JHU HLTCOE Submission to the Voice Privacy Challenge 2024

Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola García-Perera, Sanjeev Khudanpur, Nicholas Andrews, Matthew Wiesner

JOHNS HOPKINS UNIVERSITY

human language technology
center of excellence

Baselines

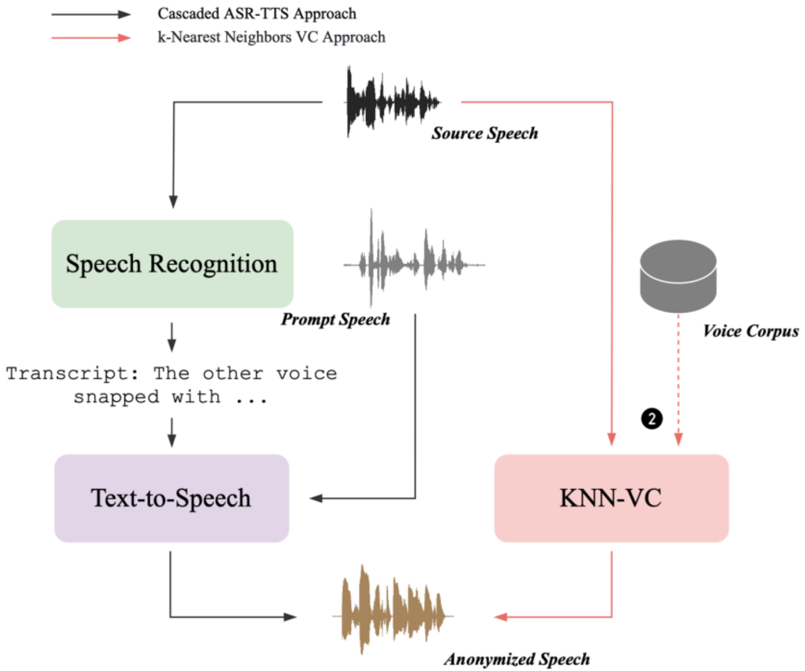
- Strong performance in the privacy objective (EER) necessitates the removal of acoustic characteristics, like duration, speaking style, from source speech. (STTTS, ASRBN, NAC)
- while strong performance in the utility objective (emotion preservation, UAR) requires more acoustic characteristics from the source utterance (McAdams)

Baselines

- Strong performance in the privacy objective (EER) necessitates the removal of acoustic characteristics, like duration, speaking style, from source speech. (STTTS, ASRBN, NAC)
- while strong performance in the utility objective (emotion preservation, UAR) requires more acoustic characteristics from the source utterance (McAdams)

Our Approaches

- kNN-Voice Conversion (8.0% EER, 56.7% UAR)
  - Good at preserving Emotion
- Cascading ASR and TTS (48.4% EER, 30.4% UAR)
  - Good at concealing speaker identity

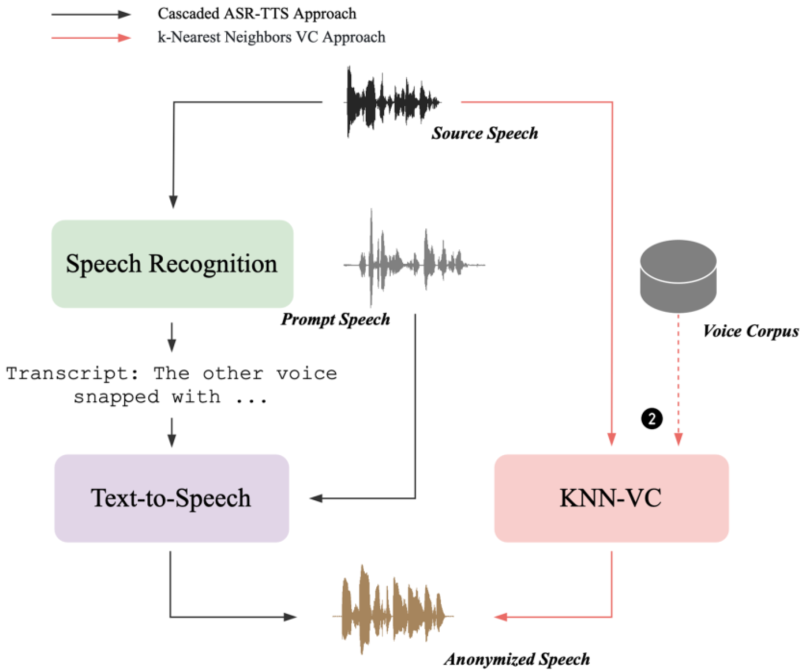


Baselines

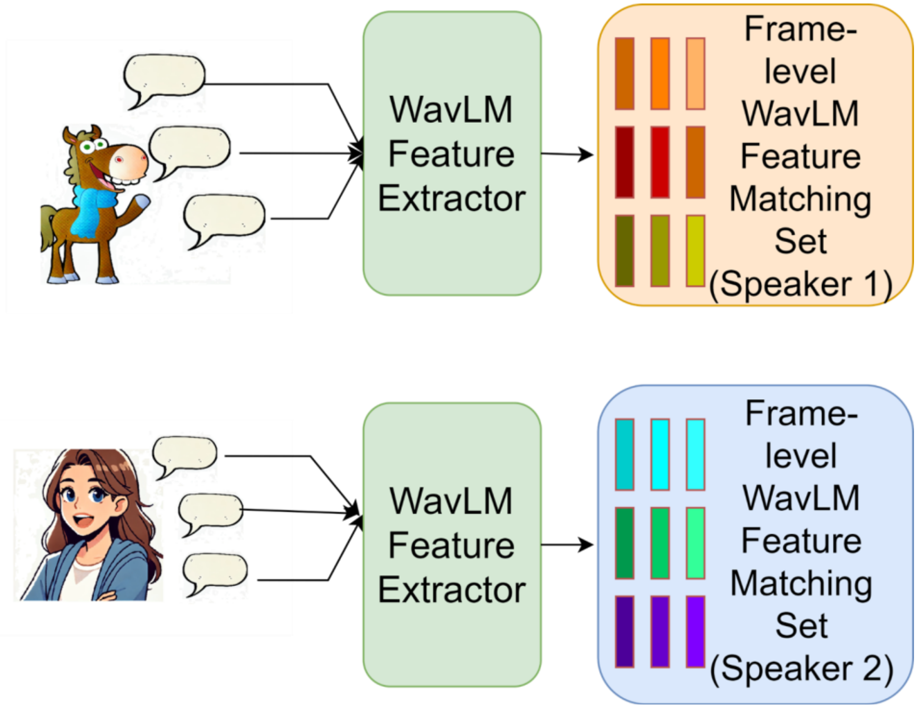
- Strong performance in the privacy objective (EER) necessitates the removal of acoustic characteristics, like duration, speaking style, from source speech. (STTTS, ASRBN, NAC)
- while strong performance in the utility objective (emotion preservation, UAR) requires more acoustic characteristics from the source utterance (McAdams)

Our Approaches

- kNN-Voice Conversion (8.0% EER, 56.7% UAR)
    - Good at preserving Emotion
- Cascading ASR and TTS (48.4% EER, 30.4% UAR)
    - Good at concealing speaker identity
- **Random Admixture (40.81% EER, 47.1% UAR)**
    - **Achieve the best of both worlds**

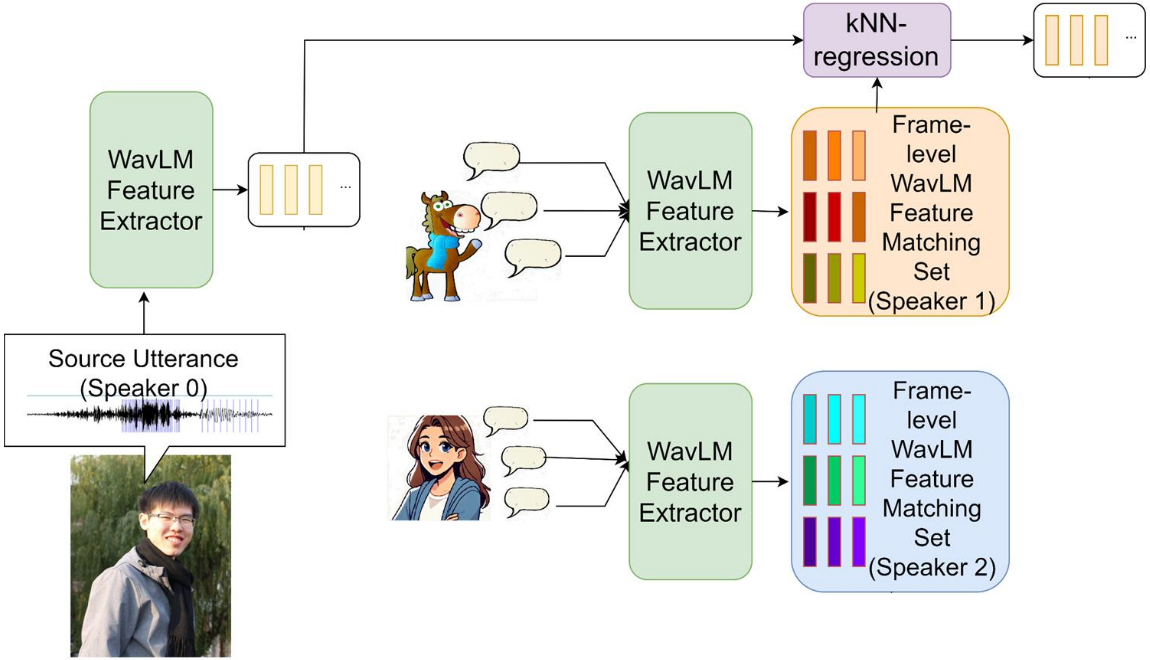


kNN-Voice Conversion [1]



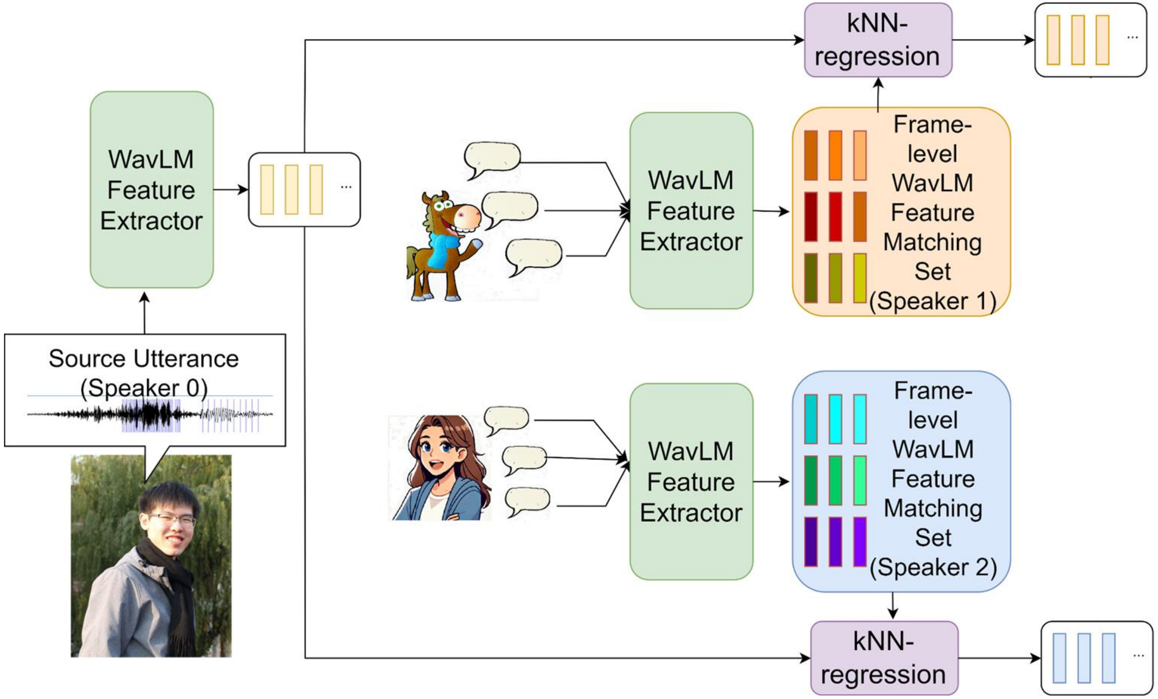
[1] Voice Conversion With Just Nearest Neighbors, Baas et al., 2023

kNN-Voice Conversion [1]



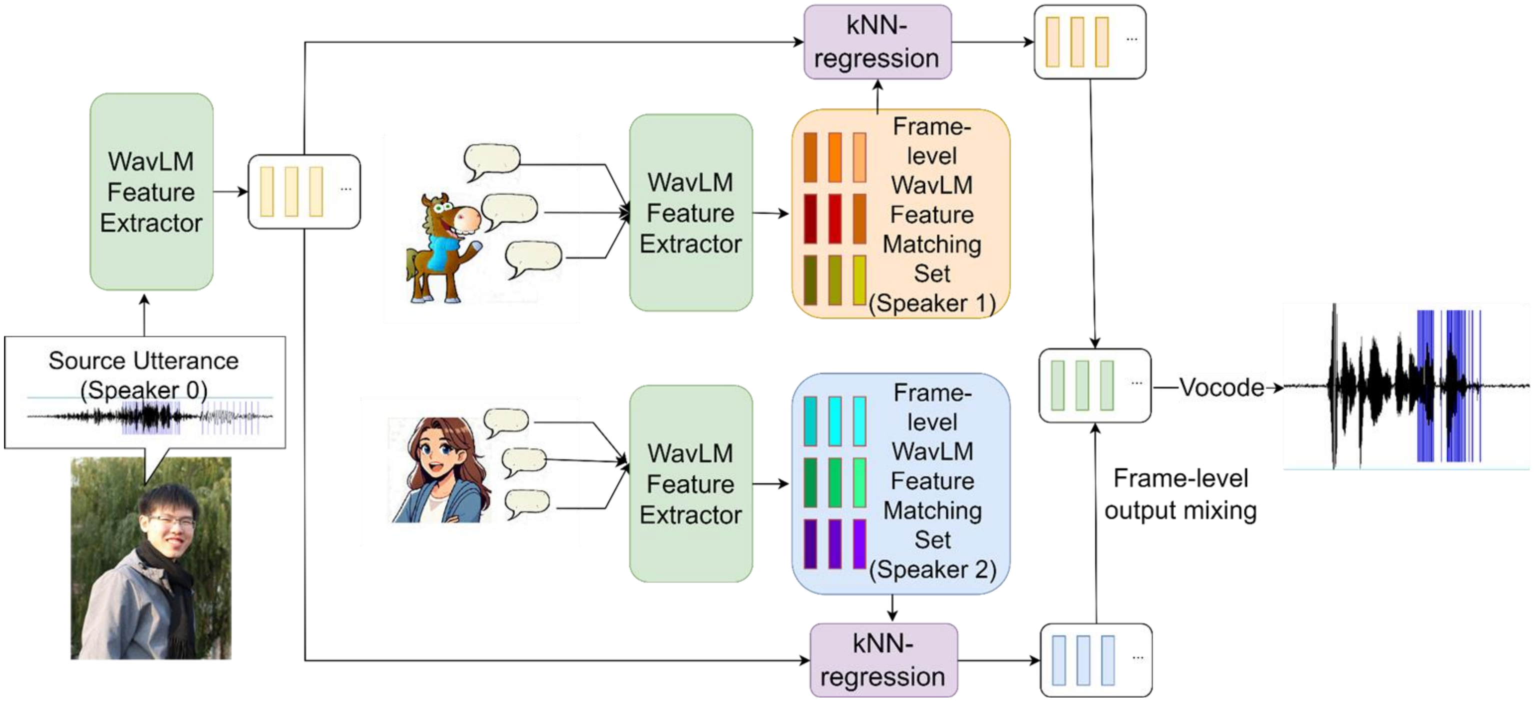
[1] Voice Conversion With Just Nearest Neighbors, Baas et al., 2023

kNN-Voice Conversion [1]



[1] Voice Conversion With Just Nearest Neighbors, Baas et al., 2023

kNN-Voice Conversion [1]



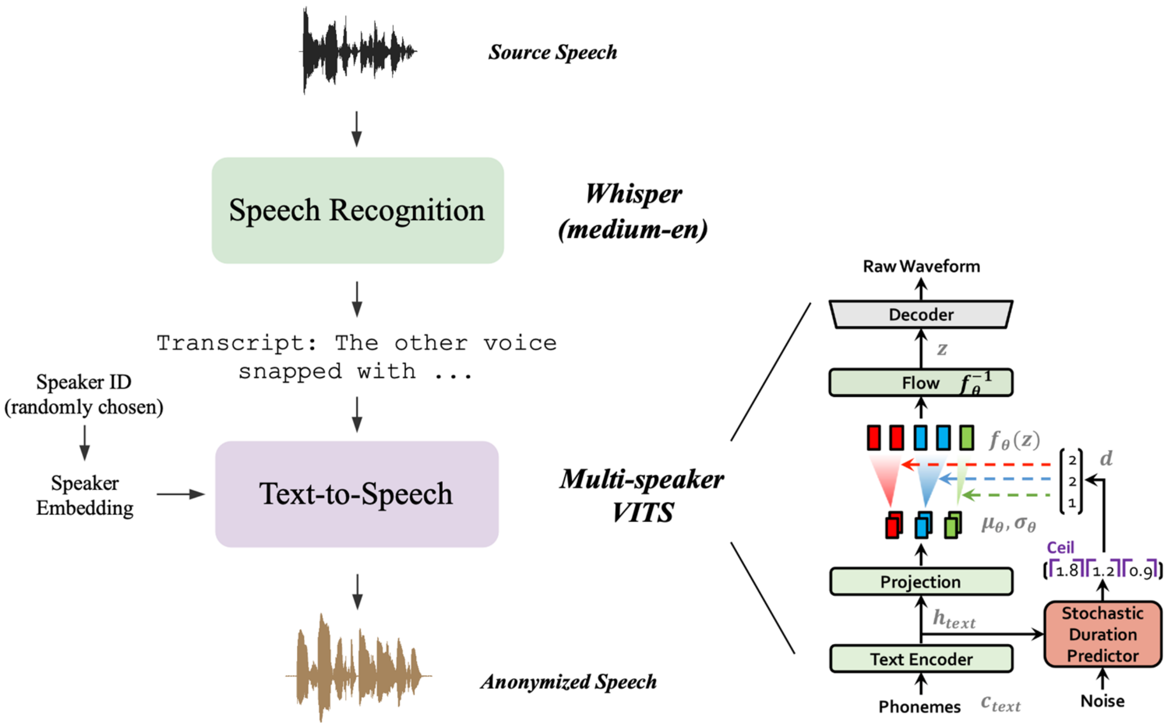
[1] Voice Conversion With Just Nearest Neighbors, Baas et al., 2023

Cascading ASR and TTS

- ASR: Whisper
- TTS: Multispeaker-VITS [2]

Anonymized speakers:

- Randomly selected from LibriTTS

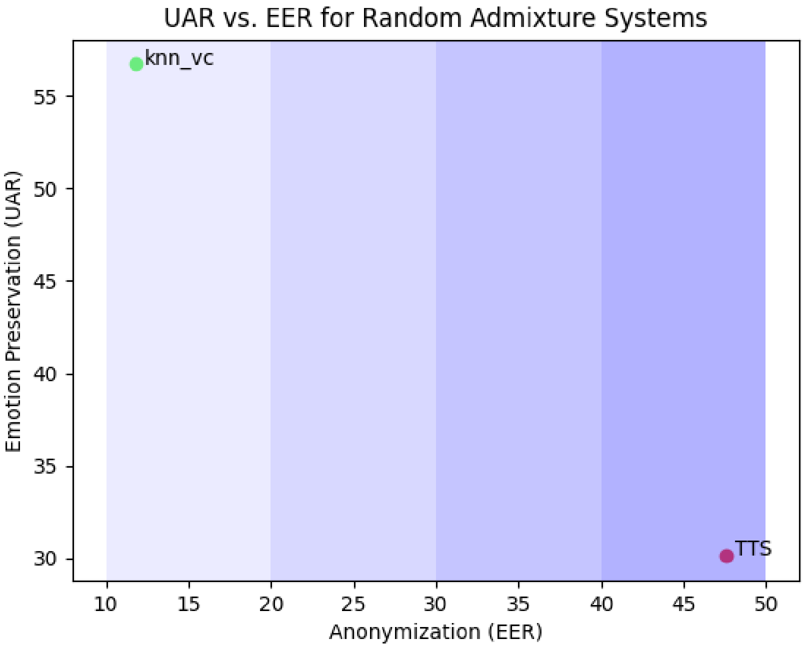


https://github.com/openai/whisper
https://huggingface.co/datasets/rhasspy/piper-checkpoints/blob/main/en/en_US/libritts_r/medium

[2] Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, J. Kim et al., 2021

Random Admixture - Getting the best of both worlds

- Created in response to the adversarial training setup for the Voice Privacy Challenge
- Inspired by data poisoning attacks, which demonstrate that a small amount of poisoned data can alter the decision boundary sufficiently that the model performance degrades significantly



Random Admixture - Getting the best of both worlds

- Created in response to the adversarial training setup for the Voice Privacy Challenge
- Inspired by data poisoning attacks, which demonstrate that a small amount of poisoned data can alter the decision boundary sufficiently that the model performance degrades significantly

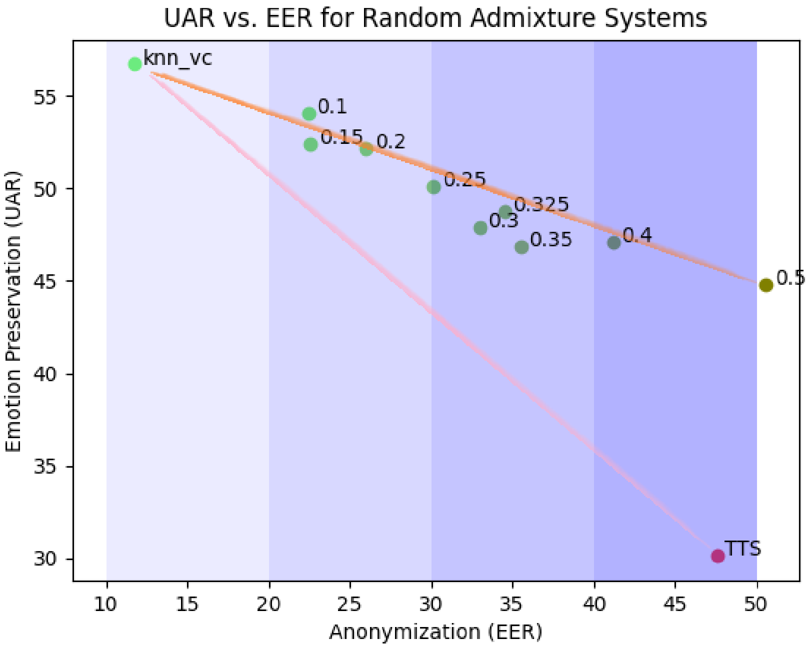


Table 1: *Privacy and Utility Performance of Various Anonymization Approaches (Darker Color Indicates Better Performance)*

ID	System	Privacy - EER (%) ↑					Utility - UAR mean (%) ↑			Utility - WER (%) ↓		
		libri-dev-f	libri-dev-m	libri-test-f	libri-test-m	avg.	IEMOCAP-dev	IEMOCAP-test	avg.	libri-dev	libri-test	avg.
0	origin	10.511	0.931	8.761	0.418	5.16	69.0796	71.0618	70.07	1.807	1.844	1.83
1*	kNN-VC	11.789	5.141	9.307	5.570	7.95	56.7330	56.6740	56.70	3.275	3.048	3.16
2	kNN-VC + len variation	11.192	5.125	10.218	5.793	8.08	56.9488	55.638	56.29	3.28	3.387	3.33
3	kNN-VC+ len var + noise-in	24.681	18.624	19.891	19.115	20.58	44.1260	42.3846	43.26	11.993	10.008	11.00
4*	whisper-VITS	47.584	49.233	47.445	48.750	48.25	30.1074	30.5932	30.35	3.743	3.755	3.75
1 + 4*	Admixture ($p = 0.2$)	26.003	16.155	20.776	24.722	21.91	51.2840	52.1324	51.71	3.300	3.290	3.31
1 + 4*	Admixture ($p = 0.325$)	34.518	32.918	34.532	33.676	33.91	49.3398	48.7304	49.04	3.514	3.336	3.43
1 + 4*	Admixture ($p = 0.4$)	41.192	40.660	42.182	39.225	40.81	47.0784	47.1046	47.09	3.454	3.199	3.33
5	WavLM Conv (base)	13.622	6.987	9.307	4.231	8.54	55.5458	53.9522	54.75	3.044	2.982	3.01
6	WavLM Conv + Adv Spk Loss	17.472	9.005	12.773	7.164	11.60	50.7706	50.4628	50.62	4.442	4.015	4.23
7	WavLM Conv + Discrete Loss	18.041	12.268	13.716	10.913	13.73	44.5292	42.5980	43.56	10.313	10.014	10.16
8	WavLM Conv + Adv + Discrete Loss	19.308	11.645	13.870	10.690	13.88	44.0936	42.9102	43.50	10.811	10.850	10.83

* marks submitted systems

Thanks and Questions