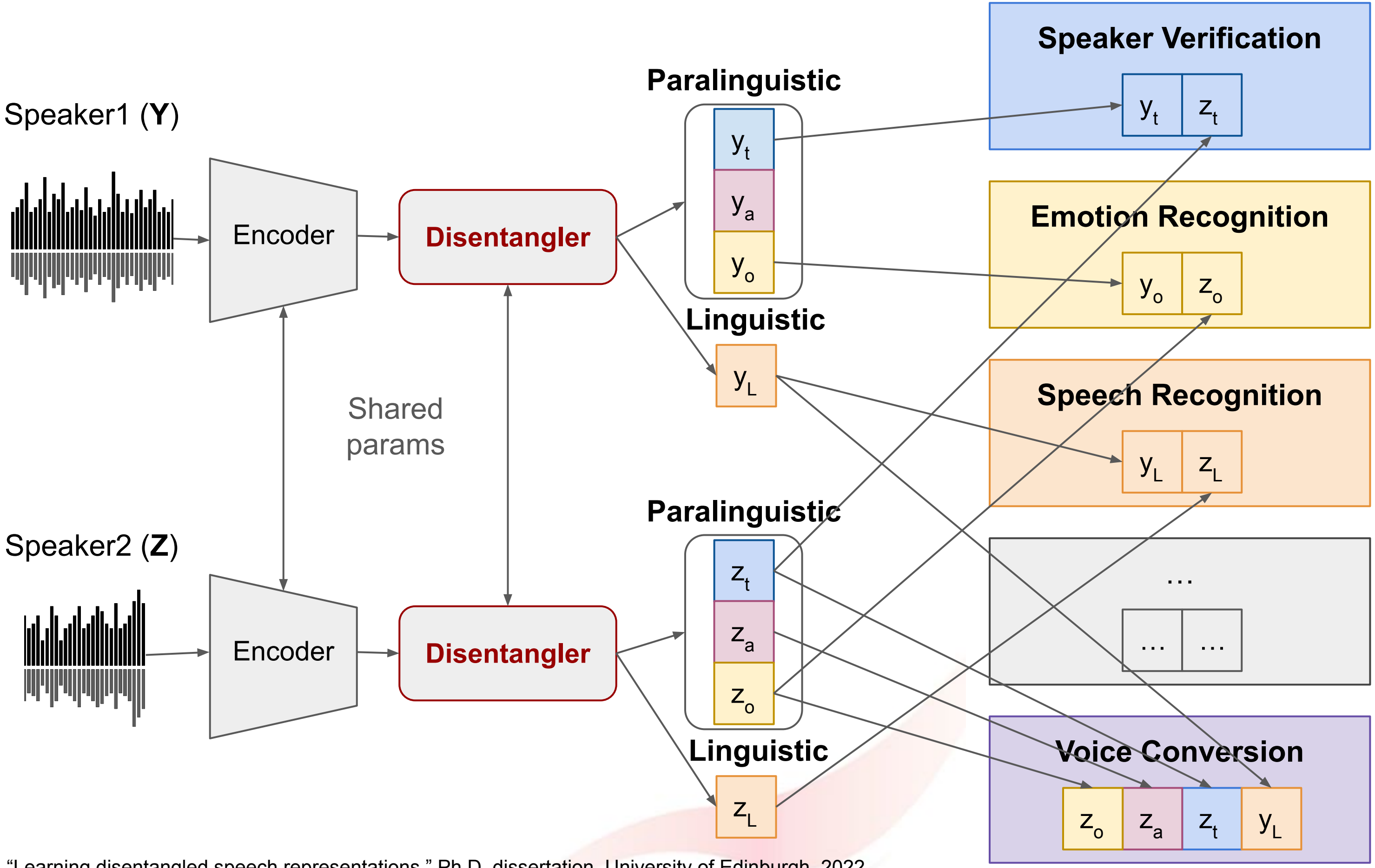
Speech Recognition

…

Voice Conversion

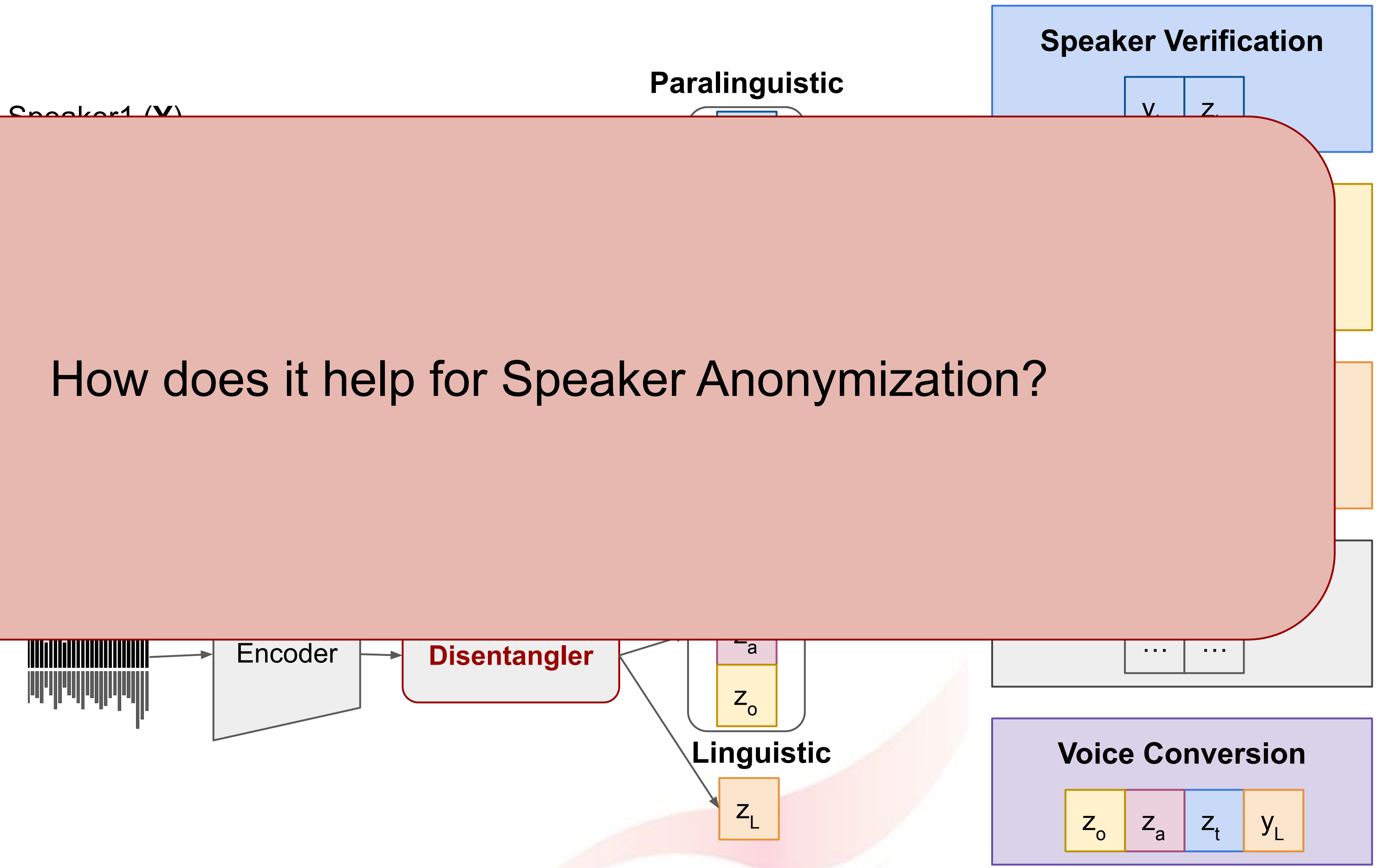
J. Williams, “Learning disentangled speech representations,” Ph.D. dissertation, University of Edinburgh, 2022.

1.2. Intro to Disentanglement Learning



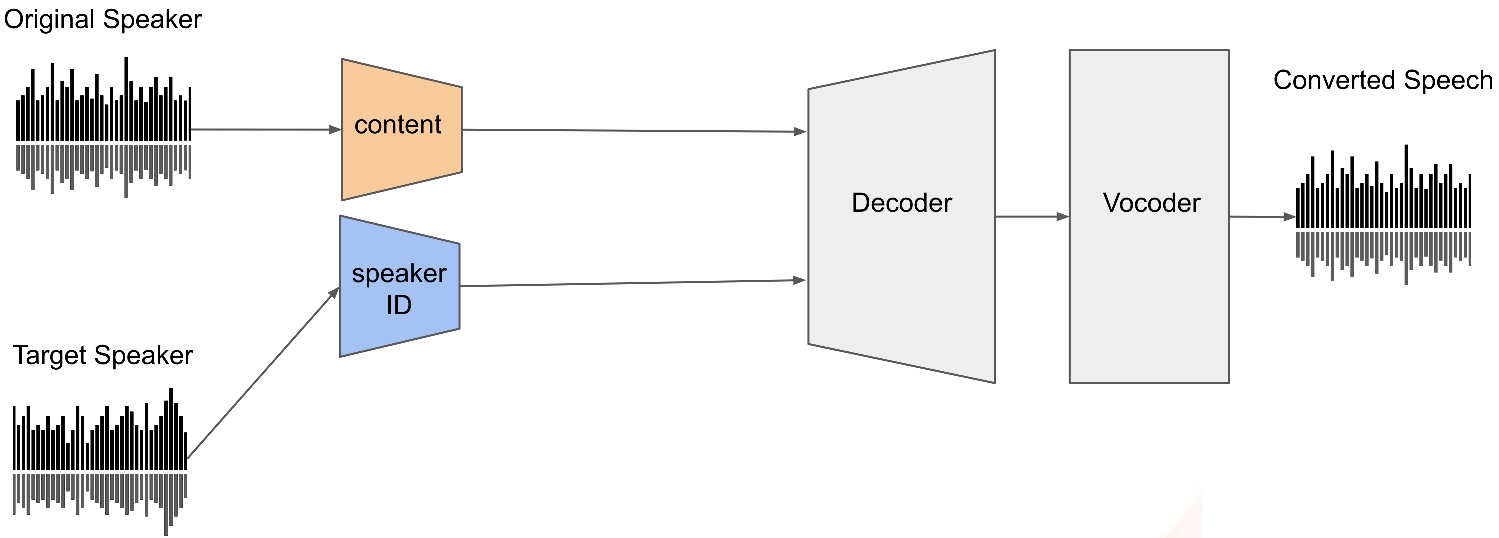
J. Williams, "Learning disentangled speech representations," Ph.D. dissertation, University of Edinburgh, 2022.

1.2. Intro to Disentanglement Learning



How does it help for Speaker Anonymization?

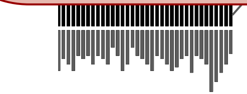
1.3. Voice Conversion



H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, “Disentangled speech representation learning for one-shot cross-lingual voice conversion using β-vae,” Jan. 2023, pp. 814–821

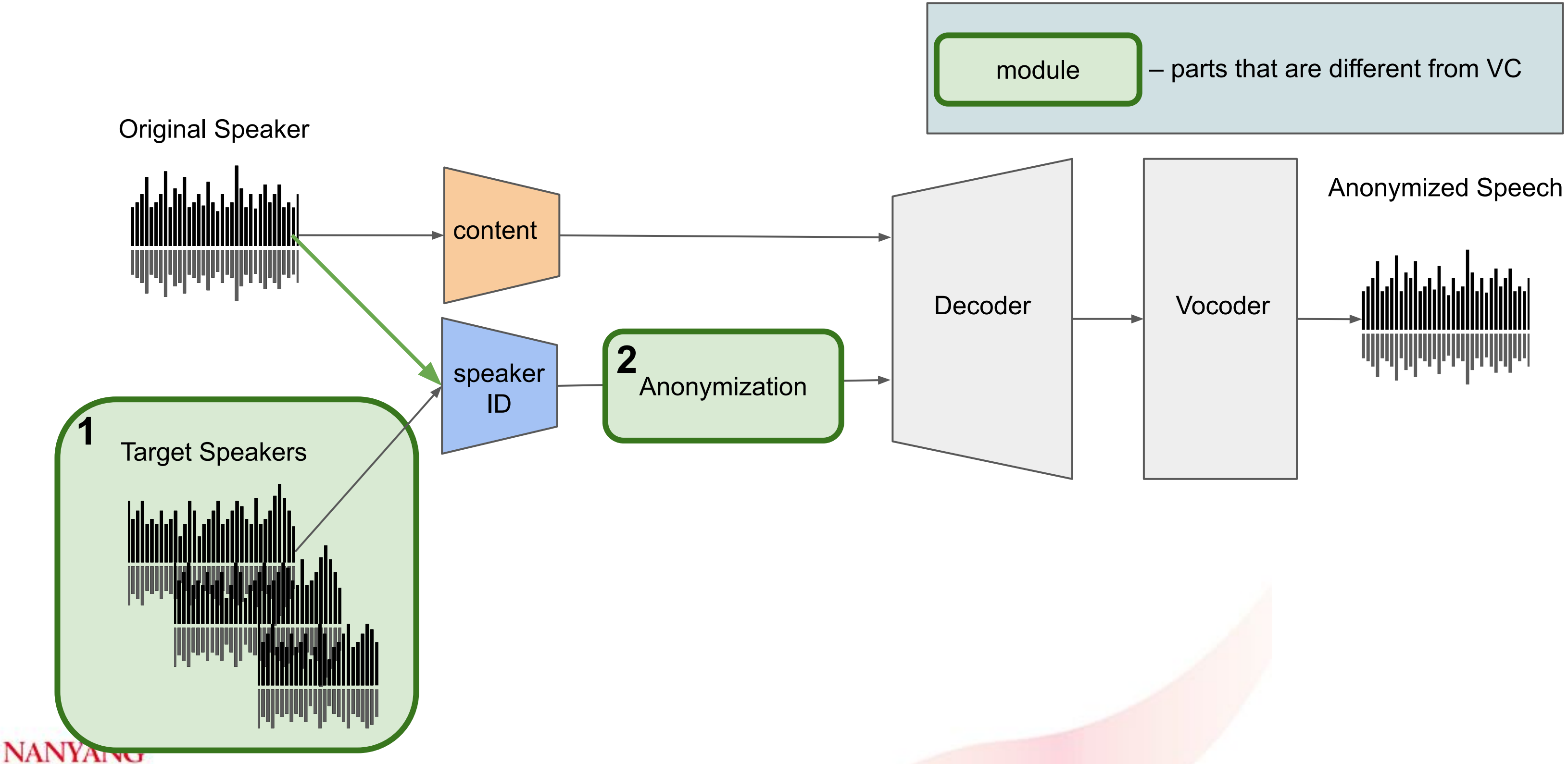
1.3. Voice Conversion vs Speaker Anonymization

What are the similarities between VC and Speaker Anonymization?



H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, “Disentangled speech representation learning for one-shot cross-lingual voice conversion using β-vae,” Jan. 2023, pp. 814–821

1.3. Speaker Anonymization Pipeline



S. Meyer, P. Tilli, F. Lux, P. Denisov, J. Koch, and N. T. Vu, “Cascade of phonetic speech recognition, speaker embeddings gan and multispeaker speech synthesis for the VoicePrivacy 2022 Challenge,” in Proc. 2nd Symposium on Security and Privacy in Speech Communication, 2022.

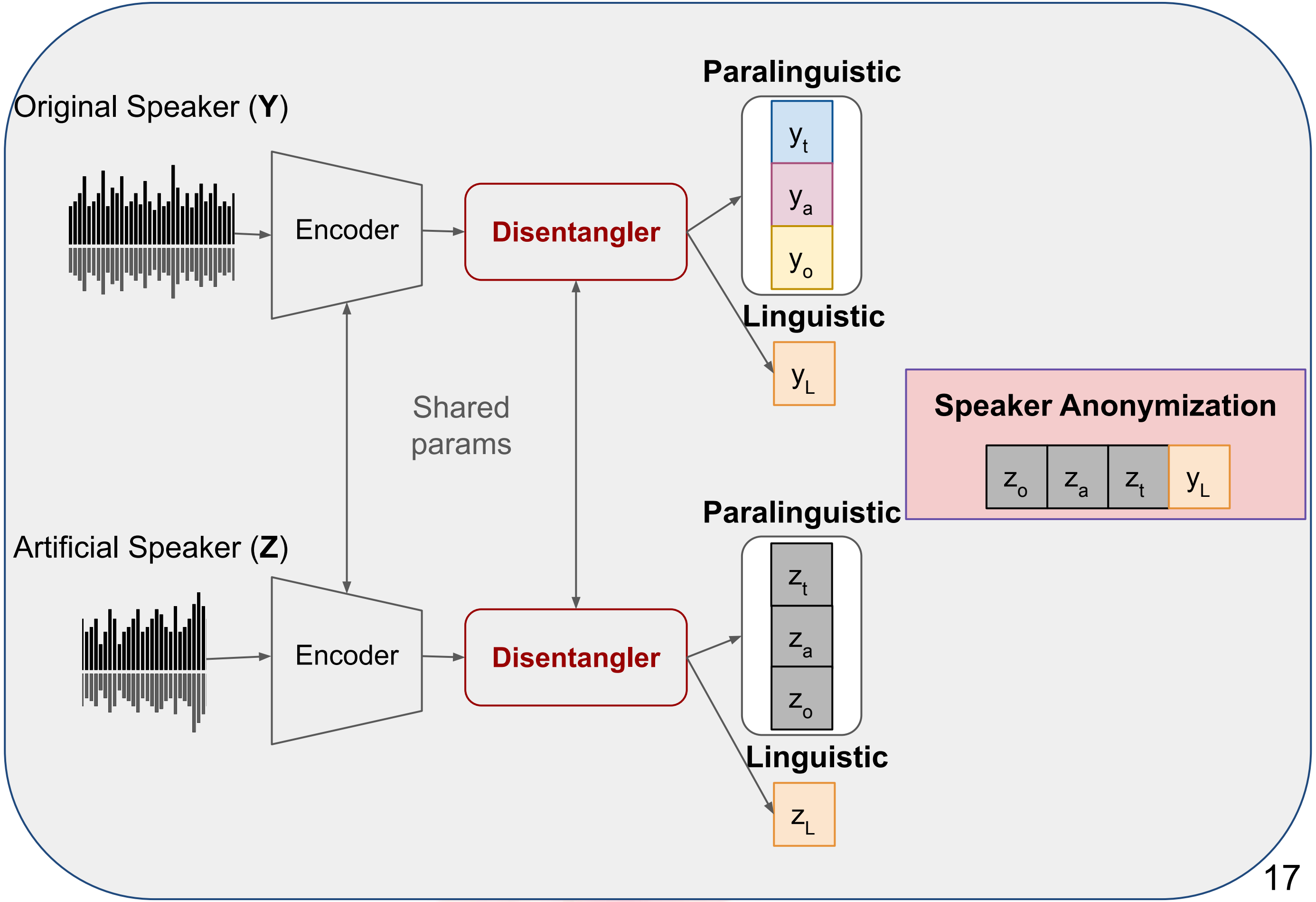
Connection between Disentanglement and Anonymization

Feature Separation:

Isolates linguistic content from speaker-specific (sensitive) attributes.

Targeted Privacy:

Enables removal of identity information while retaining speech quality. We can choose which components we want to conceal.



M. Baas and H. Kamper, “Disentanglement in a gan for unconditional speech synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing

Outline

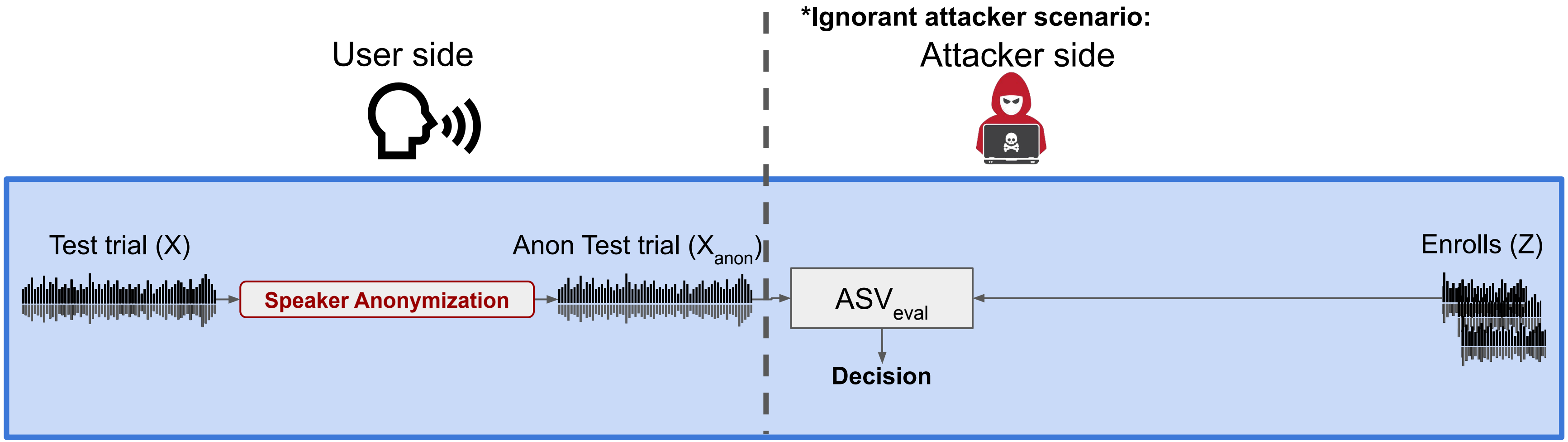
1. Introduction

2. **Literature Review**
    a. **Speaker Anonymization**
    b. **Disentanglement Learning**

3. Disentanglement-based Approaches for Anonymization
    a. Problem 1: Anonymization models do not preserve emotions
        i. Contribution 1.1: NS3 FACodec
        ii. Contribution 1.2: Emotion embeddings
    b. Problem 2: Identity Leakage in B5
        i. Contribution 2: Mean-reversion + Noise

4. Conclusions and Future Work

Speaker Anonymization: Privacy protection



N. Tomashenko et al., “Introducing the voiceprivacy initiative,” in Interspeech 2020.

Speaker Anonymization: Privacy protection

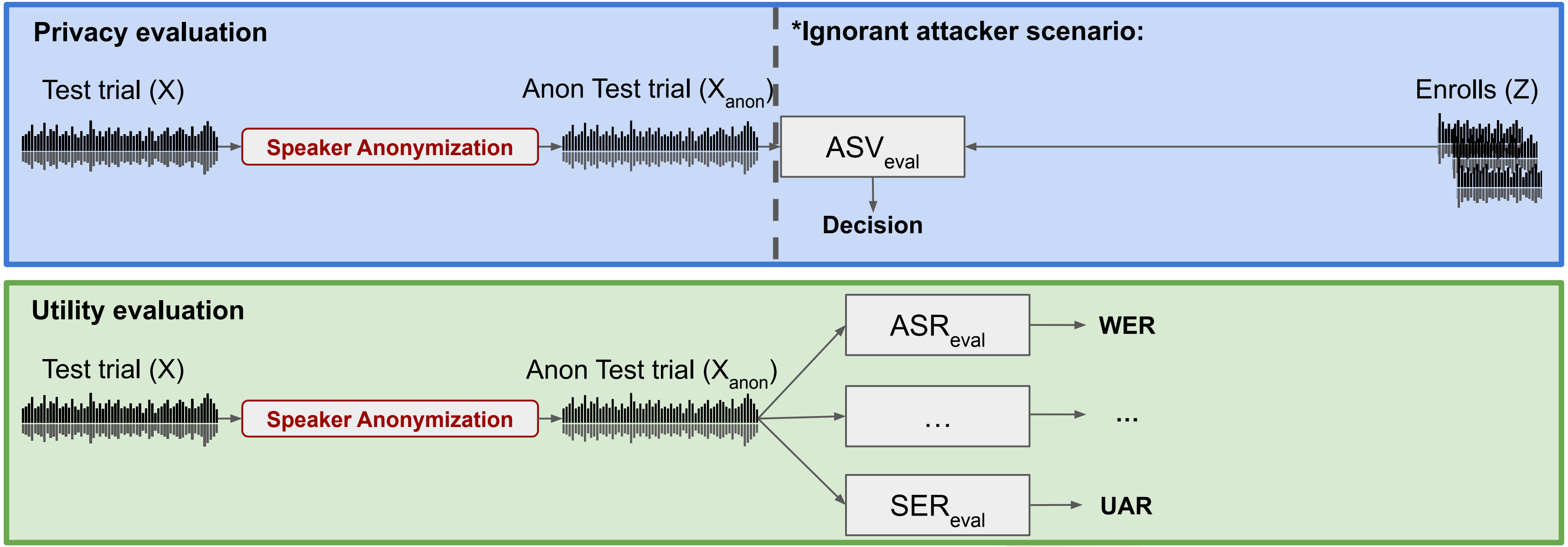
Best defence: transform each utterance to noise?

N. Tomashenko et al., “Introducing the voiceprivacy initiative,” in Interspeech 2020.

Speaker Anonymization: Privacy vs Utility

🔥 – finetuned/adapted on train data

❄ – no finetuning/adaptation is allowed

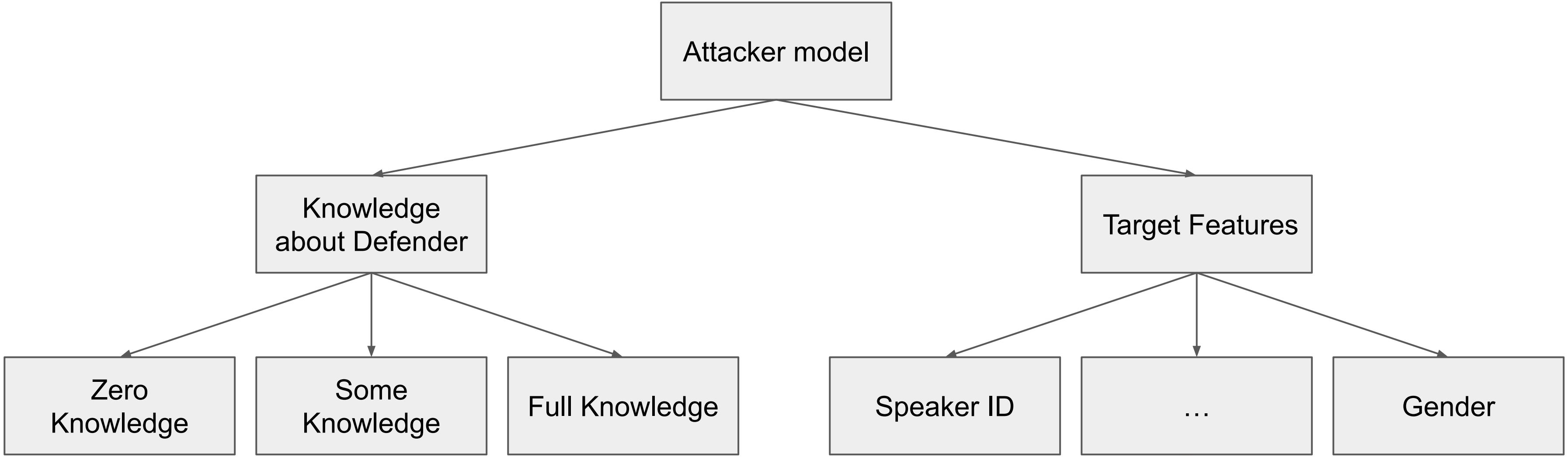


N. Tomashenko et al., "Introducing the voiceprivacy initiative," in Interspeech 2020.

Privacy vs Utility tradeoff: Qualitative Examples

Original Speech	Anonymized Speech		
	Poor content preservation	Poor emotion preservation	Proposed
🔊	🔊	🔊	🔊

Literature Review. Types of Attacker models



Srivastava, Brij Mohan Lal, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, « Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers », in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020

Literature Review. Types of Attacker models

So many different setups, how to compare methods?

Knowledge

Knowledge

der

Srivastava, Brij Mohan Lal, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, « Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers », in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020

Literature Review. Voice Privacy Challenge

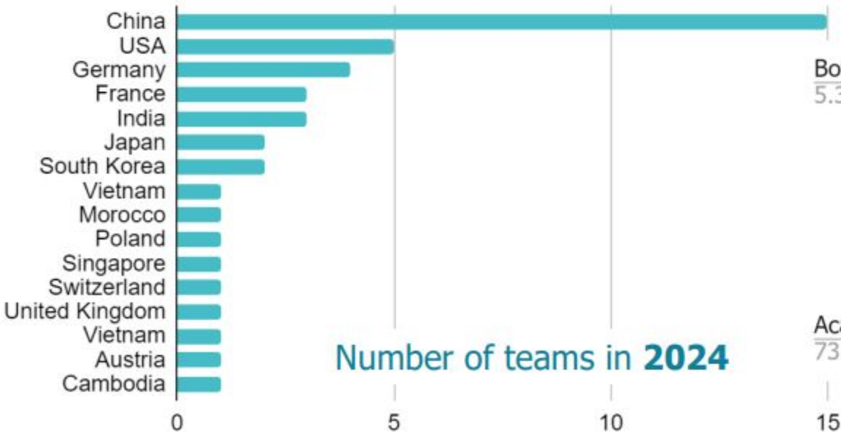
Why was the VoicePrivacy Initiative started?

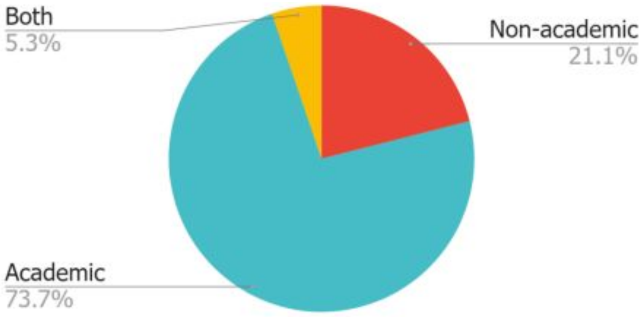
- Existing privacy **methods were incomparable**, and **there was no standard way to evaluate** anonymization solutions.
- Started the **Voice Privacy Challenge (VPC)** series to provide researchers the platform to compete in building the best anonymization system.

N. Tomashenko et al., “Introducing the voiceprivacy initiative,” in Interspeech 2020.
N. Tomashenko et al., “The voiceprivacy 2020 challenge evaluation plan,”
N. Tomashenko et al., The voiceprivacy 2022 challenge evaluation plan, 2022.
N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.

VPC evolution: Participants

	VPC2020	VPC2022	VPC2024
Statistics	Registered teams: 25 Participants: 45 Countries: 13 Submitted systems: 16	Registered teams: 43 Participants: 79 Countries: 17 Submitted systems: 16	Registered teams: 40 Participants: 107 Countries: 16 Submitted systems: 36





Srivastava, Brij Mohan Lal, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, « Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers », in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020
N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.

VPC evolution: Summary

+ – new elements

− – removed elements

= – unchanged elements

	VPC2020	VPC2022	VPC2024
Anon Level:	+ Speaker-level	= Speaker-level	+ Utterance-level − Speaker-level
Baseline systems:	+ 2 baseline systems: x-vector and neural waveform (2 systems in total)	+ 1 baseline system: x-vector with HiFi-GAN (3 systems in total)	+ 3 more baseline systems: NAC and ASR-BN with VQ (6 systems in total)
Attackers:	+ Lazy-informed	+ Semi-informed = Lazy-informed	= Semi-informed − Lazy-informed
Datasets:	+ Provided common datasets: VCTK, LibriSpeech, LibriTTS, VoxCeleb	= Same datasets as in VPC2020	+ Extended datasets and pretrained models
Metrics:	+ Metrics: EER and WER, subjective metrics	+ New metrics: GVD and pitch correlation = Metrics: EER and WER, subjective metrics	+ New utility metric: UAR for SER − GVD, pitch correlation and subjective metrics

NANYANG TECHNOLOGICAL UNIVERSITY

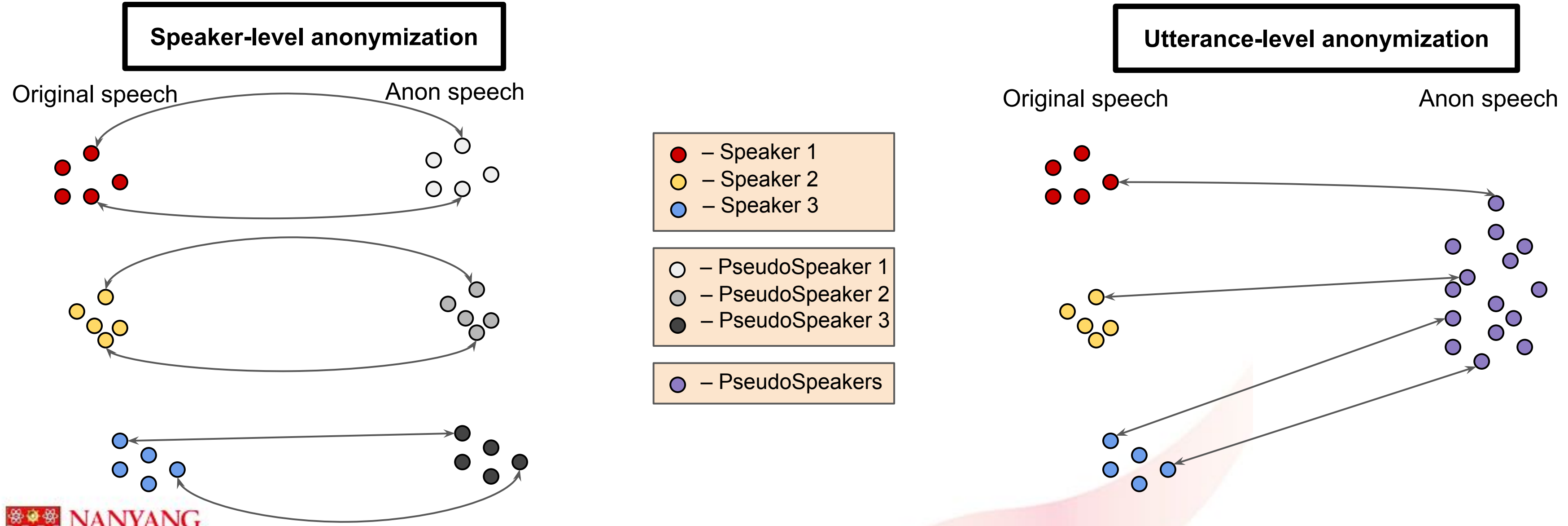
VPC evolution: Anonymization Levels

➕ – new elements

🟥 – removed elements

🟦 – unchanged elements

Anon Level:	➕ Speaker-level	🟦 Speaker-level	➕ Utterance-level 🟥 Speaker-level
	VPC2020	**VPC2022**	**VPC2024**

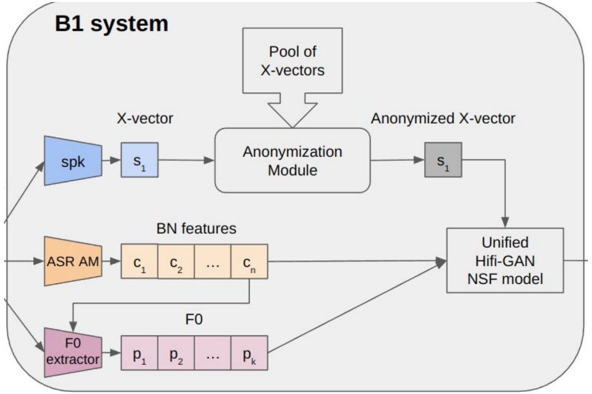


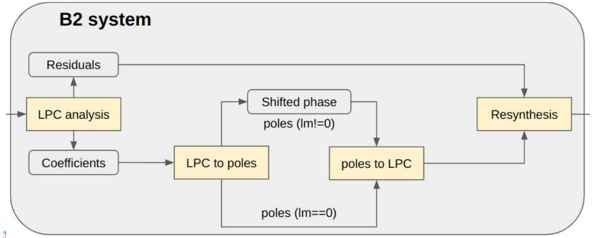
N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.

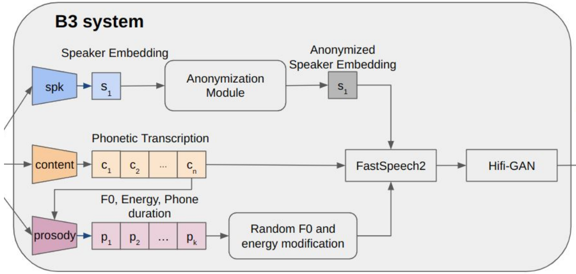
VPC evolution: Baseline Systems

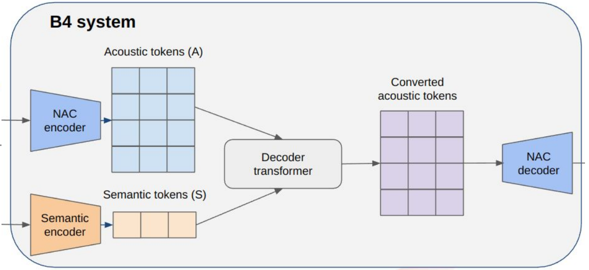
Baseline systems:

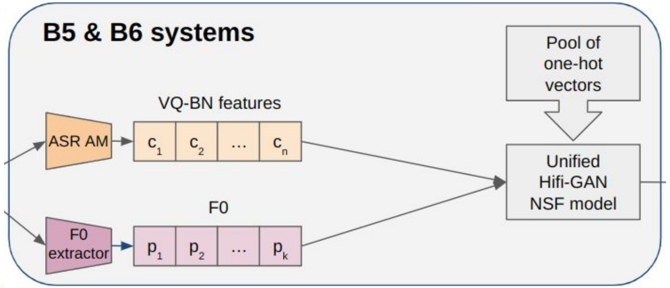
VPC2020	VPC2022	VPC2024
2 baseline systems: x-vector and neural waveform (2 systems in total)	1 baseline system: x-vector with HiFi-GAN (3 systems in total)	3 more baseline systems: NAC and ASR-BN with VQ (6 systems in total)











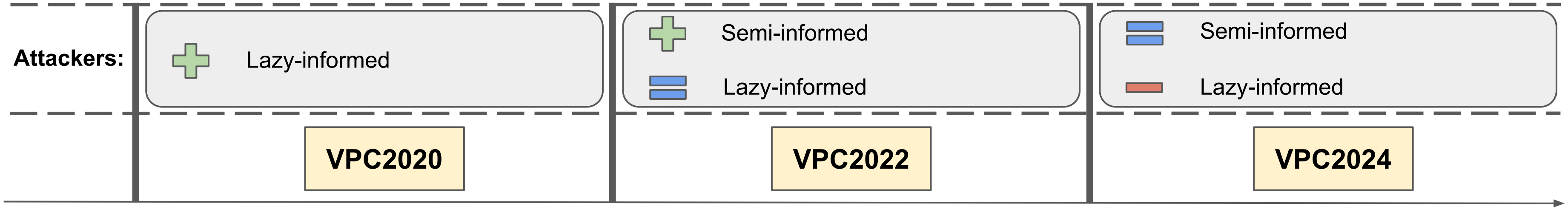
F. Fang et al., "Speaker anonymization using x-vector and neural waveform models," in Speech Synthesis Workshop 2019
M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024
J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," in Proc. Interspeech 2021, 2021
S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in ICASSP 2023
P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," PhD dissertation, Université de Lorraine, Jun 2023

VPC evolution: Attacker Types

✚ – new elements

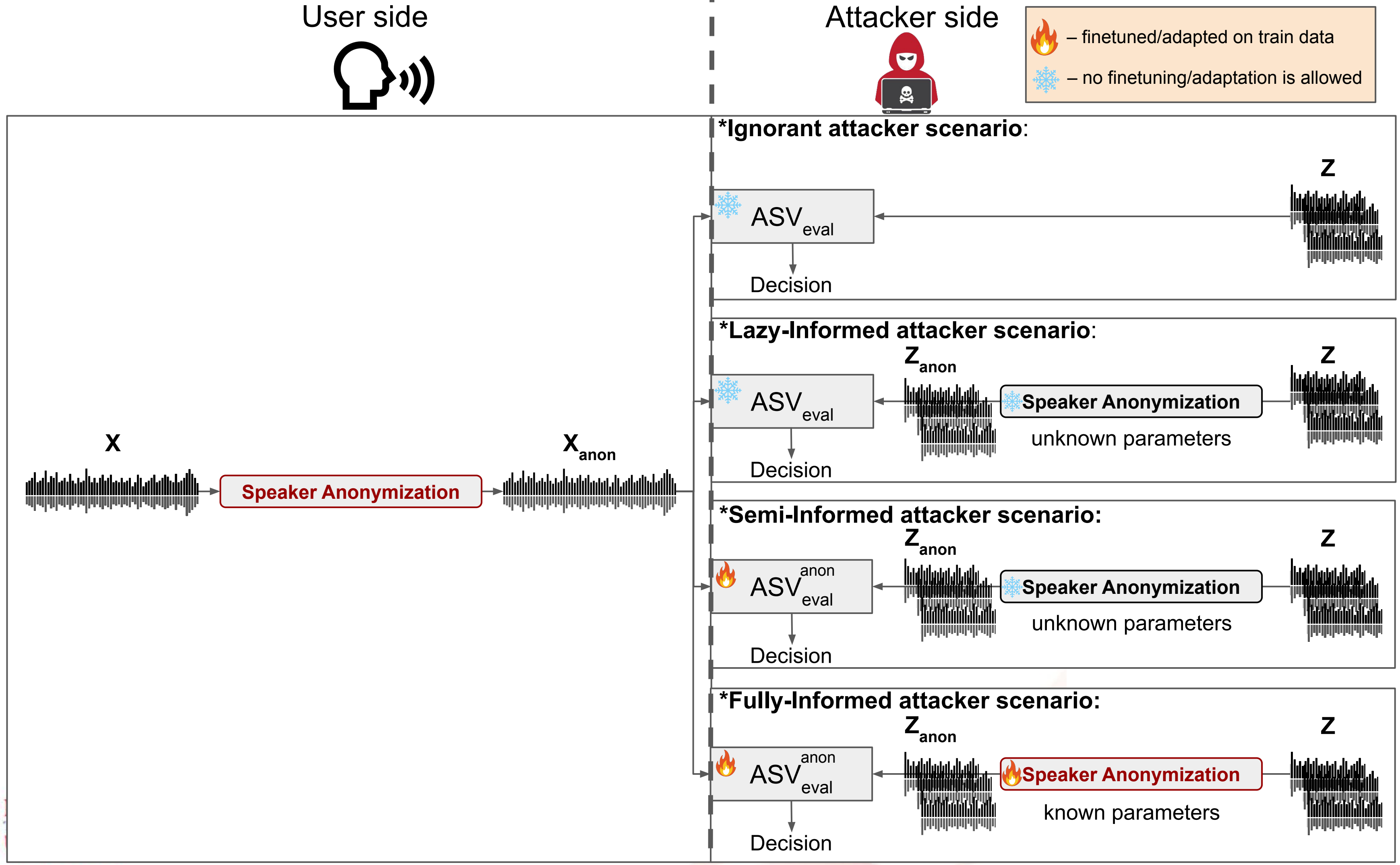
🟥 – removed elements

🟦 – unchanged elements



	VPC2020	VPC2022	VPC2024
Attackers:	✚ Lazy-informed	✚ Semi-informed 🟦 Lazy-informed	🟦 Semi-informed 🟥 Lazy-informed

Srivastava, Brij Mohan Lal, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, « Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers », in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020
N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.



User side

Attacker side

🔥 – finetuned/adapted on train data

❄ – no finetuning/adaptation is allowed

X → Speaker Anonymization → X_{anon}

*Ignorant attacker scenario:

❄ ASV_{eval} ← Z

Decision

*Lazy-Informed attacker scenario:

❄ ASV_{eval} ← Z_{anon} ← ❄ Speaker Anonymization ← Z

unknown parameters

Decision

*Semi-Informed attacker scenario:

🔥 ASV^{anon}_{eval} ← Z_{anon} ← ❄ Speaker Anonymization ← Z

unknown parameters

Decision

*Fully-Informed attacker scenario:

🔥 ASV^{anon}_{eval} ← Z_{anon} ← 🔥 Speaker Anonymization ← Z

known parameters

Decision

VPC evolution: Datasets

➕ – new elements

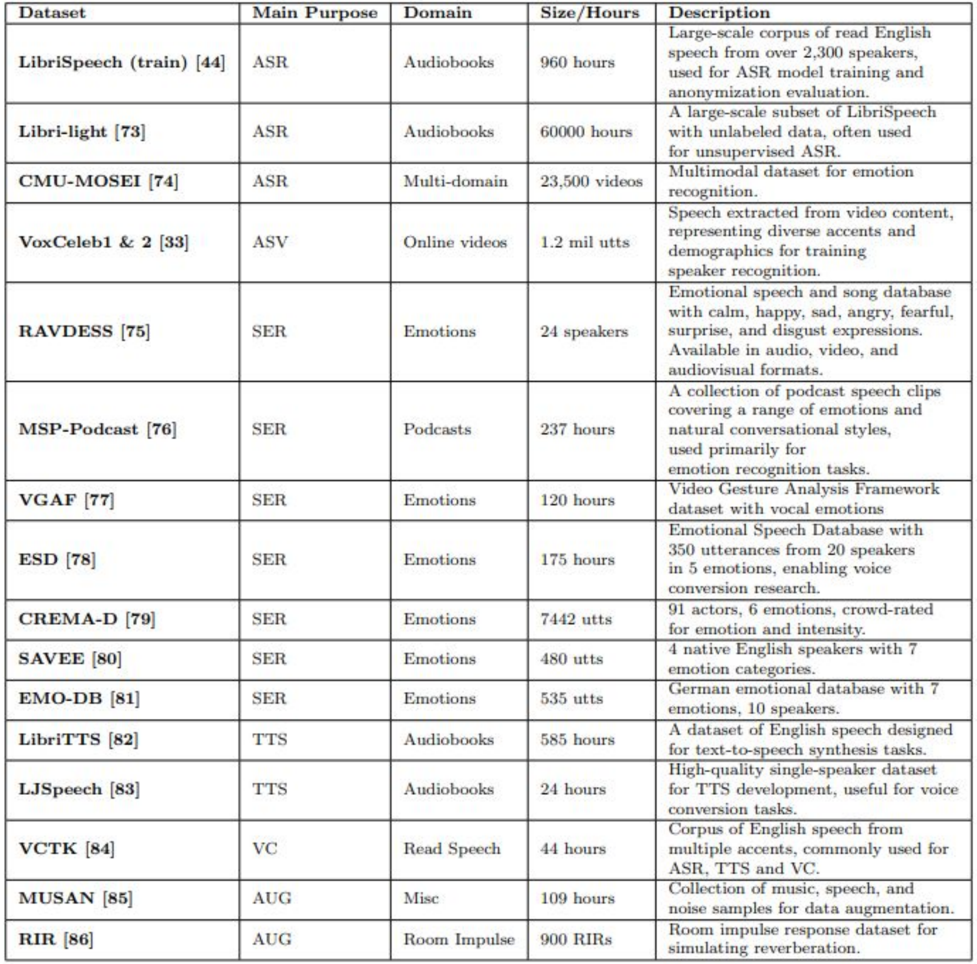
▬ – removed elements

〓 – unchanged elements

	VPC2020	VPC2022	VPC2024
Datasets:	➕ Provided common datasets: VCTK, LibriSpeech, LibriTTS, VoxCeleb	〓 Same datasets as in VPC2020	➕ Extended datasets and pretrained models
Privacy:	**train:** LibriSpeech train-clean 360 **eval:** LibriSpeech test-clean, VCTK test		**train: *check the table on the next slide** **eval:** LibriSpeech **dev**, test-clean
Utility:	**train:** LibriSpeech train-clean 360 **eval:** LibriSpeech test-clean, VCTK test		**train: *check the table on the next slide** **eval:** LibriSpeech **dev,** test-clean; **IEMOCAP**

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multispeaker corpus for cstr voice cloning toolkit (version 0.92),” 2019
C. Busso et al., “Iemocap: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation,. 2008
N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.

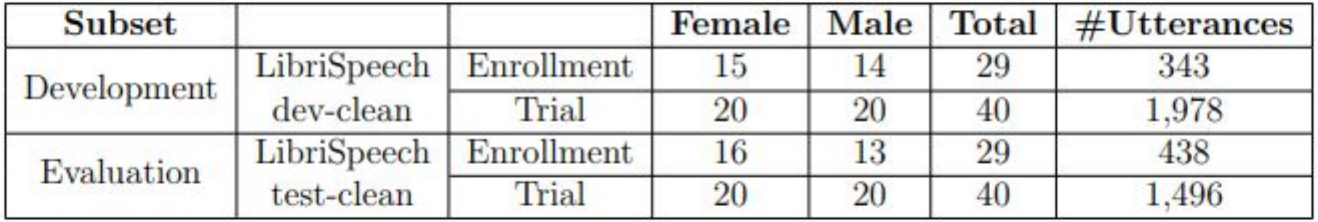
VPC evolution: Main Train Datasets



Dataset	Main Purpose	Domain	Size/Hours	Description
LibriSpeech (train) [44]	ASR	Audiobooks	960 hours	Large-scale corpus of read English speech from over 2,300 speakers, used for ASR model training and anonymization evaluation.
Libri-light [73]	ASR	Audiobooks	60000 hours	A large-scale subset of LibriSpeech with unlabeled data, often used for unsupervised ASR.
CMU-MOSEI [74]	ASR	Multi-domain	23,500 videos	Multimodal dataset for emotion recognition.
VoxCeleb1 & 2 [33]	ASV	Online videos	1.2 mil utts	Speech extracted from video content, representing diverse accents and demographics for training speaker recognition.
RAVDESS [75]	SER	Emotions	24 speakers	Emotional speech and song database with calm, happy, sad, angry, fearful, surprise, and disgust expressions. Available in audio, video, and audiovisual formats.
MSP-Podcast [76]	SER	Podcasts	237 hours	A collection of podcast speech clips covering a range of emotions and natural conversational styles, used primarily for emotion recognition tasks.
VGAF [77]	SER	Emotions	120 hours	Video Gesture Analysis Framework dataset with vocal emotions
ESD [78]	SER	Emotions	175 hours	Emotional Speech Database with 350 utterances from 20 speakers in 5 emotions, enabling voice conversion research.
CREMA-D [79]	SER	Emotions	7442 utts	91 actors, 6 emotions, crowd-rated for emotion and intensity.
SAVEE [80]	SER	Emotions	480 utts	4 native English speakers with 7 emotion categories.
EMO-DB [81]	SER	Emotions	535 utts	German emotional database with 7 emotions, 10 speakers.
LibriTTS [82]	TTS	Audiobooks	585 hours	A dataset of English speech designed for text-to-speech synthesis tasks.
LJSpeech [83]	TTS	Audiobooks	24 hours	High-quality single-speaker dataset for TTS development, useful for voice conversion tasks.
VCTK [84]	VC	Read Speech	44 hours	Corpus of English speech from multiple accents, commonly used for ASR, TTS and VC.
MUSAN [85]	AUG	Misc	109 hours	Collection of music, speech, and noise samples for data augmentation.
RIR [86]	AUG	Room Impulse	900 RIRs	Room impulse response dataset for simulating reverberation.

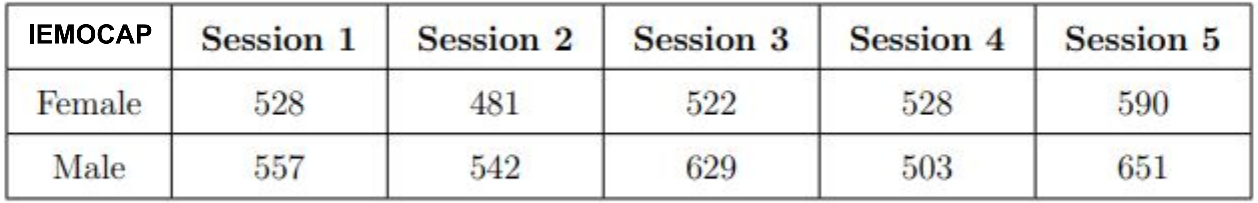
VPC evolution: Main Eval Datasets

Automatic Speech Recognition (ASR) and Privacy Evaluation (ASV):



Subset			Female	Male	Total	#Utterances
Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496

Speech Emotion Recognition (SER):



IEMOCAP	Session 1	Session 2	Session 3	Session 4	Session 5
Female	528	481	522	528	590
Male	557	542	629	503	651

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multispeaker corpus for cstr voice cloning toolkit (version 0.92),” 2019
C. Busso et al., “Iemocap: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation,. 2008
N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.

VPC evolution: Evaluation Metrics

➕	– new elements
▬	– removed elements
〓	– unchanged elements

	VPC2020	VPC2022	VPC2024
Metrics:	➕ Metrics: EER and WER, subjective metrics	➕ New metrics: GVD and pitch correlation New eval: EER$_1$, EER$_2$, EER$_3$, EER$_4$	➕ New utility metric: UAR for SER
		〓 Metrics: EER and WER, subjective metrics	▬ GVD, pitch correlation and subjective metrics

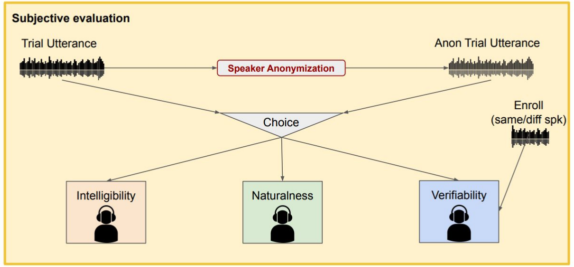
$$\mathrm{EER} = P_{\mathrm{fa}}(\theta_{\mathrm{EER}}) = P_{\mathrm{miss}}(\theta_{\mathrm{EER}})$$

$$G_{\mathrm{VD}} = 10 \log_{10} \frac{D_{\mathrm{diag}}(S_{\mathrm{anon}})}{D_{\mathrm{diag}}(S_{\mathrm{orig}})}$$

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

$$\rho_{F_0} = \frac{\sum_{t=1}^{T}(P_t - \bar{P})(Q_t - \bar{Q})}{\sqrt{\sum_{i=t}^{T}(P_t - \bar{P})^2}\sqrt{\sum_{t=1}^{T}(Q_t - \bar{Q})^2}},$$

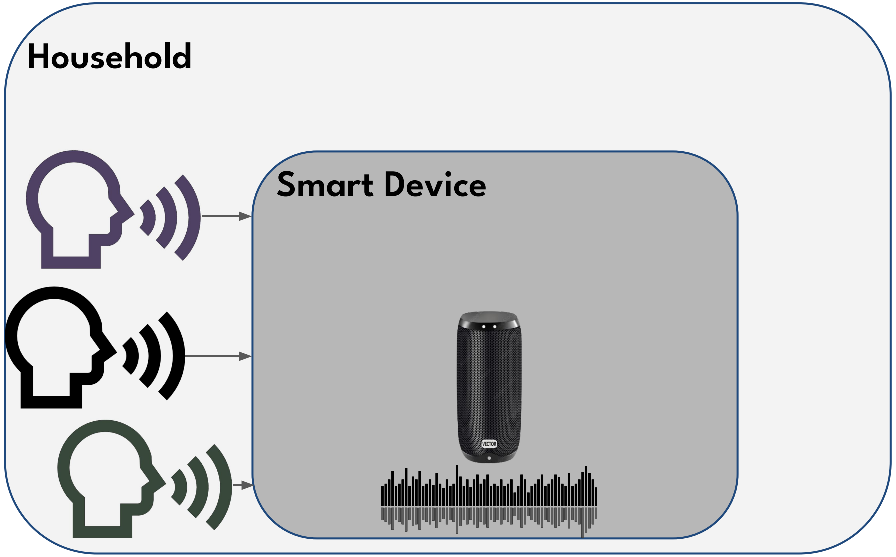
$$UAR = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}$$



VPC evolution: Different Privacy Requirements

❗ **High utility** requirements ❗

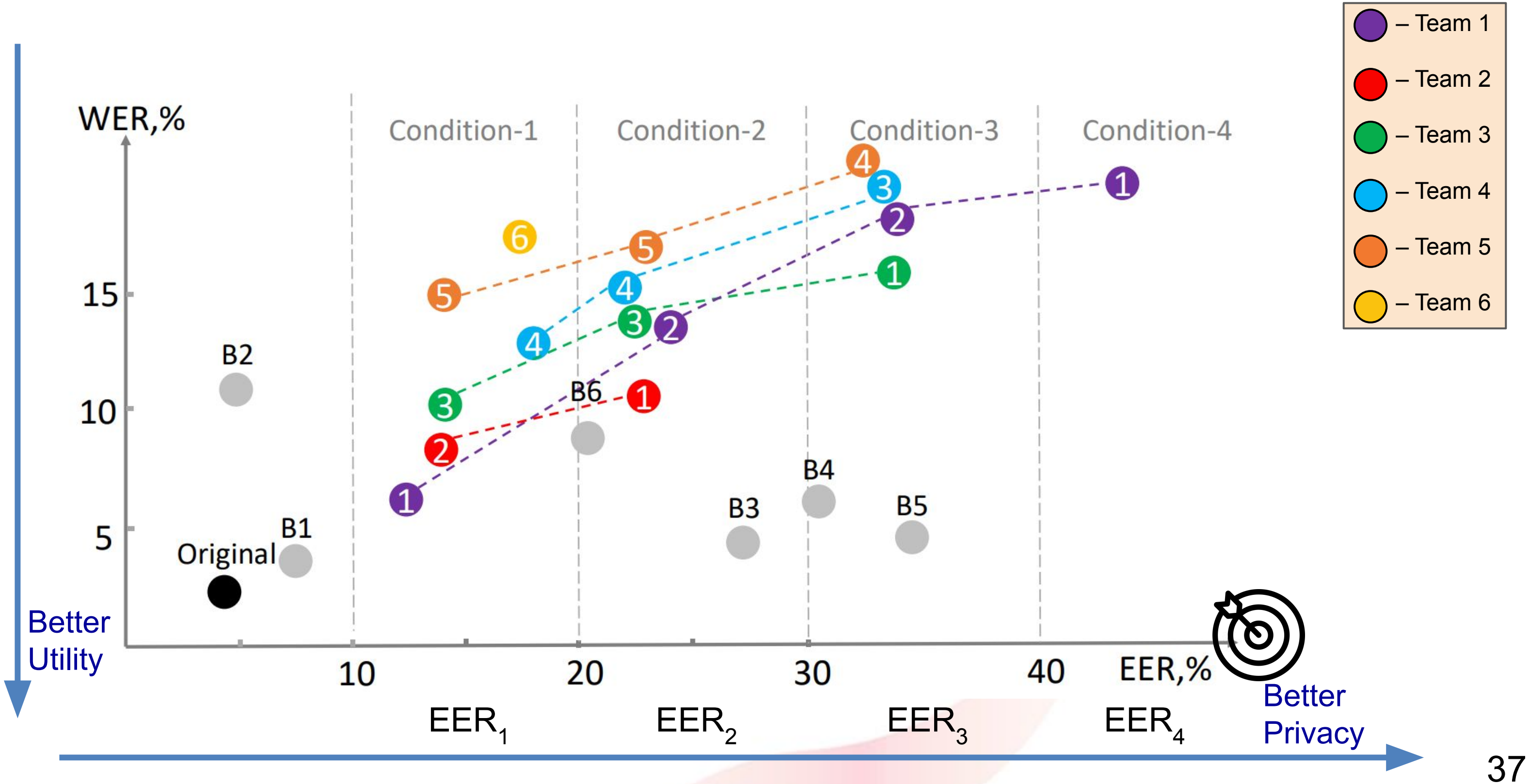
Low privacy requirements



Low utility requirements

❗ **High privacy** requirements ❗

VPC evolution: Different Privacy Requirements



N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.

Outline

1. Introduction

2. Literature Review
   a. Speaker Anonymization
   b. Disentanglement Learning

3. **Disentanglement-based Approaches for Anonymization**
   a. **Problem 1:** Anonymization models do not preserve emotions
      i. **Contribution 1.1:** NS3 FACodec
      ii. **Contribution 1.2:** Emotion embeddings
   b. **Problem 2:** Identity Leakage in B5
      i. **Contribution 2:** Mean-reversion + Noise

4. Conclusions and Future Work

Motivation

What is the motivation?

- **Problem1:** Current models do not preserve emotions
- **Problem2:** Identity Leakage in B5 system.
- Cover all Privacy conditions in VPC

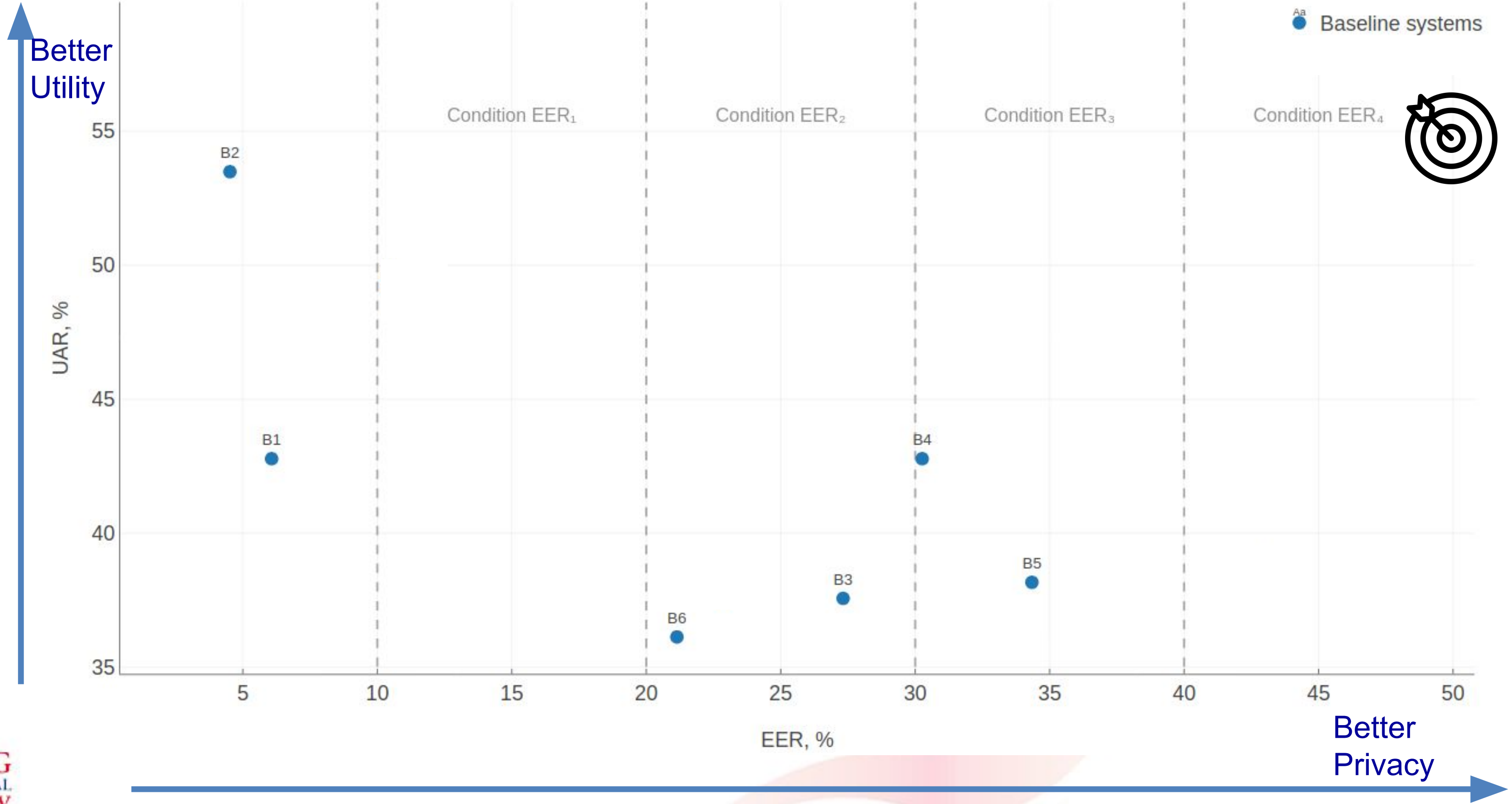
How do we tackle these problems?

- Enhance multiple approaches:
  - NaturalSpeech3 FACodec for Speaker Anonymization with emotion preservation
  - Emotion Embedding to preserve emotions
  - MeanReversion + AWGN for F0 to enhance prosody protection

N. Tomashenko et al., “The VoicePrivacy 2024 challenge evaluation plan,” 2024.
Z. Ju et al., “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models”, in ICML 2024
P. Champion, “Anonymizing speech: Evaluating and designing speaker anonymization techniques,” Ph.D. dissertation, Universit´e de Lorraine, 2023

Problem 1: Anonymization models do not preserve emotions

Problem 1: Anonymization models do not preserve emotions

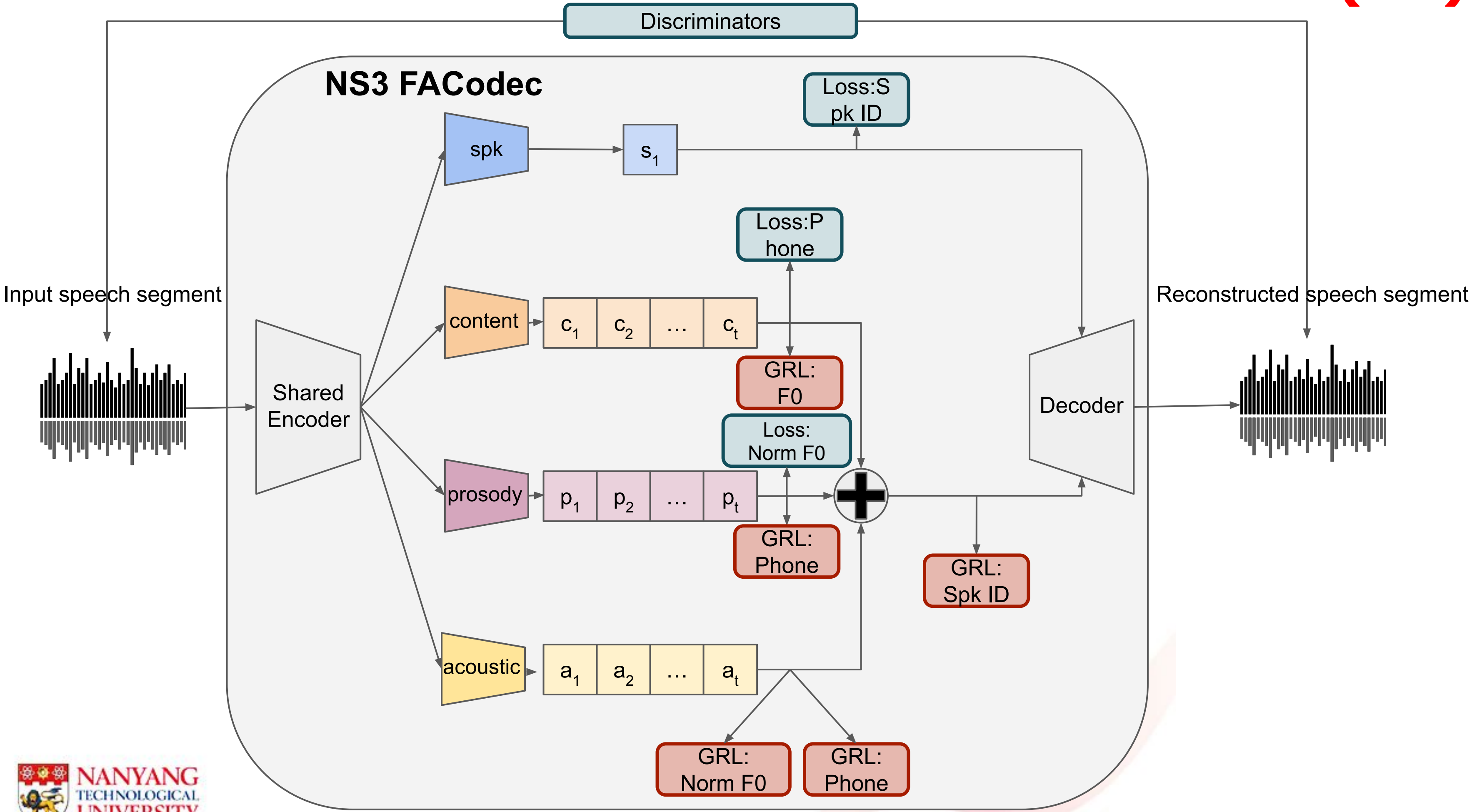


Problem 1: Qualitative Examples

Original	B3	Proposed (1.1)	Proposed (1.2)
🔊	🔊	🔊	🔊

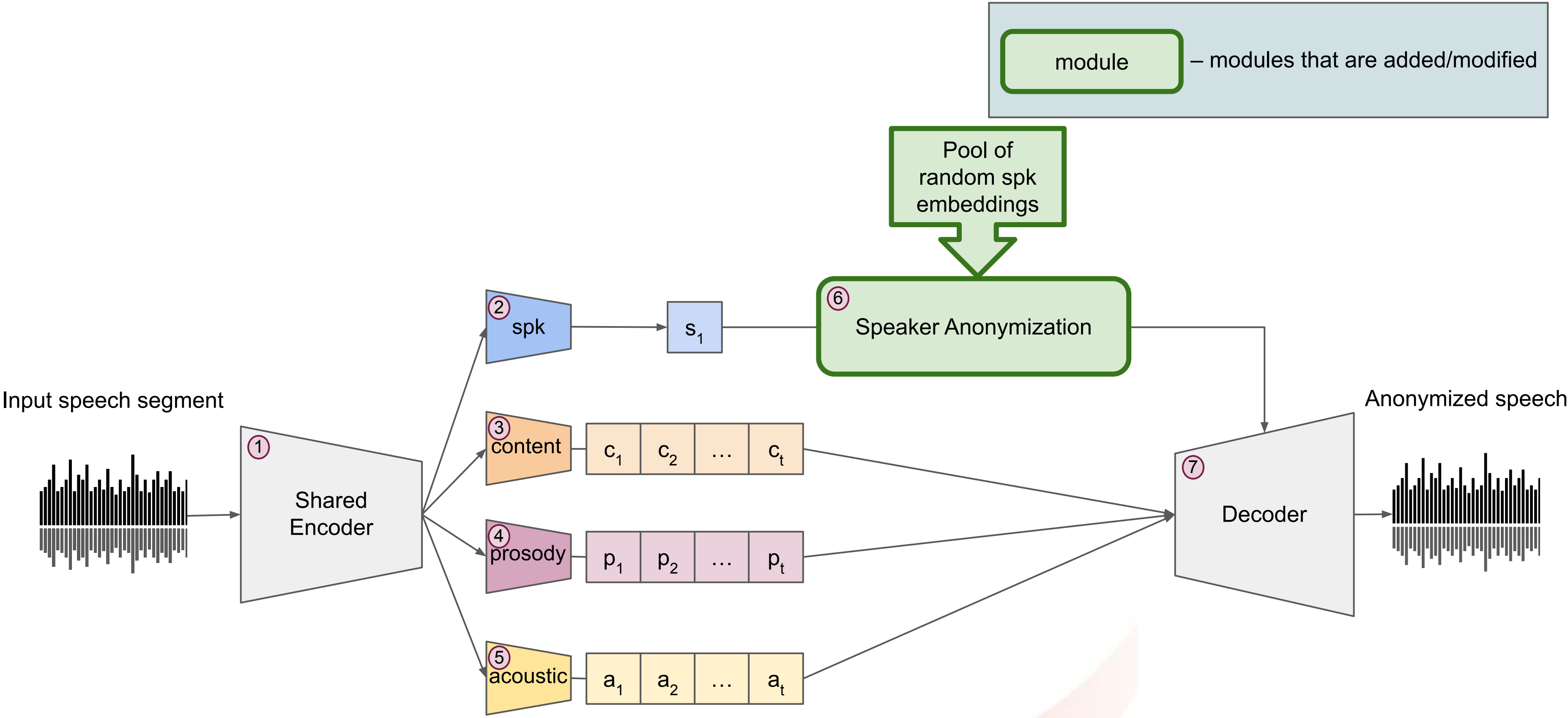
Contribution 1.1: Modified NaturalSpeech3 FACodec

Contribution 1.1: Modified NS3 FACodec (1a)



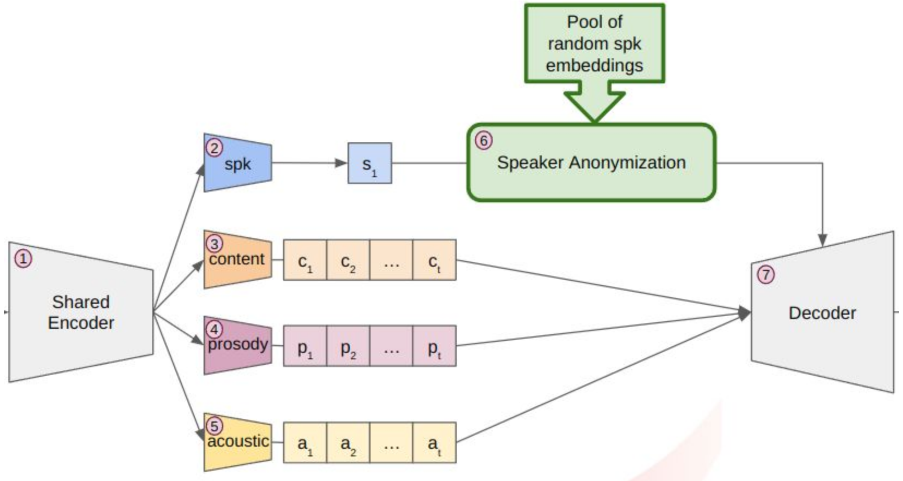
Z. Ju et al., “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models”, in ICML 2024

Contribution 1.1: Modified NS3 FACodec (1a)



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution 1.1: Modified NS3 FACodec (1a)

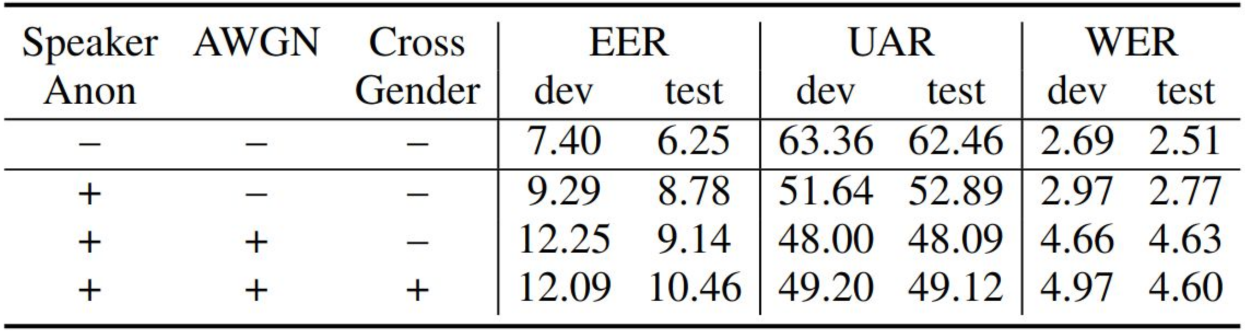


#	Module	Description	Output features	Data
①	Encoder [124]	4 Downsampling Convolution-based Layers with Snake activation function Input: speech waveform	Output vector^{256}	Librilight train [125]
②	Speaker embedding extractor	Several Conformer blocks	Speaker embedding^{256}	Librilight train
③	Content extractor	Factorized Vector Quantization with 2 quantizers, codebook size: 1024	Content vector^{256}	Librilight train
④	Prosody extractor	Factorized Vector Quantization with 1 quantizer, codebook size: 1024	Prosody vector^{256}	Librilight train
⑤	Acoustic extractor	Factorized Vector Quantization with 3 quantizer, codebook size: 1024	Acoustic vector^{256}	Librilight train
⑥	Speaker anonymization module	Averaged 100 embeddings randomly selected from a pool of 200 farthest embeddings from source by cosine scoring AWGN with scale= 0.075 Cross-gender	Anonymized speaker embedding^{256}	LibriTTS: train-clean-100
⑦	Decoder [124]	Upsampling Convolution-based Layers with Snake activation function	speech waveform	Librilight train

NANYANG TECHNOLOGICAL UNIVERSITY

N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

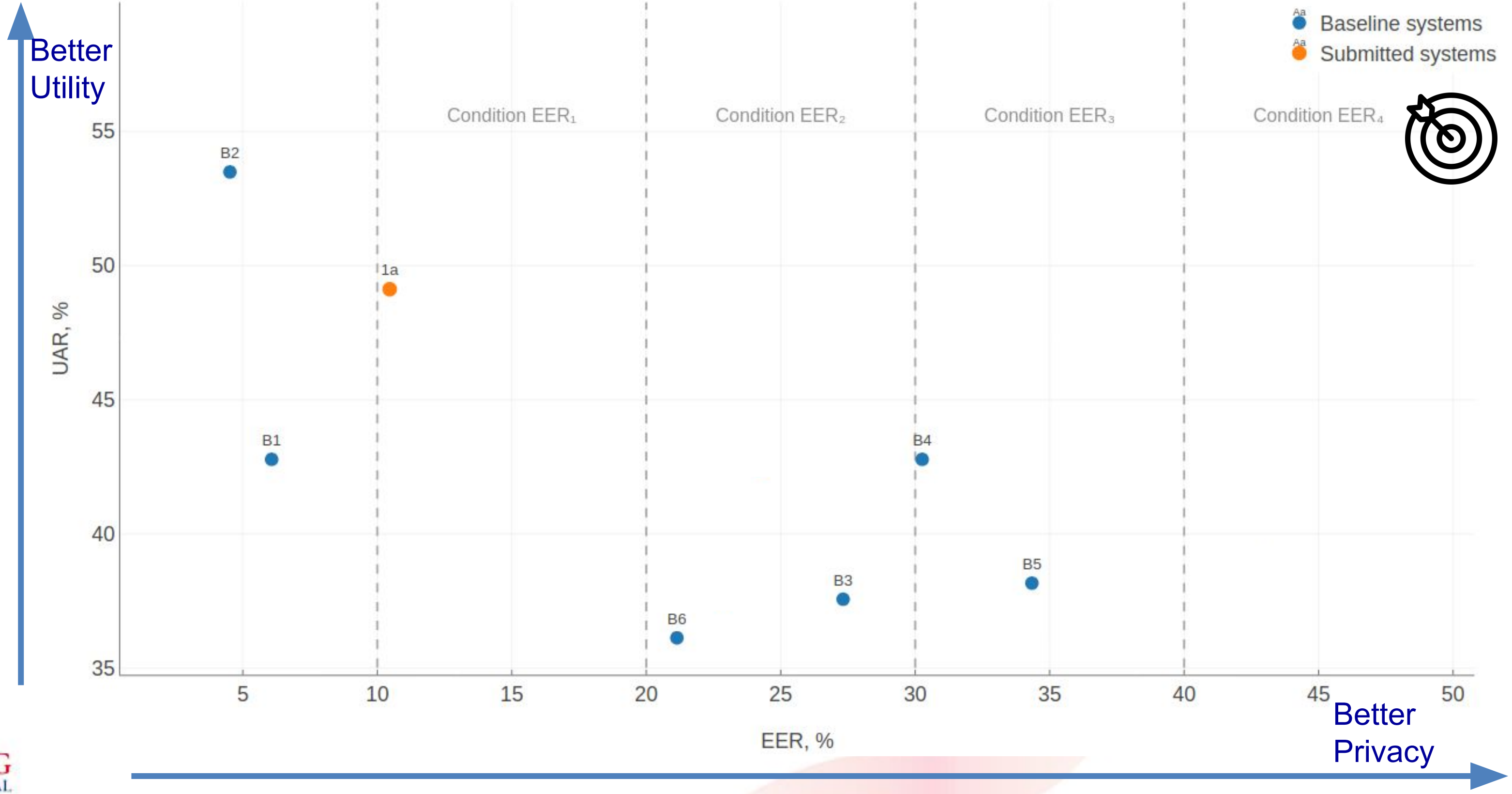
Experiments. NS3 + AWGN to Speaker Embedding + Cross Gender



Speaker	AWGN	Cross	EER		UAR		WER	
Anon		Gender	dev	test	dev	test	dev	test
–	–	–	7.40	6.25	63.36	62.46	2.69	2.51
+	–	–	9.29	8.78	51.64	52.89	2.97	2.77
+	+	–	12.25	9.14	48.00	48.09	4.66	4.63
+	+	+	12.09	10.46	49.20	49.12	4.97	4.60

N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

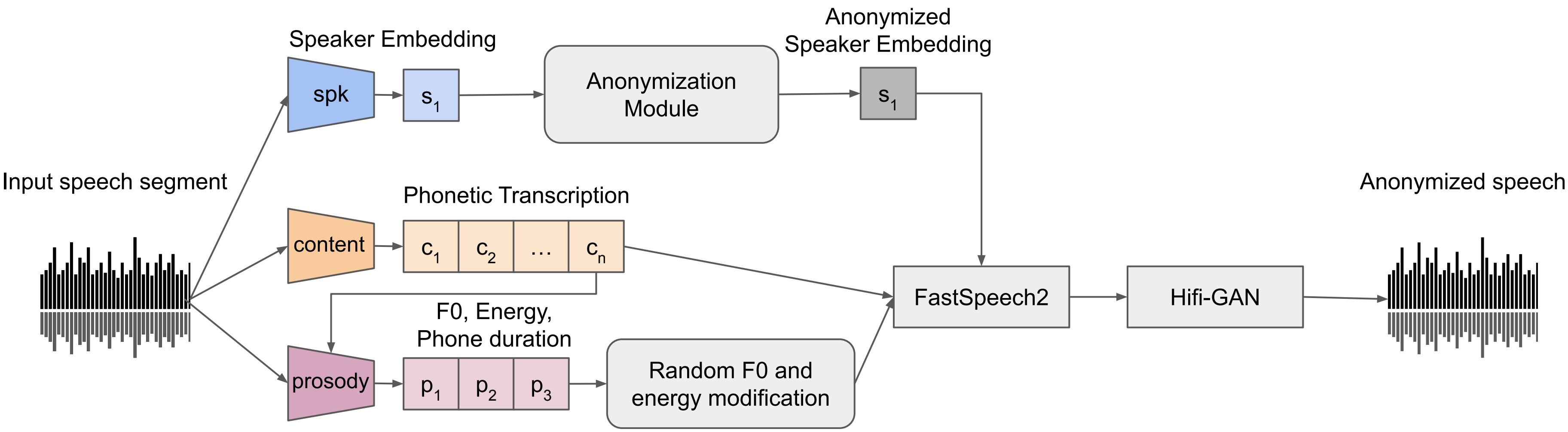
Contribution 1.1: Modified NS3 FACodec (1a). SER Results.



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

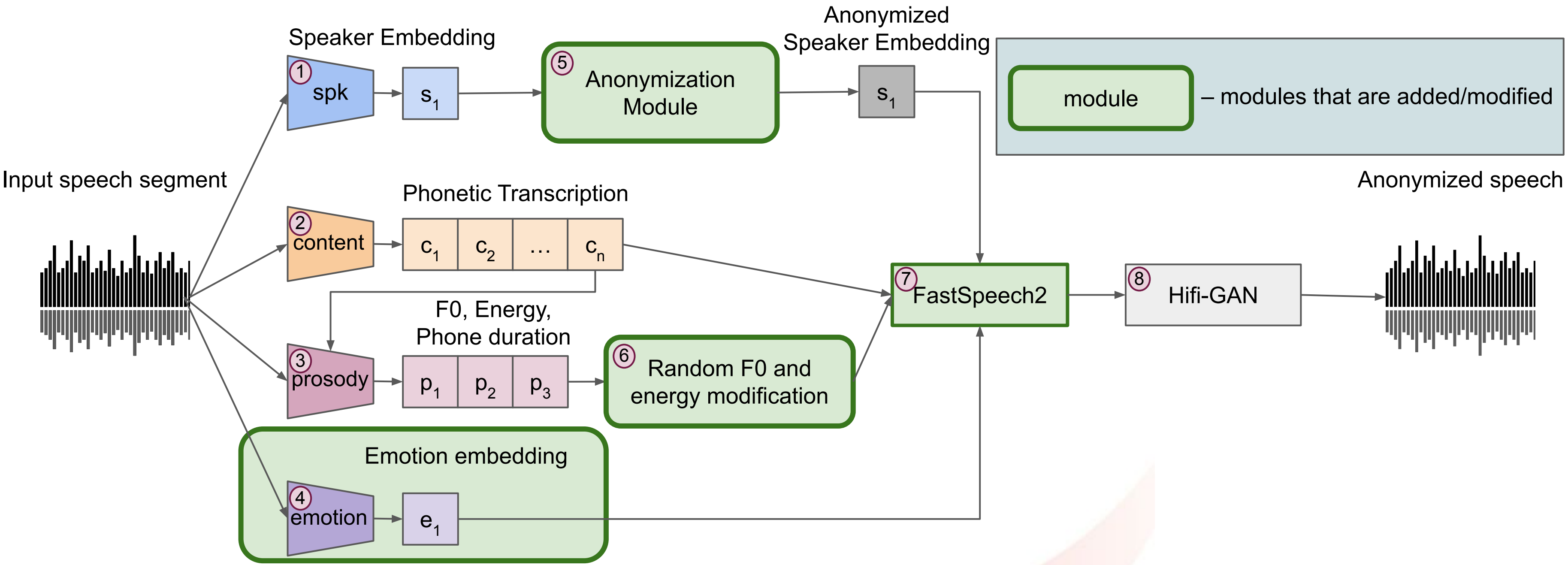
Contribution 1.2: Emotion embeddings for B3

Contribution 1.2: Emotion embeddings for B3 (Sys 1b, 2a)



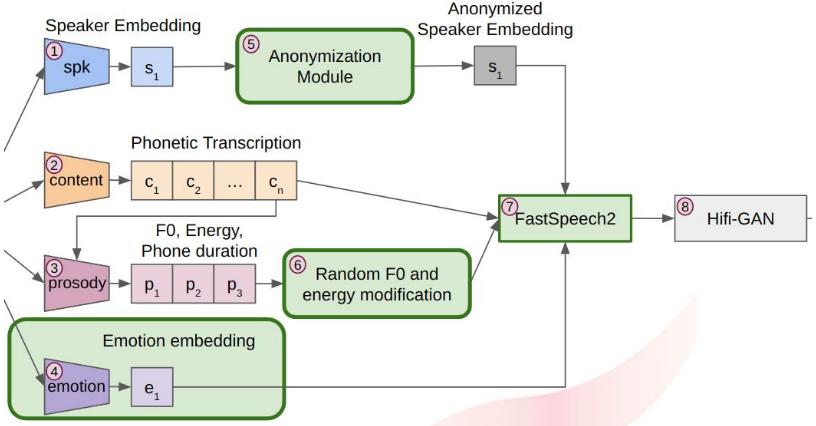
S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and T. Vu, “Prosody is not identity: A speaker anonymization approach using prosody cloning,” ICASSP, 2023
S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, “Anonymizing speech with generative adversarial networks to preserve speaker privacy,” in IEEE Spoken Language Technology Workshop (SLT), 2023

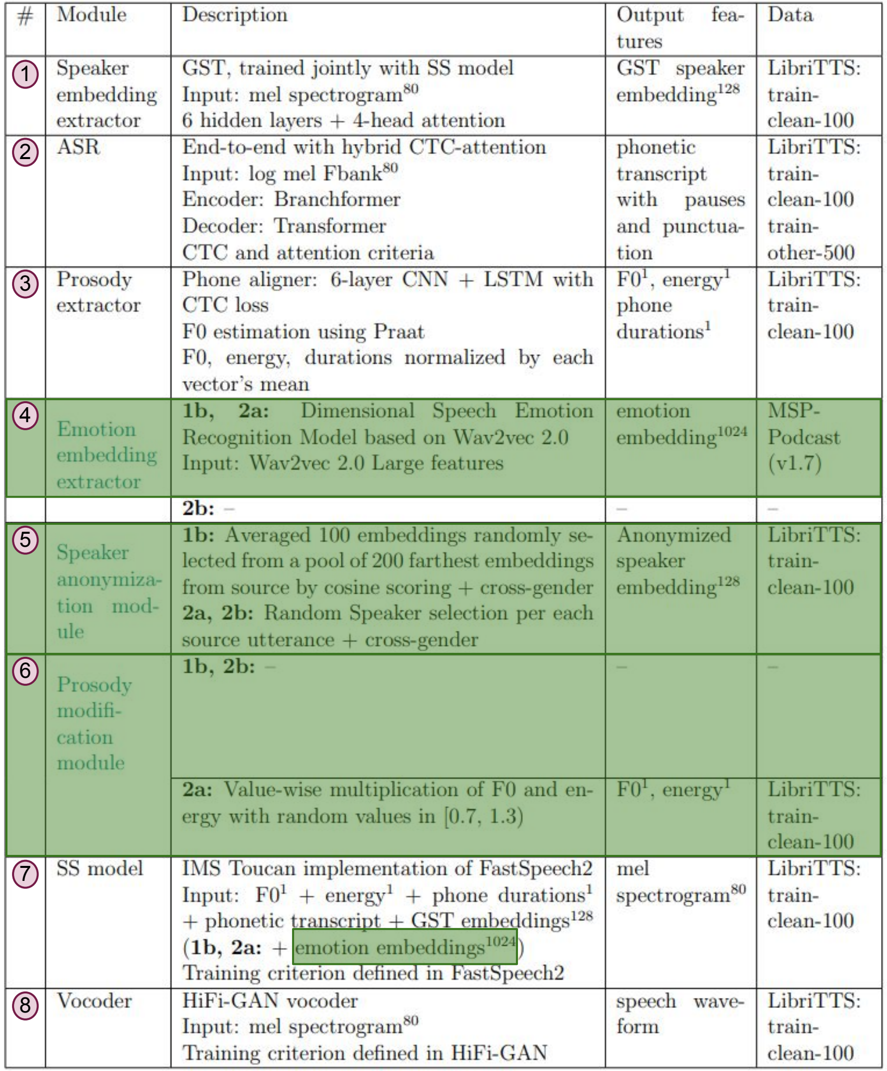
Contribution 1.2: Emotion embeddings for B3 (Sys 1b, 2a)



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

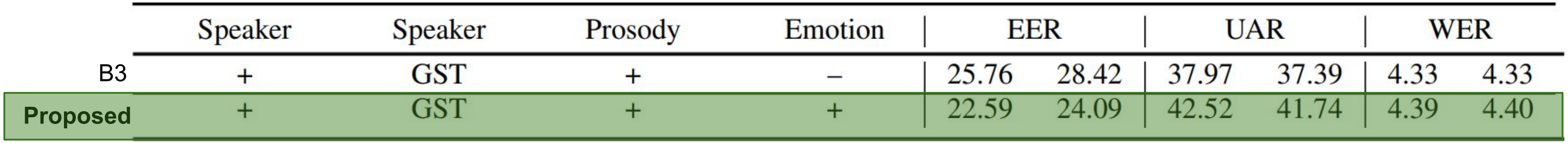
Modified B3





#	Module	Description	Output features	Data
①	Speaker embedding extractor	GST, trained jointly with SS model Input: mel spectrogram^{80} 6 hidden layers + 4-head attention	GST speaker embedding^{128}	LibriTTS: train-clean-100
②	ASR	End-to-end with hybrid CTC-attention Input: log mel Fbank^{80} Encoder: Branchformer Decoder: Transformer CTC and attention criteria	phonetic transcript with pauses and punctuation	LibriTTS: train-clean-100 train-other-500
③	Prosody extractor	Phone aligner: 6-layer CNN + LSTM with CTC loss F0 estimation using Praat F0, energy, durations normalized by each vector's mean	F0^1, energy^1 phone durations^1	LibriTTS: train-clean-100
④	Emotion embedding extractor	**1b, 2a:** Dimensional Speech Emotion Recognition Model based on Wav2vec 2.0 Input: Wav2vec 2.0 Large features	emotion embedding^{1024}	MSP-Podcast (v1.7)
		2b: –	–	–
⑤	Speaker anonymization module	**1b:** Averaged 100 embeddings randomly selected from a pool of 200 farthest embeddings from source by cosine scoring + cross-gender **2a, 2b:** Random Speaker selection per each source utterance + cross-gender	Anonymized speaker embedding^{128}	LibriTTS: train-clean-100
⑥	Prosody modification module	**1b, 2b:** –	–	–
		2a: Value-wise multiplication of F0 and energy with random values in [0.7, 1.3)	F0^1, energy^1	LibriTTS: train-clean-100
⑦	SS model	IMS Toucan implementation of FastSpeech2 Input: F0^1 + energy^1 + phone durations^1 + phonetic transcript + GST embeddings^{128} (**1b, 2a:** + emotion embeddings^{1024}) Training criterion defined in FastSpeech2	mel spectrogram^{80}	LibriTTS: train-clean-100
⑧	Vocoder	HiFi-GAN vocoder Input: mel spectrogram^{80} Training criterion defined in HiFi-GAN	speech waveform	LibriTTS: train-clean-100

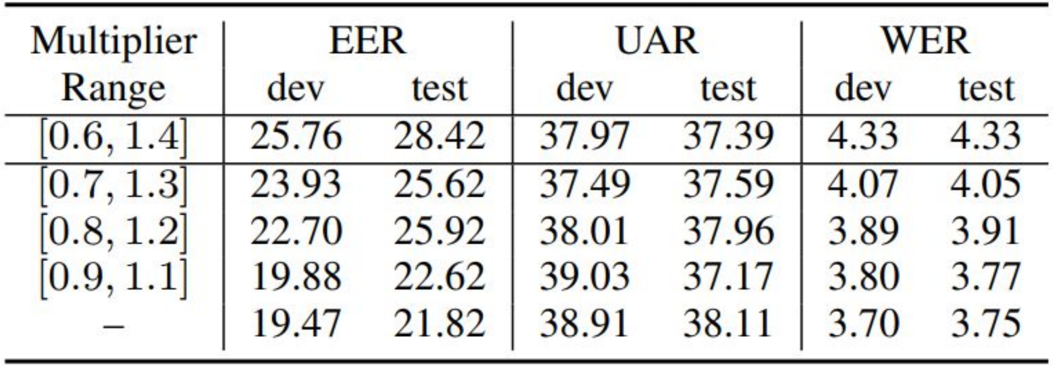
Experiments. B3 + Emotion embedding



	Speaker	Speaker	Prosody	Emotion	EER		UAR		WER	
B3	+	GST	+	–	25.76	28.42	37.97	37.39	4.33	4.33
Proposed	+	GST	+	+	22.59	24.09	42.52	41.74	4.39	4.40

N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

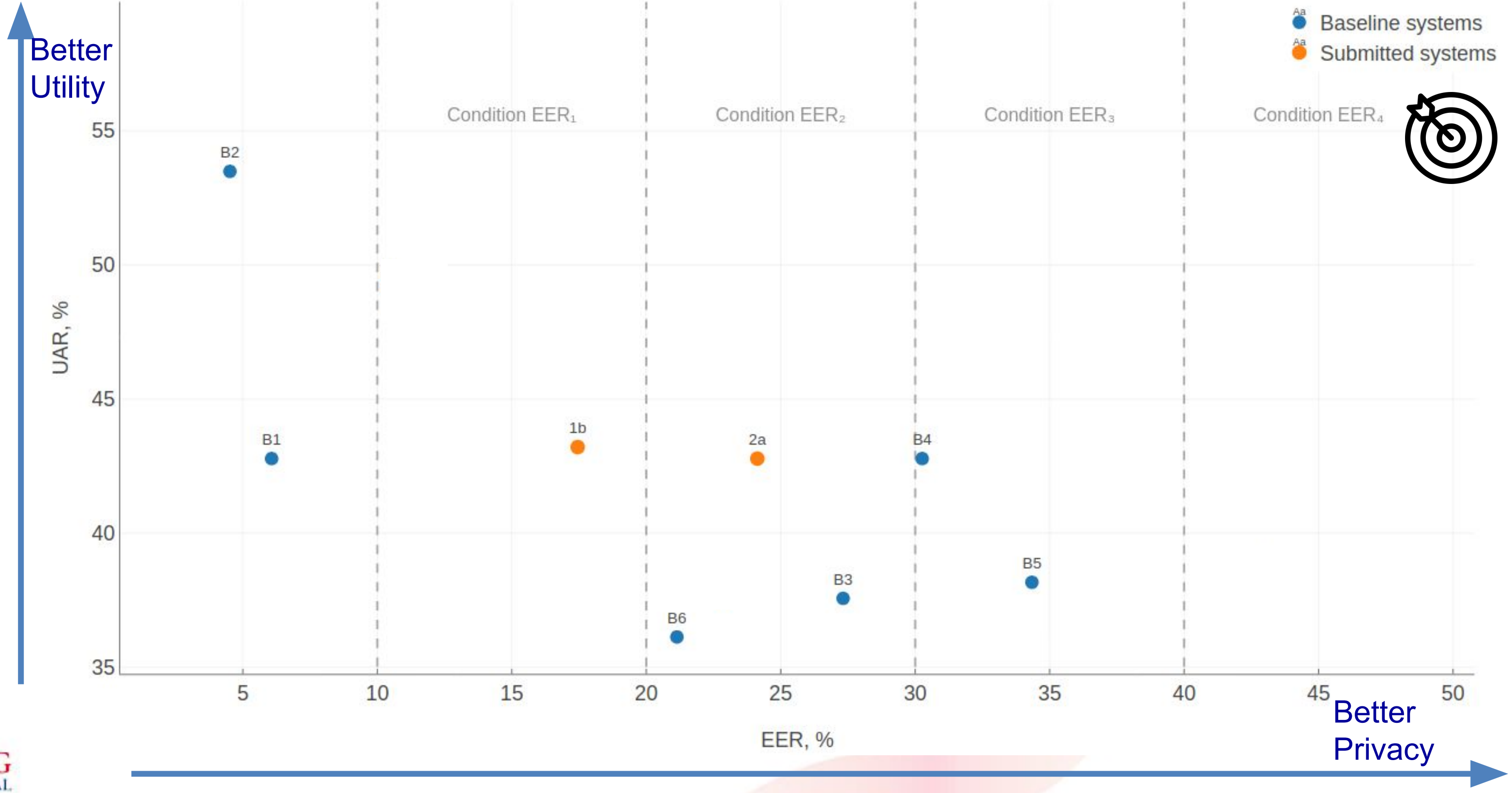
Experiments. B3 + Prosody Modification



Multiplier Range	EER		UAR		WER	
	dev	test	dev	test	dev	test
[0.6, 1.4]	25.76	28.42	37.97	37.39	4.33	4.33
[0.7, 1.3]	23.93	25.62	37.49	37.59	4.07	4.05
[0.8, 1.2]	22.70	25.92	38.01	37.96	3.89	3.91
[0.9, 1.1]	19.88	22.62	39.03	37.17	3.80	3.77
–	19.47	21.82	38.91	38.11	3.70	3.75

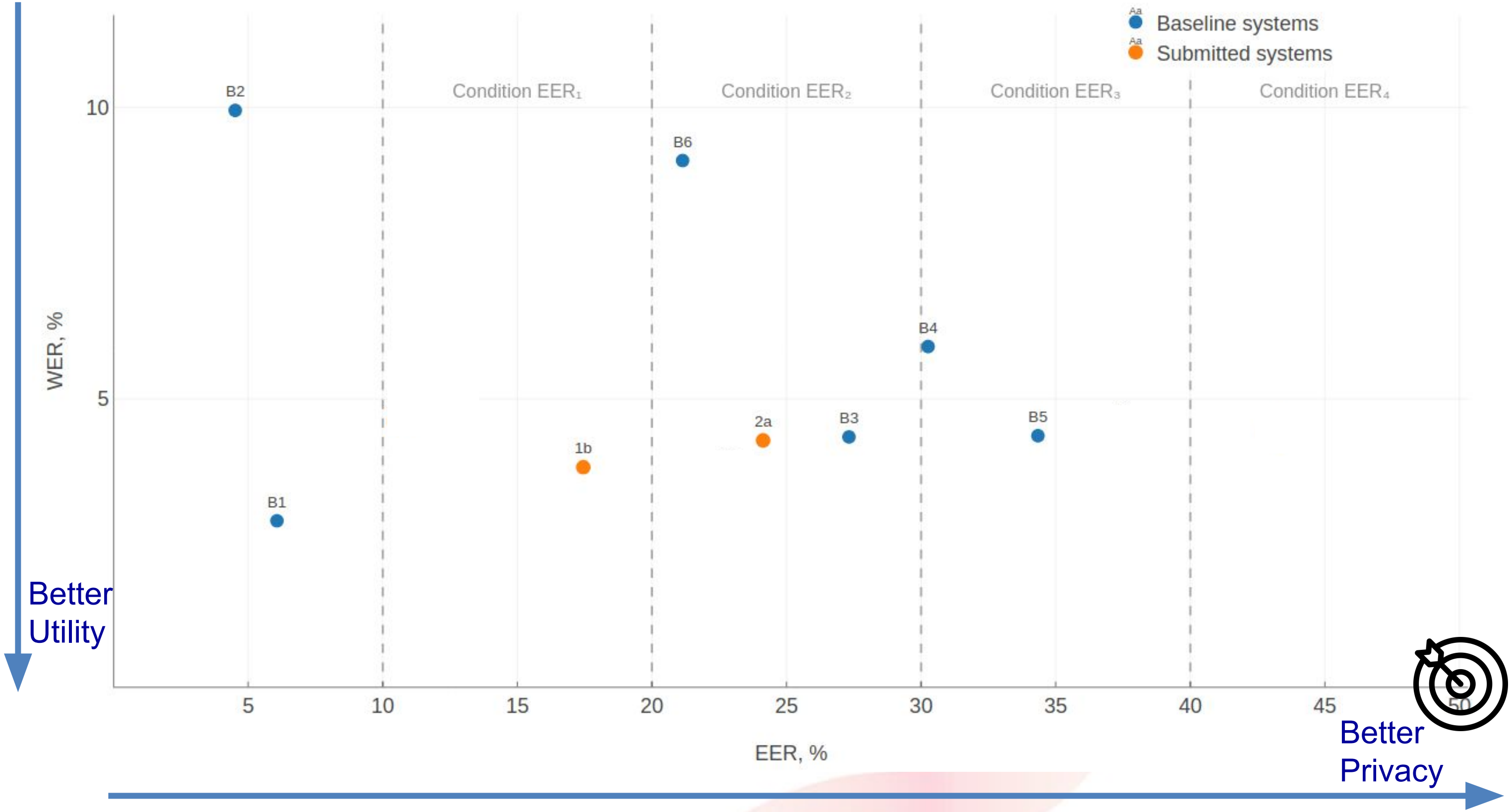
N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution 1.2: SER performance (Sys 1b, 2a)



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution 1.2: ASR performance (Sys 1b, 2a)



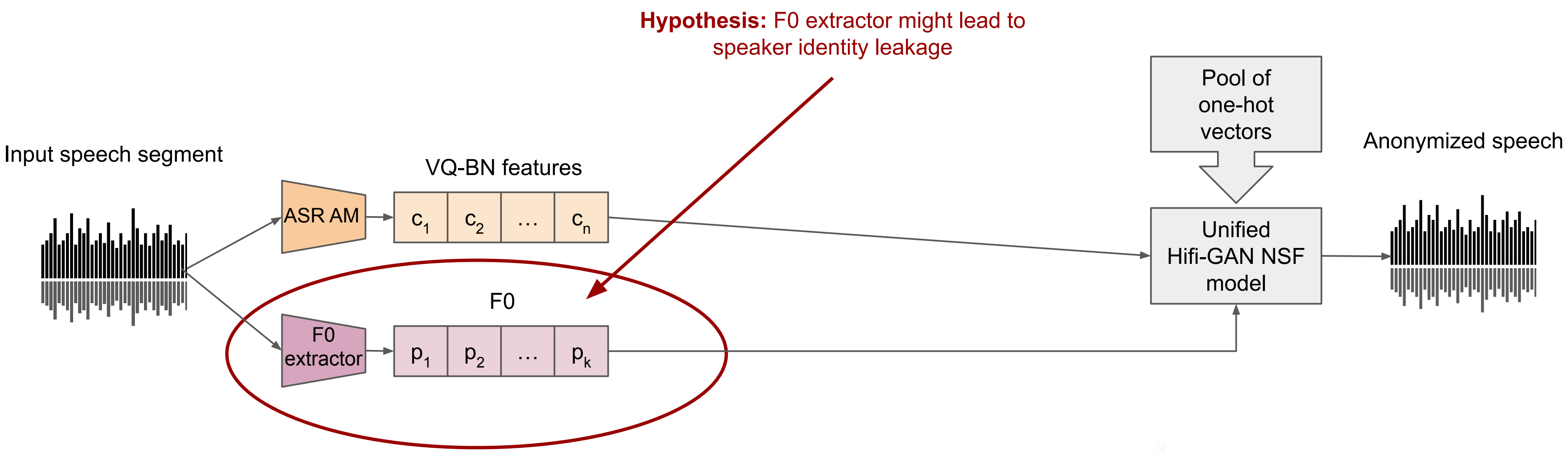
N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution: Examples

Original	B3	Proposed (1.1)	Proposed (1.2)
🔊	🔊	🔊	🔊

Problem2: Identity Leakage in B5 system

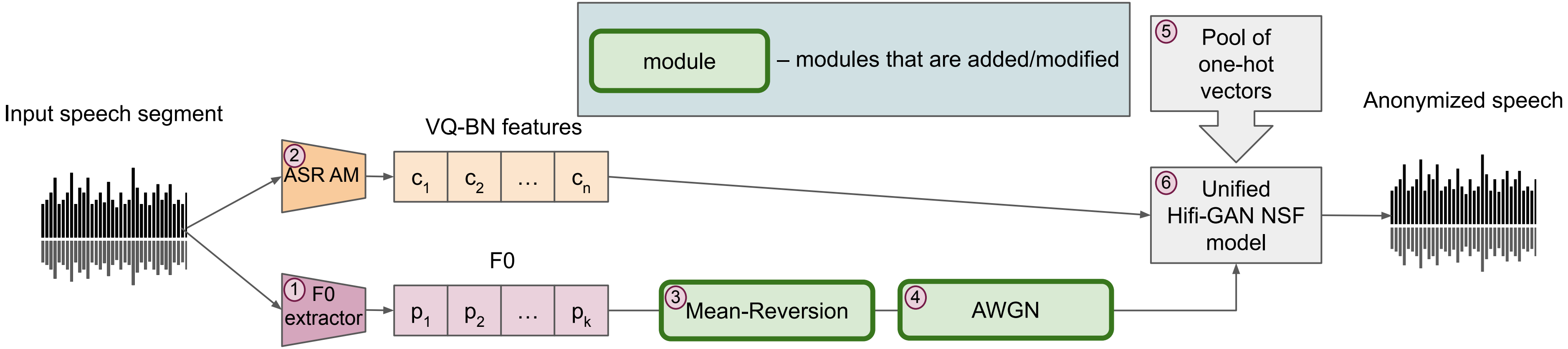
Problem2: Identity Leakage in B5 system



Hypothesis: F0 extractor might lead to speaker identity leakage

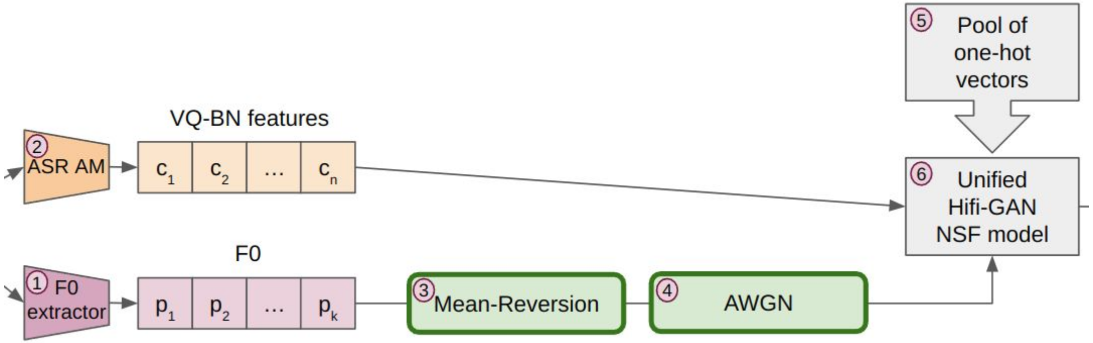
P. Champion, “Anonymizing speech: Evaluating and designing speaker anonymization techniques,” Ph.D. dissertation, Universit´e de Lorraine, 2023

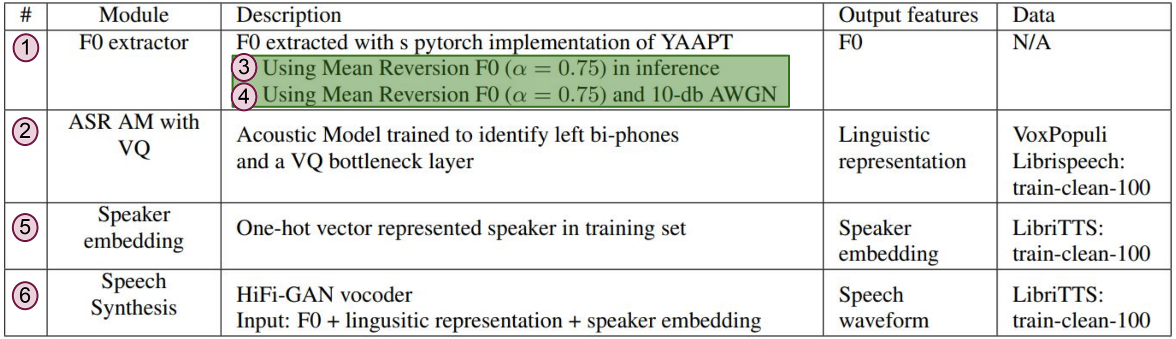
Contribution 2: Mean-reversion + Noise



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution 2: Details about Modified B5

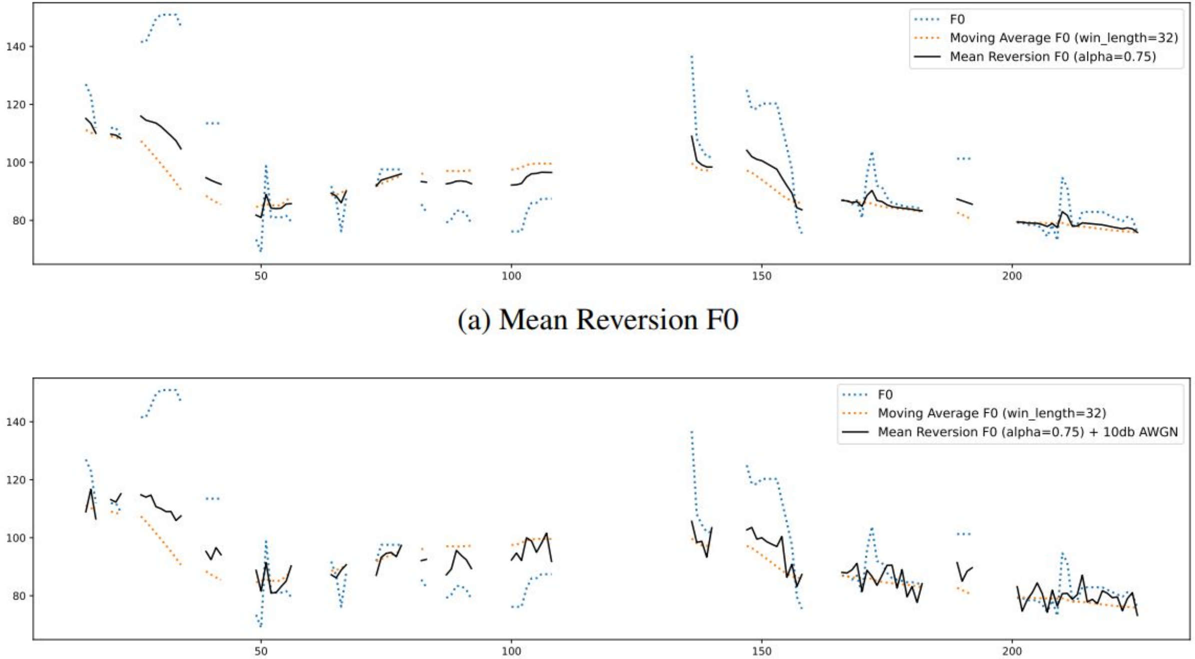




#	Module	Description	Output features	Data
①	F0 extractor	F0 extracted with s pytorch implementation of YAAPT ③ Using Mean Reversion F0 ($\alpha = 0.75$) in inference ④ Using Mean Reversion F0 ($\alpha = 0.75$) and 10-db AWGN	F0	N/A
②	ASR AM with VQ	Acoustic Model trained to identify left bi-phones and a VQ bottleneck layer	Linguistic representation	VoxPopuli Librispeech: train-clean-100
⑤	Speaker embedding	One-hot vector represented speaker in training set	Speaker embedding	LibriTTS: train-clean-100
⑥	Speech Synthesis	HiFi-GAN vocoder Input: F0 + lingusitic representation + speaker embedding	Speech waveform	LibriTTS: train-clean-100

N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution 2: Mean Reversion + AWGN

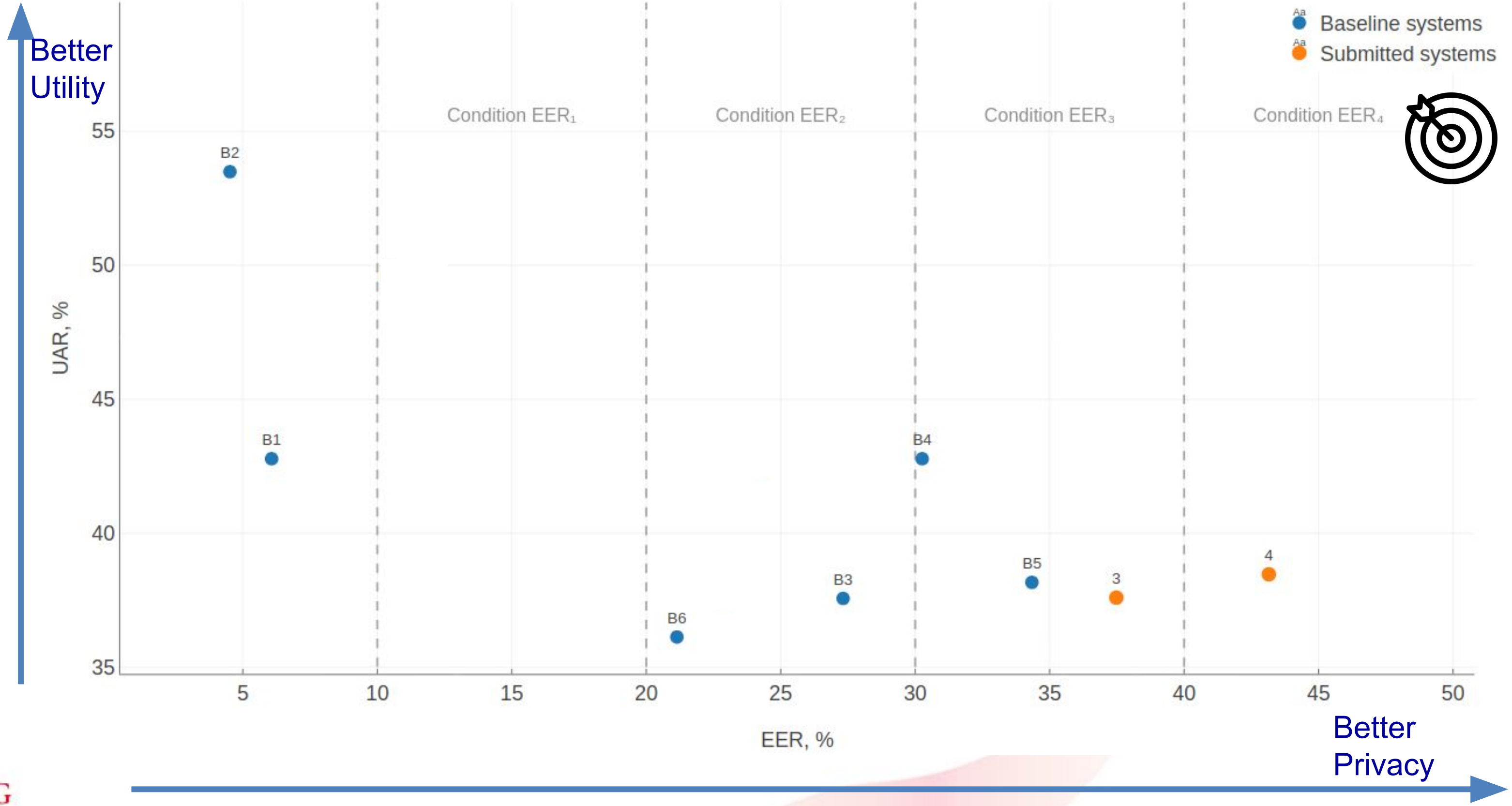


(a) Mean Reversion F0

(b) Mean Reversion F0 with a 10-dB white gaussian noise

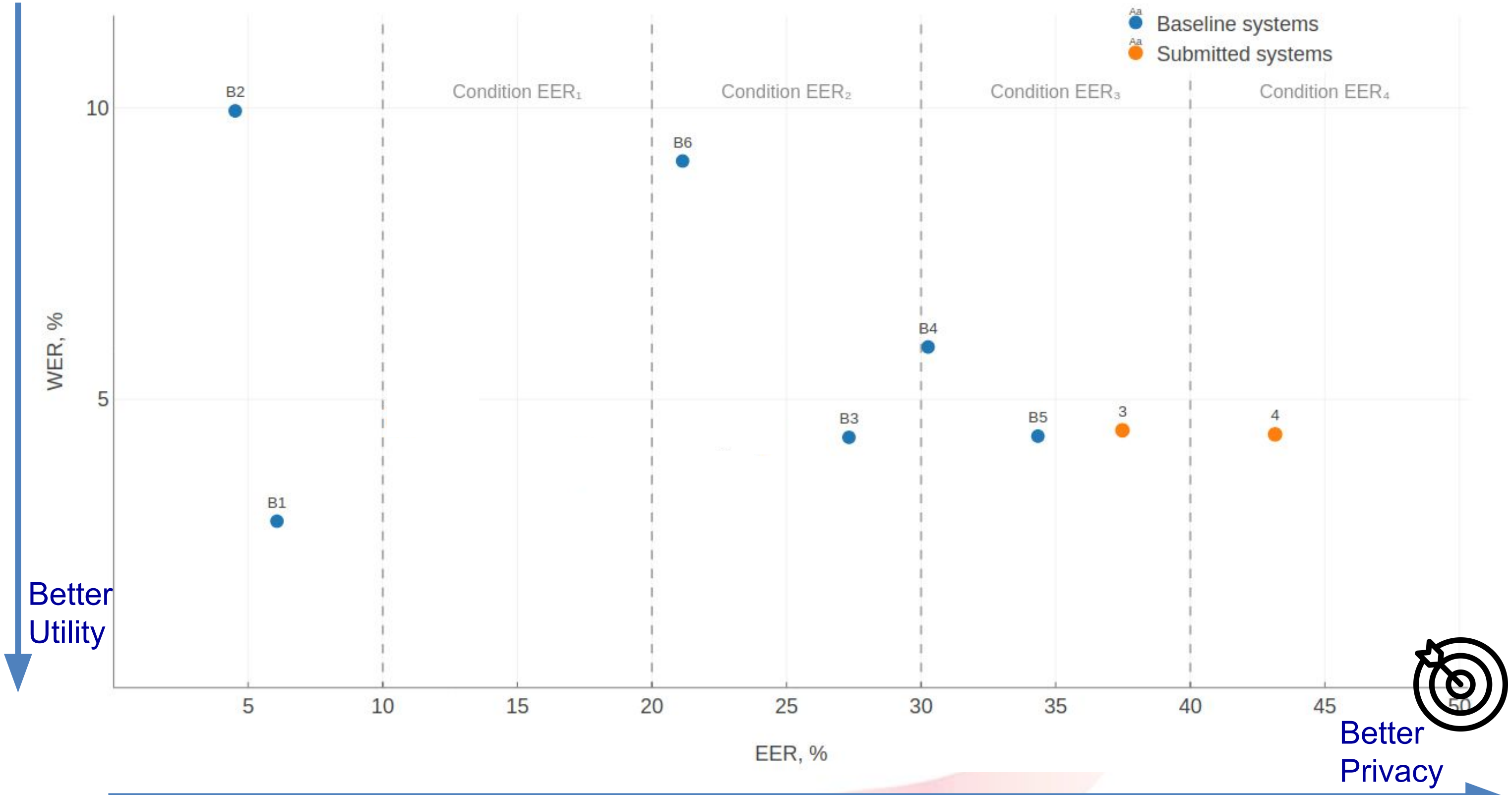
Figure 2: *Examples of Mean Reversion F0 with and without addictive noise*

Contribution 2: SER performance (Sys 3, 4)



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Contribution 2: ASR performance (Sys 3, 4)



N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

Key Takeaways

1. We achieved **3rd place** (out of 36 submitted systems) in Privacy Protection performance on VPC2024

2. NaturalSpeech3 FACodec:
   - Promising results for ER and ASR
   - But there may be leakage of speaker identity in other branches (content/acoustic)

3. Emotion Embeddings:
   - Helps to improve ER performance
   - But leads to speaker identity leakage

4. Mean-reversion of F_0 and AWGN:
   - Improves privacy protection while keeping ASR and ER

NANYANG TECHNOLOGICAL UNIVERSITY

N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

5. Conclusion & Future work

1. Introduction

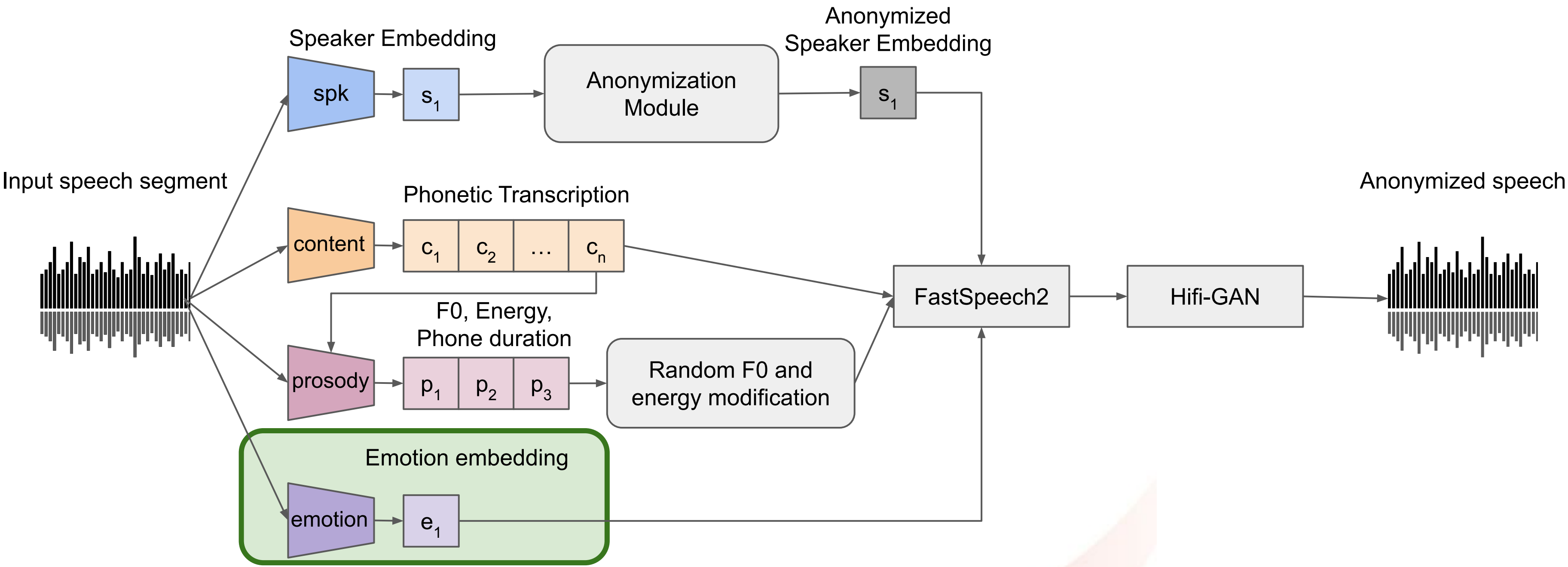
2. Literature Review
   a. Speaker Anonymization
   b. Disentanglement Learning

3. Disentanglement-based Approaches for Anonymization
   a. Problem1: Anonymization models do not preserve emotions
      i. Contribution1: NS3 FACodec, Emotion embeddings
   b. Problem2: Prosody Leakage
      i. Contribution2: Mean-reversion + Noise

4. **Future Work**
   a. **Future Directions**

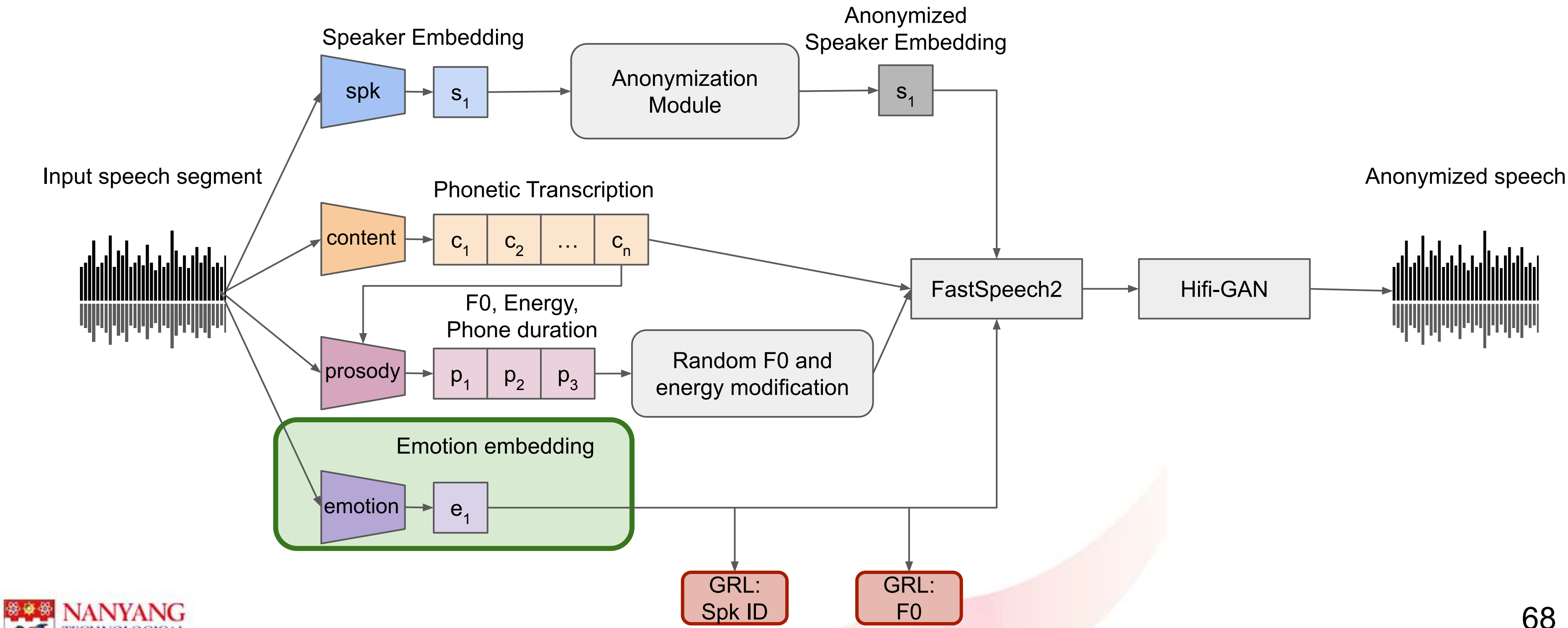
NANYANG TECHNOLOGICAL UNIVERSITY

5.1. ID leakage in Emotion Embedding



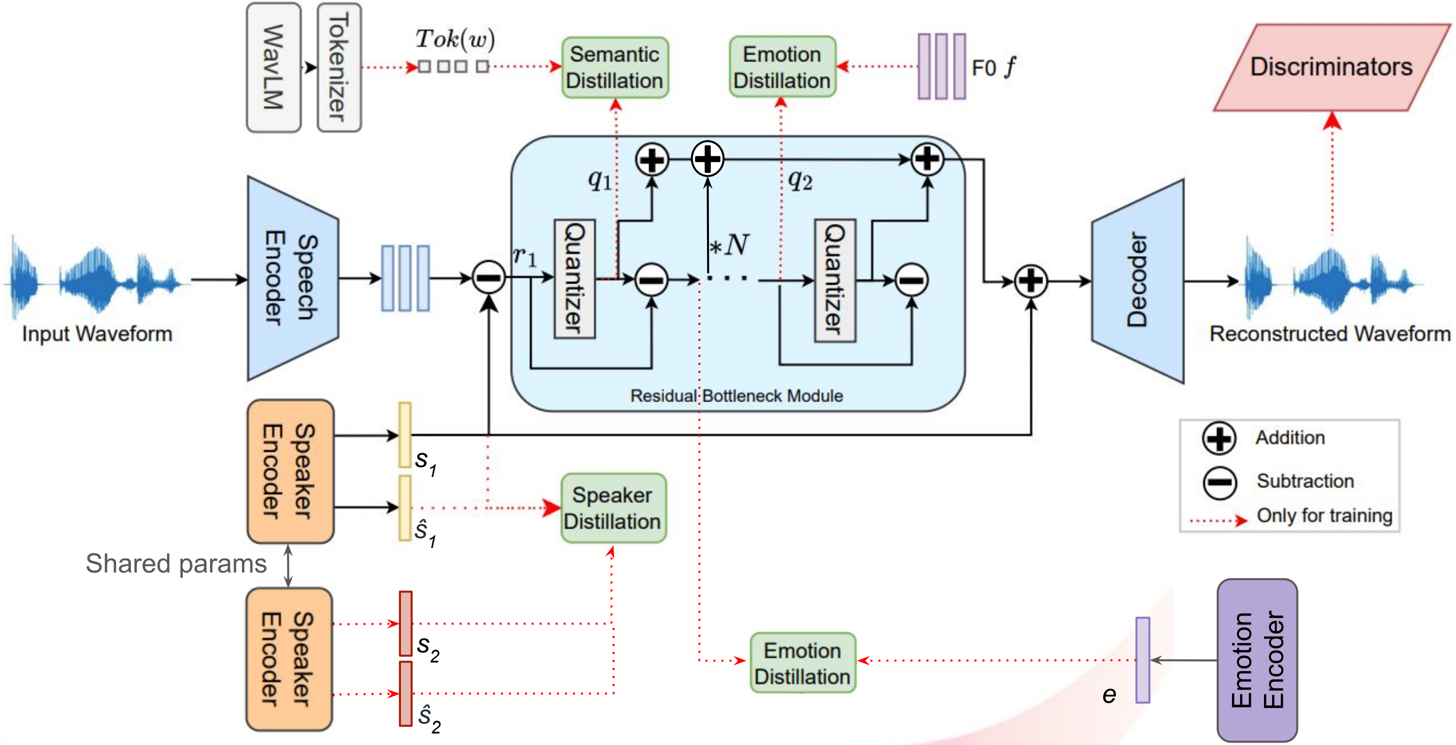
N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

5.1. ID leakage in Emotion Embedding

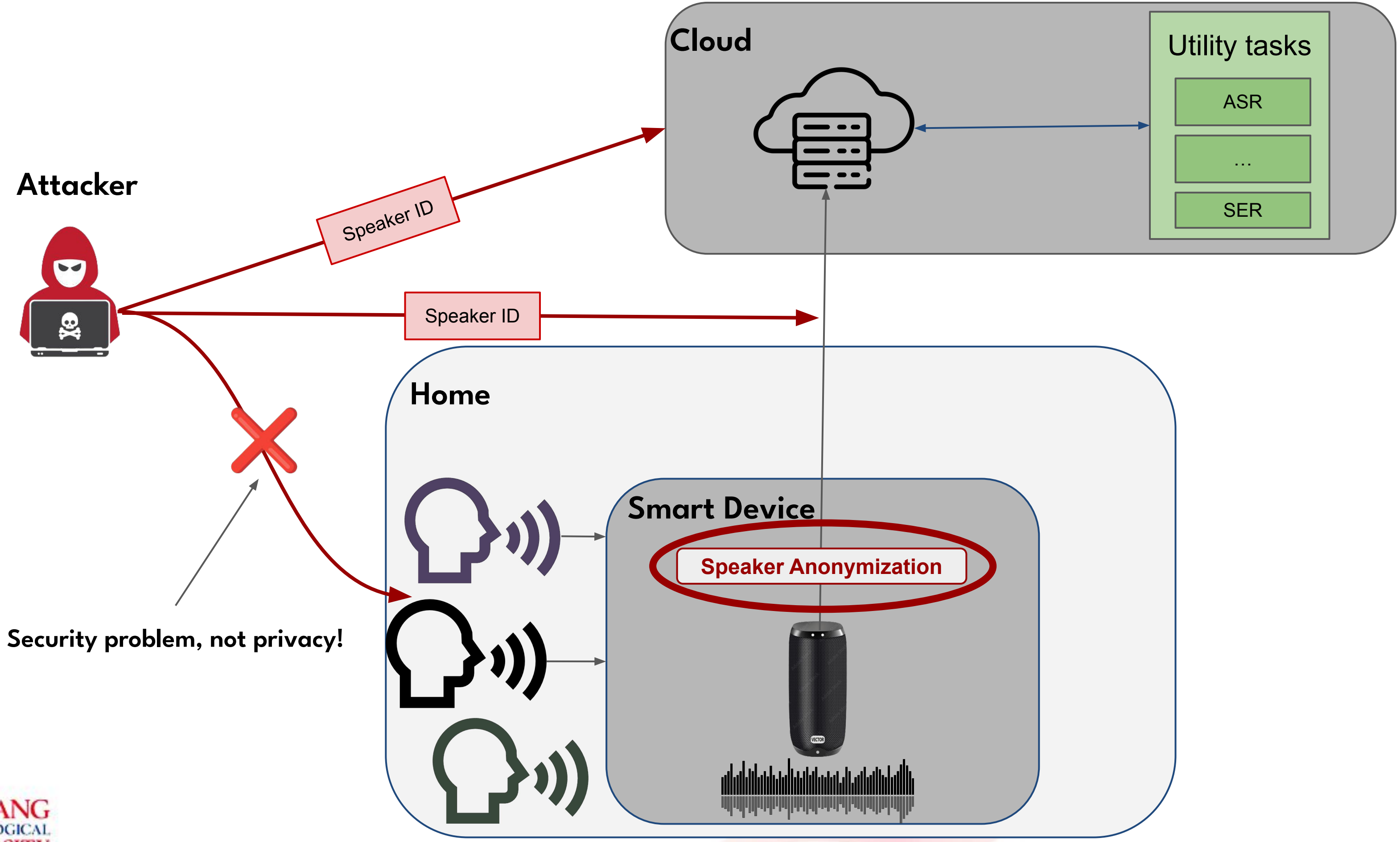


N. Kuzmin, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S., NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 2024

5.2. Codec models + Diffusion models

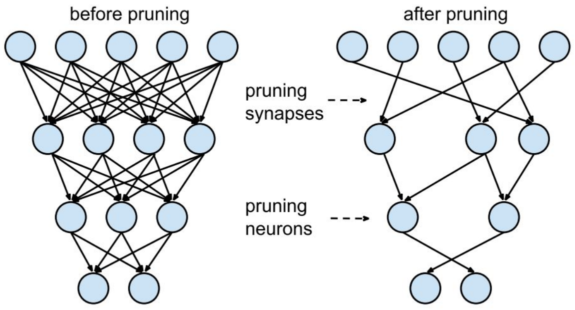


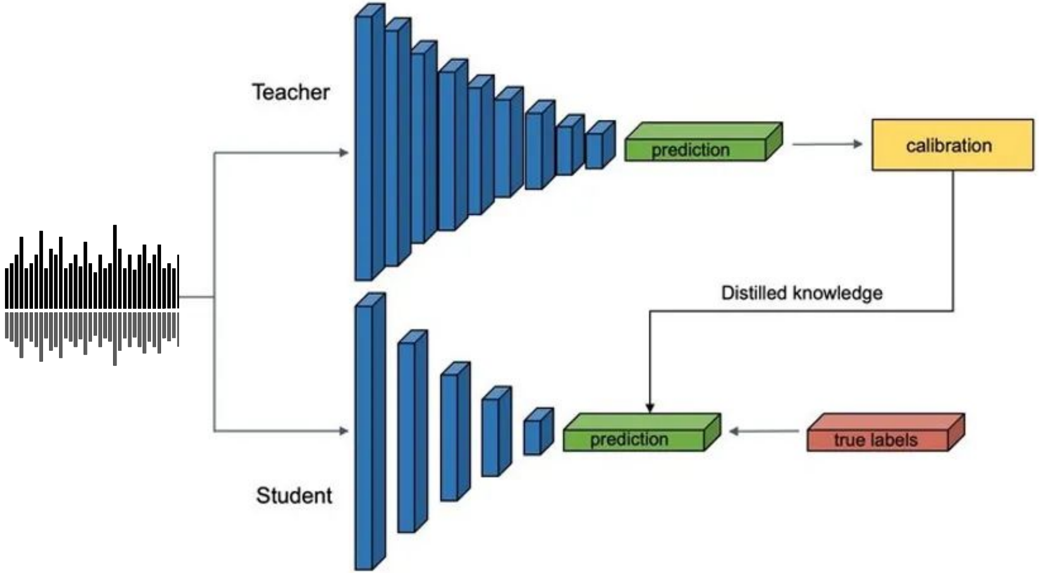
5.3. HouseHold Speaker Anonymization



N. Kuzmin*, A. Sholokhov*, K. A. Lee and E. S. Chng, ”Probabilistic Back-ends for Online Speaker Recognition and Clustering,” ICASSP, 2023

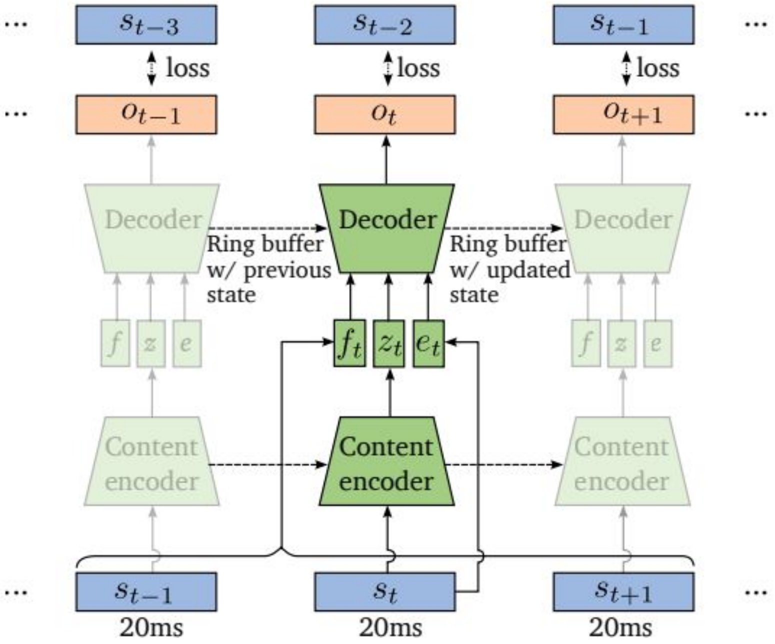
Pruning and Knowledge Distillation





Gongfan Fang et al ”DepGraph: Towards Any Structural Pruning”, CVPR, 2023
Jianping Gou et al “Knowledge Distillation: A Survey”, International Journal of Computer Vision, 2021

Streaming (online) speaker anonymization



Streaming inference.

Y. Yang et al "StreamVC: Real-Time Low-Latency Voice Conversion", ICASSP, 2024

Thank you for your attention!
Feel free to ask questions.

About me

1. Name: Nikita Kuzmin
2. Status:
   a. Matriculated on 08-Aug-2022
   b. 1, 2, 3 TAC appraisal passed
   c. All school requirement fulfilled for QE (GAP hours, TA courses)
3. CGPA: 4.67

4. Publications:
   a. **N. Kuzmin,** Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S. (2024) NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 72-79, doi: 10.21437/SPSC.2024-13
   b. **N. Kuzmin*,** A. Sholokhov*, K. A. Lee and E. S. Chng, ”Probabilistic Back-ends for Online Speaker Recognition and Clustering,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10097032.
   c. Yao, J., **Kuzmin, N.**, Wang, Q., Guo, P., Ning, Z., Guo, D., Lee, K.A., Chng, E.-S., Xie, L. (2024) NPU-NTU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 67-71, doi: 10.21437/SPSC.2024-12