NWPU-ASLP System for the VoicePrivacy 2022 Challenge

Jixun Yao, Qing Wang, Li Zhang, Pengcheng Guo, Yuhao Liang, Lei Xie

Audio, Speech and Language Processing Group (ASLP@NPU),
School of Computer Science, Northwestern Polytechnical University, Xi’an, China
http://www.npu-aslp.org

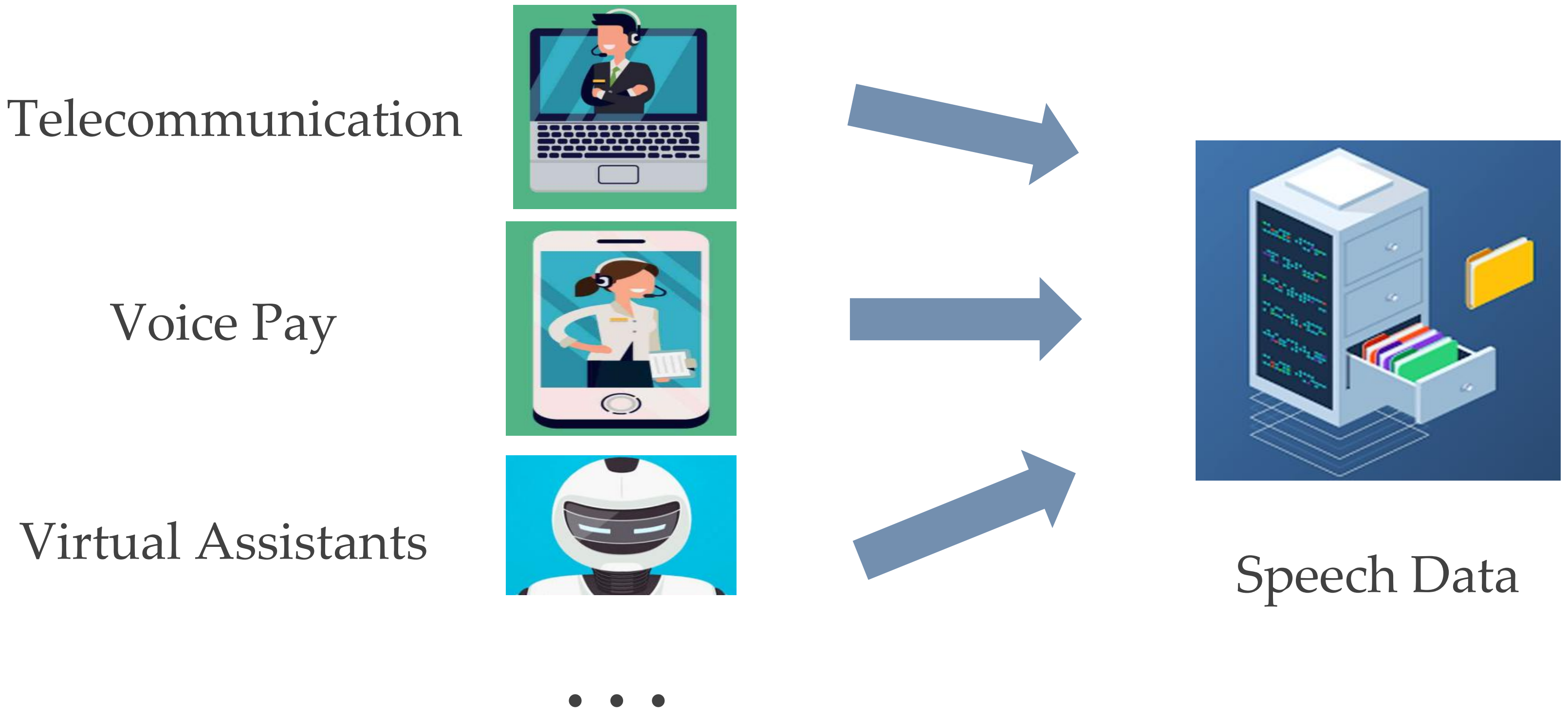
Content

Content

Introduction

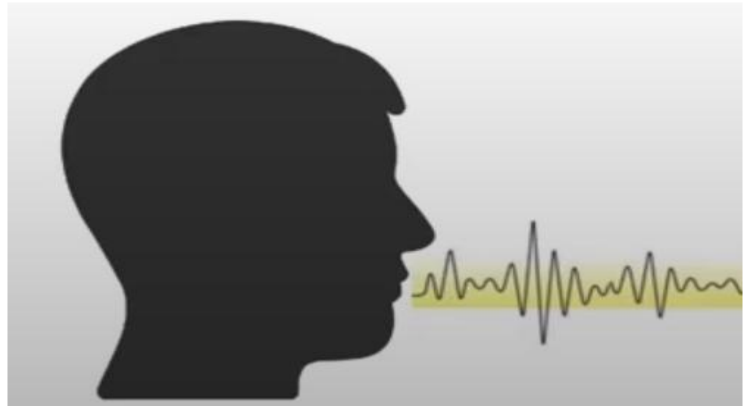
Speech Data

- Speech data are proliferating exponentially
- Applications record personal speech data which have rick to be stolen by attacker



Speech Information

- Speech data contain rich personal sensitive information

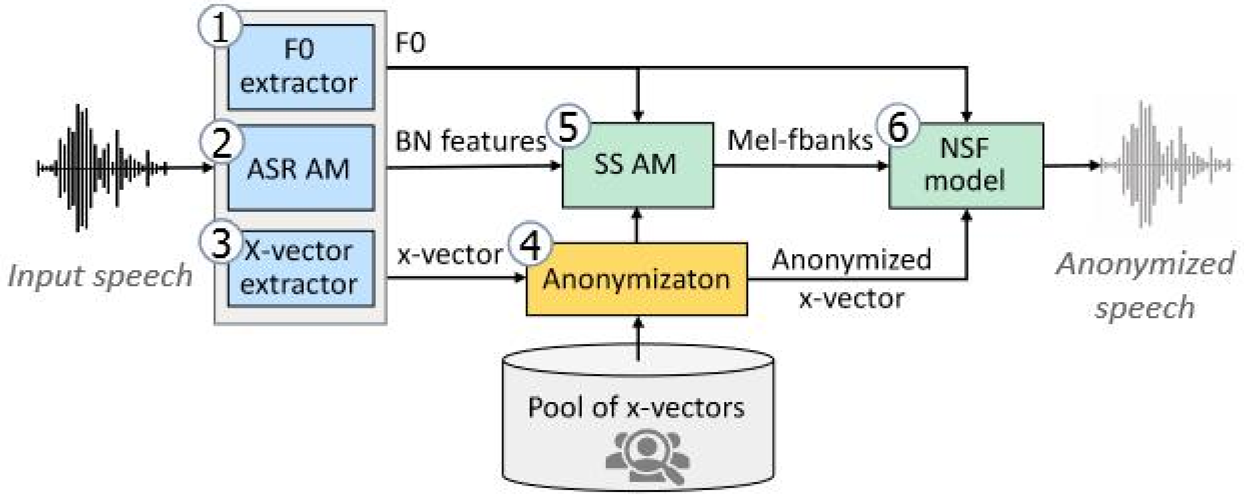


Age
Gender
Health State
Religious Beliefs
Other sensitive attributes

Introduction

- ## How to protect the personal speech data
  - Cryptography: may be hacked
  - Anonymization: the hiding of speaker identity

- ## VoicePrivacy challenge
  - Provide metrics, protocols, benchmarks and evaluation datasets



https://www.voiceprivacychallenge.org/

VoicePrivacy baseline anonymization system

Introduction

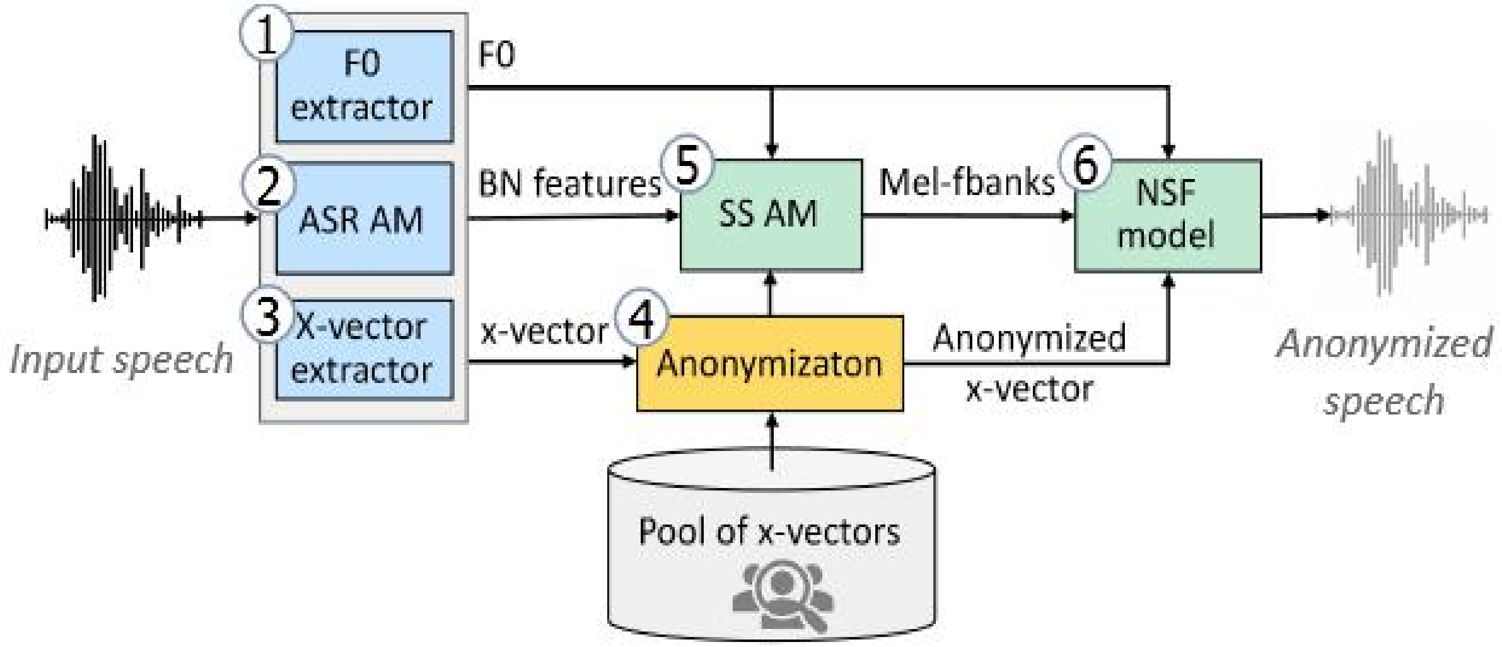
- ## Our method -- Different from the baseline system
  - Our system **DOES NOT** involve additional ASV models or an x-vectors pool
  - Also reduce the risk of insufficient generalization of the ASV model and the complexity of anonymization computation
- ## ASV-model-free approach for speaker anonymization
  - Look-up-table (LUT) for speakers in training set as speaker pool
  - Reserve a pseudo speaker ID in LUT to generate pseudo speaker embedding
  - Average the randomly selected speaker embeddings in LUT
  - Pseudo speaker embedding + Averaged embedding → anonymized embedding

Introduction

Compare with the baseline system

- Our work **focuses** on the **anonymization module**

	Baseline system	Our system
(1)	orign F0 extractor	YIN algorithm
(2)	Kaldi PPG extractor	WeNet tools
(3)	X-vector	look-up table + speaker encoder
(4)	average candidate speaker vectors	**combine two types of speaker embedding**
(5)	SS AM	CBHGAR
(6)	NFS model	our modified version of HifiGAN



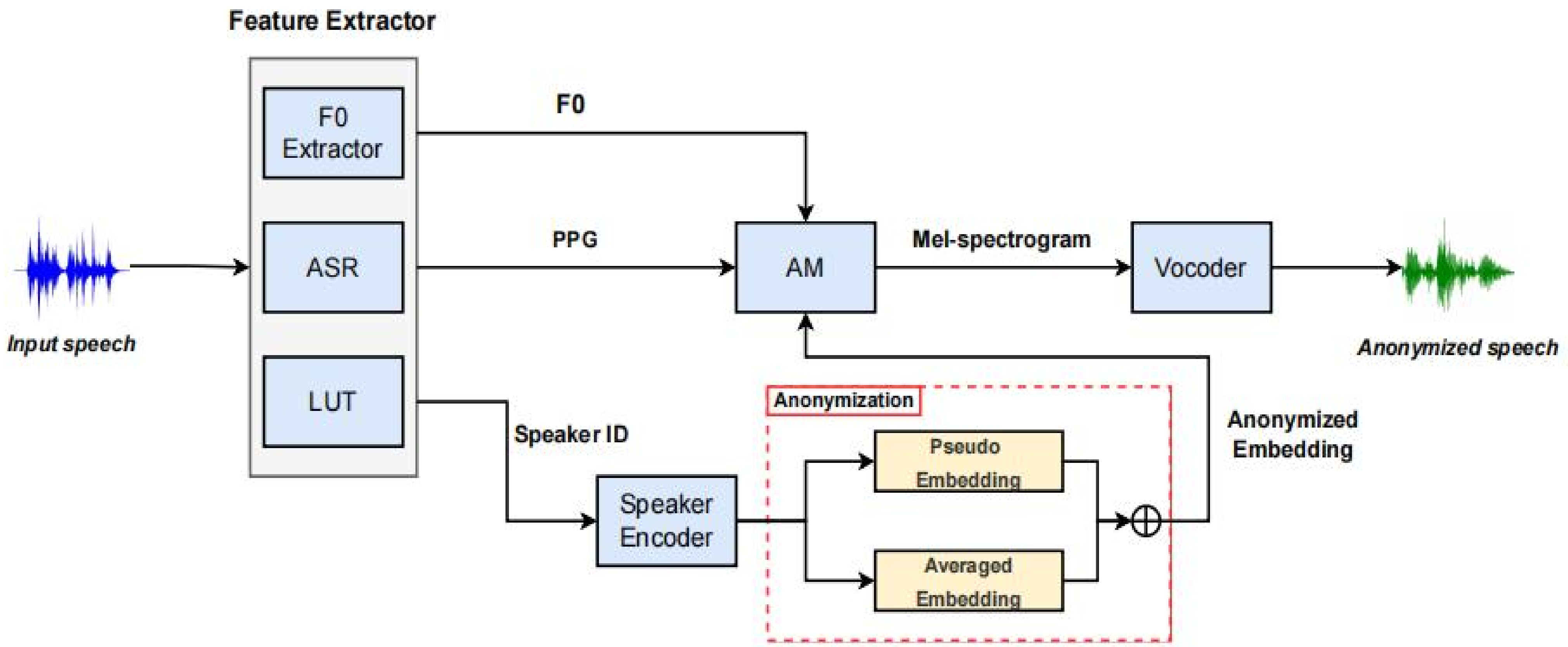
VoicePrivacy baseline anonymization system

Content

Proposed method

System overview

- Our anonymization system consists of four modules:
  - (a) Feature extractor
  - (b) Acoustic model (AM)
  - (c) Anonymization module
  - (d) Vocoder
- Anonymization process in three steps
  - Extract F0, PPG and Speaker ID
  - Predict anonymized mel-spectrogram
  - Reconstruct mel-spectrogram to anonymized speech



Proposed method

Anonymization strategy

Averaged embedding

- Average the randomly selected K speaker embeddings

Pseudo embedding

- generated by pseudo Speaker ID

spkid	spk name
1	Speaker 1
2	Speaker 2
…	…
1407	Speaker 1407
1408	**Pseudo Speaker**

Look-Up Table (LUT)

random select
K speakers

averaged embedding

pseudo embedding

Anonymized embedding

- weighted concatenation with averaged embedding and pseudo embedding
- anonymized embedding = α * averaged embedding $\oplus$ β * pseudo embedding

  α and β are hyperparameters

Content

Evaluations

Dataset

- Our proposed system follows VoicePrivacy 2022 Challenge data configuration
- LibriSpeech-test-clean and VCTK-test for ASR and ASV evaluation tasks

Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
	VCTK-dev	Enrollment	15	15	30	600
		Trial (different)				10,677
		Trial (common)				695
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496
	VCTK-test	Enrollment	15	15	30	600
		Trial (different)				10,748
		Trial (common)				700

Number of speakers and utterances in the development and evaluation sets

Results

Primary EER results

- Our approach leads to a notable increase in average EER of up to 18.34% compared with B1.a and 20.97% compared with B1.b
- Different genders perform similarly in our approach, as compared with significant difference in baseline systems

Table 2: *Privacy results on different conditions. EER achieved by ASV_{eval}^{anon} on data processed by our anonymization method vs. EER achieved by baseline B1.a or B1.b and original(Orig).*

Dataset	Gender	Weight	Orig	B1.a	B1.b	Condition1	Condition2	Condition3	Condition4
LibriSpeech-dev	female	0.25	8.67	17.76	19.03	13.92	21.02	25.28	26.28
	male	0.25	1.24	6.37	5.59	15.53	19.57	22.05	23.45
VCTK-dev (diff)	female	0.20	2.86	12.46	8.25	18.36	29.14	38.80	40.31
	male	0.20	1.44	9.33	6.01	22.28	31.46	36.92	37.77
VCTK-dev (comm)	female	0.05	2.62	13.95	9.01	19.19	26.45	34.59	35.76
	male	0.05	1.43	13.11	9.40	21.37	29.91	37.04	37.89
Weighted average dev			3.54	11.74	9.93	17.51	25.08	30.55	31.73
LibriSpeech-test	female	0.25	7.66	12.04	9.49	16.61	17.88	20.99	22.08
	male	0.25	1.11	8.91	7.80	10.69	14.03	17.37	19.15
VCTK-test (diff)	female	0.20	4.89	16.00	10.91	23.10	34.83	40.84	40.64
	male	0.20	2.07	10.05	7.52	23.19	30.20	37.54	38.81
VCTK-test (comm)	female	0.05	2.89	17.34	15.32	23.99	34.68	40.46	40.46
	male	0.05	1.13	9.89	8.19	23.16	32.20	38.14	38.70
Weighted average test			3.79	11.81	9.18	18.44	24.32	29.19	30.15

Results

Primary WER results

- Lowest WER
  - Librispeech-test: 3.84%
  - VCTK-test:7.81%
- Absolute WER Reduction
  - 2.47% over B1.a
  - 1.74% over B1.b
- We match all the 4 EER conditions with WER substantially lower than baseline

Table 3: *Primary utility evaluation: WER achieved by* ASR_{eval}^{anon} *on data processed by our anonymization method (with the large LM). C* denotes different target EER conditions*

Dataset	Orig	B1.a	B1.b	C1	C2	C3	C4
LibriSpeech-dev	3.82	4.34	4.19	3.91	3.71	3.65	3.65
VCTK-dev	10.79	11.54	10.98	8.10	7.73	7.68	7.62
Average dev	7.31	7.94	7.59	6.00	5.72	5.66	5.63
LibriSpeech-test	4.15	4.75	4.43	3.96	3.98	3.84	3.87
VCTK-test	12.82	11.82	10.69	8.37	7.85	7.81	7.85
Average test	8.49	8.29	7.56	6.16	5.91	5.82	5.86

Results

Secondary primary evaluation

- The highest pitch correlation is achieved in c1 of 0.7 and exceeds the minimum threshold
- But the voice distinctiveness gets worse as the EER rises

Table 4: *Secondary utility evaluation: pitch correlation $\rho F0$ achieved on data processed by B1.a, B1.b and our anonymized results.*

Dataset	Gender	B1.a	B1.b	C1	C2	C3	C4
LibriSpeech-dev	female	0.77	0.84	0.70	0.71	0.71	0.71
	male	0.73	0.76	0.69	0.69	0.69	0.69
VCTK-dev (dif)	female	0.84	0.87	0.76	0.76	0.77	0.76
	male	0.78	0.76	0.71	0.71	0.71	0.71
VCTK-dev (com)	female	0.79	0.84	0.71	0.71	0.72	0.71
	male	0.72	0.72	0.67	0.67	0.67	0.67
Weighted average dev		0.77	0.80	0.71	0.71	0.72	0.72
LibriSpeech-test	female	0.77	0.85	0.71	0.72	0.72	0.72
	male	0.69	0.72	0.64	0.64	0.64	0.64
VCTK-test (dif)	female	0.87	0.87	0.77	0.76	0.77	0.77
	male	0.79	0.77	0.71	0.71	0.71	0.71
VCTK-test (com)	female	0.79	0.85	0.72	0.71	0.72	0.72
	male	0.70	0.71	0.64	0.65	0.65	0.65
Weighted average test		0.77	0.80	0.70	0.70	0.71	0.70

Table 5: *Secondary utility evaluation: gain of voice distinctiveness G_{VD} achieved on data processed by B1.a, B1.b and our anonymized results.*

Dataset	Gender	B1.a	B1.b	C1	C2	C3	C4
LibriSpeech-dev	female	-9.15	-4.92	-2.94	-10.50	-17.47	-21.35
	male	-8.94	-6.38	-2.69	-9.18	-15.78	-18.66
VCTK-dev (dif)	female	-8.82	-5.94	-2.38	-8.33	-12.19	-13.96
	male	-12.61	-9.38	-3.10	-10.68	-17.25	-20.72
VCTK-dev (com)	female	-7.56	-4.17	-1.98	-6.72	-13.33	-17.18
	male	-10.37	-6.99	-2.06	-7.70	-14.81	-19.71
Weighted average dev		-9.71	-6.44	-2.71	-9.442	-15.61	-18.86
LibriSpeech-test	female	-10.04	-5.00	-2.72	-9.21	-16.44	-20.13
	male	-9.01	-6.64	-1.64	-7.36	-13.90	-17.83
VCTK-test (dif)	female	-10.29	-6.09	-2.82	-9.18	-15.41	-17.86
	male	-11.69	-8.64	-3.85	-10.77	-15.82	-17.65
VCTK-test (com)	female	-9.31	-5.10	-2.15	-8.12	-15.55	-20.39
	male	-10.43	-6.50	-2.68	-9.43	-16.78	-21.26
Weighted average test		-10.15	-6.44	-2.67	-9.012	-15.46	-18.69

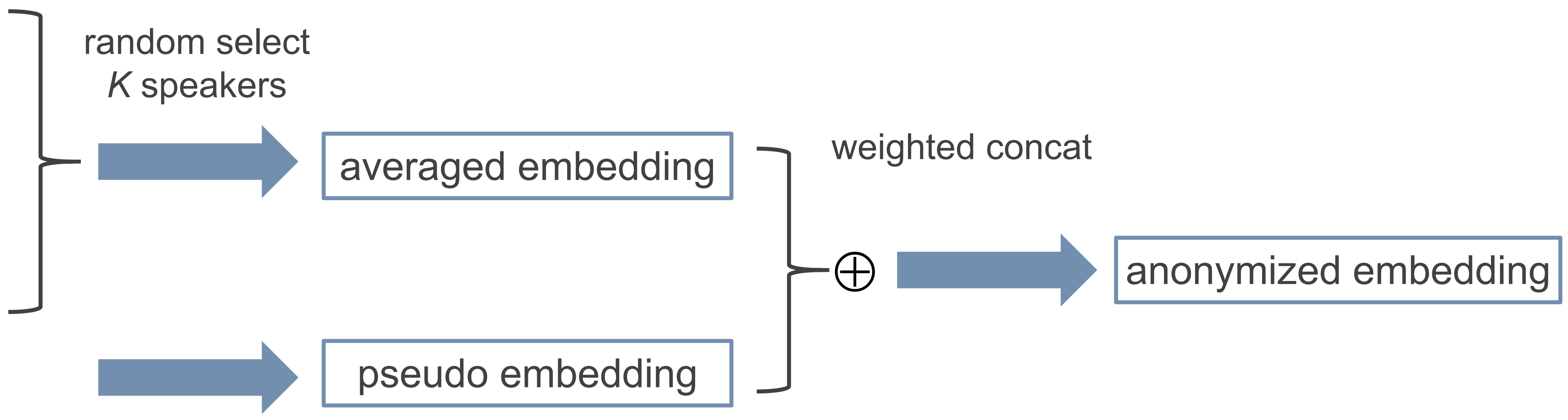
Content

Conclusions

- NWPU-ASLP anonymization system for VoicePrivacy2022 Challenge
  - Highlight: an ASV-model-free anonymization strategy

spkid	spk name
1	Speaker 1
2	Speaker 2
...	...
1407	Speaker 1407
1408	**Pseudo Speaker**

Look-Up Table (LUT)



  - Our anonymization strategy can meet all conditions by adjusting the weight h-param
    - WER substantially lower than baseline

Jixun Yao

Audio, Speech & Language Processing Group (ASLP@NPU)

www.npu-aslp.org

Thank You!