S3PRL-VC: Open-source Voice Conversion Framework with Self-supervised Speech Representations

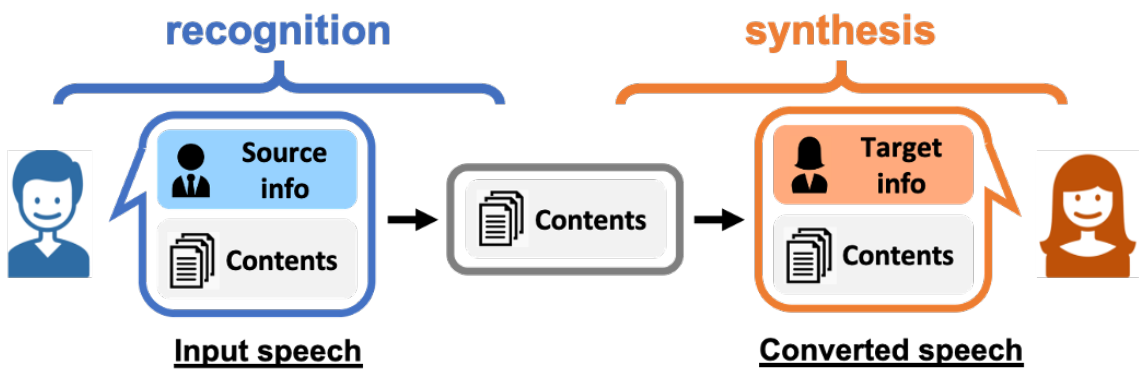
Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, Hung-Yi Lee, Shinji Watanabe, Tomoki Toda

Nagoya University, Japan
National Taiwan University, Taiwan
Carnegie Mellon University, USA

AAAI SAS 2022

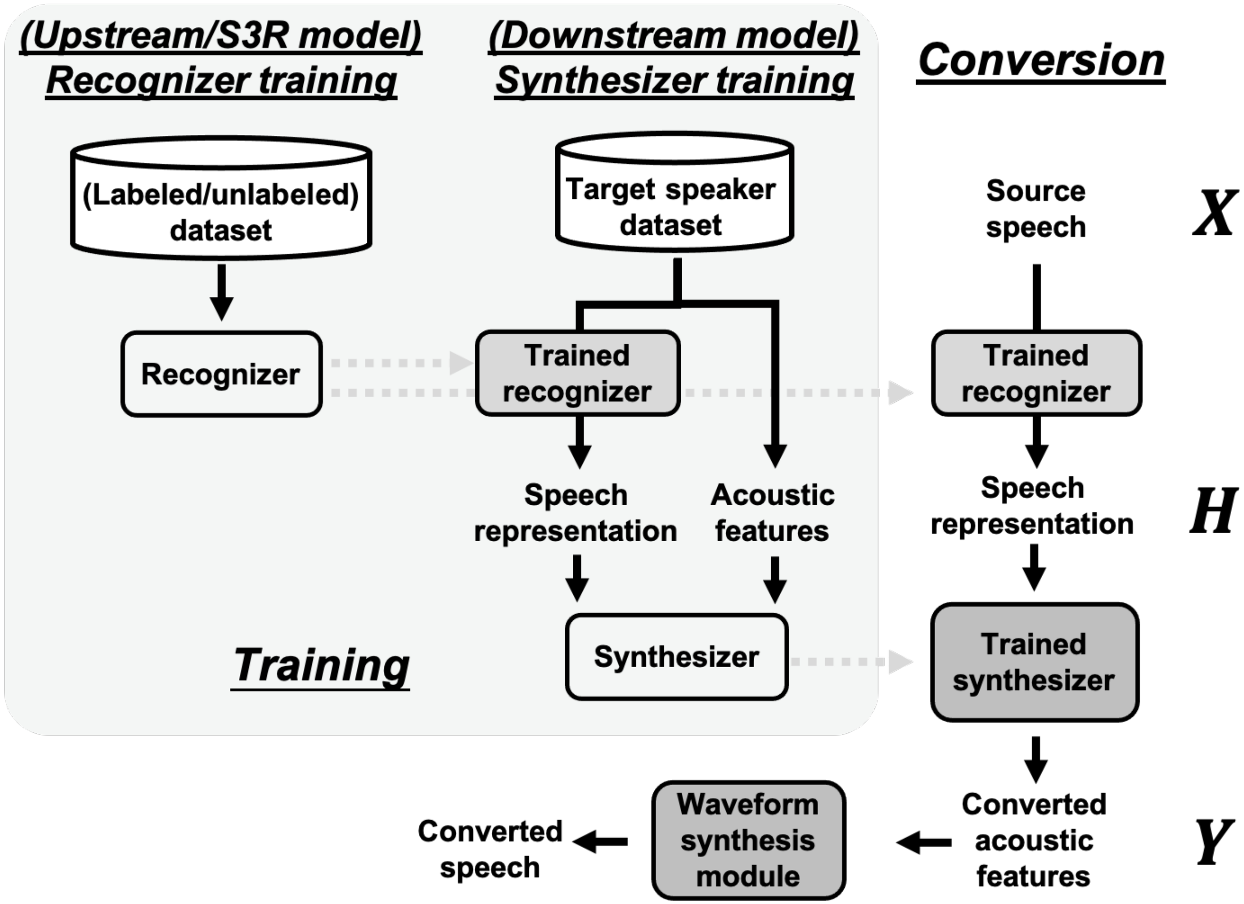
A trending paradigm in VC

- **Voice conversion** (VC)
  - A technique that converts one kind of speech to another, without changing the linguistic content.
- **Recognition-synthesis** (rec-syn) based VC
  - Information perspective: Converted = input – source + target



  - Ex. Can be realized by cascading an ASR & TTS model
  - ☺ State-of-the-art in voice conversion challenge (VCC) 2020

Training and conversion



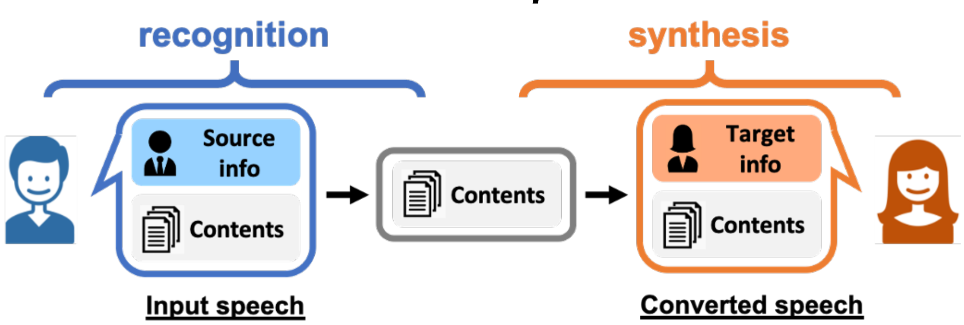
Intermediate representations

- *Supervised* speech representations
  - Ex. text, phonetic posteriorgram (PPG)
  - ☺ Accurate; ☹ Costly
- *Self-supervised* speech representations (S3Rs)
  - Learns rich, compact speech representations from large-scale **unlabeled** data.

Representation	Text	Phonetic Posteriorgram	Self-supervised speech representations
Extractor	ASR model		self-supervised model
Training data	labeled data		unlabeled data
Resolution	token level	frame level	

VC as a proxy task for S3R

- S3Rs have been studied systematically in **discriminative** tasks.
  - Speech recognition, speaker verification, etc.
- Unclear what S3Rs are optimal for **generation**.
- Hypothesis: a good S3R for VC should be
  1. Rich and compact in *content*;
  2. Contains little to none *speaker information*.



Contribution of this work

- **S3PRL-VC**: Extension of the S3PRL toolkit and SUPERB benchmark

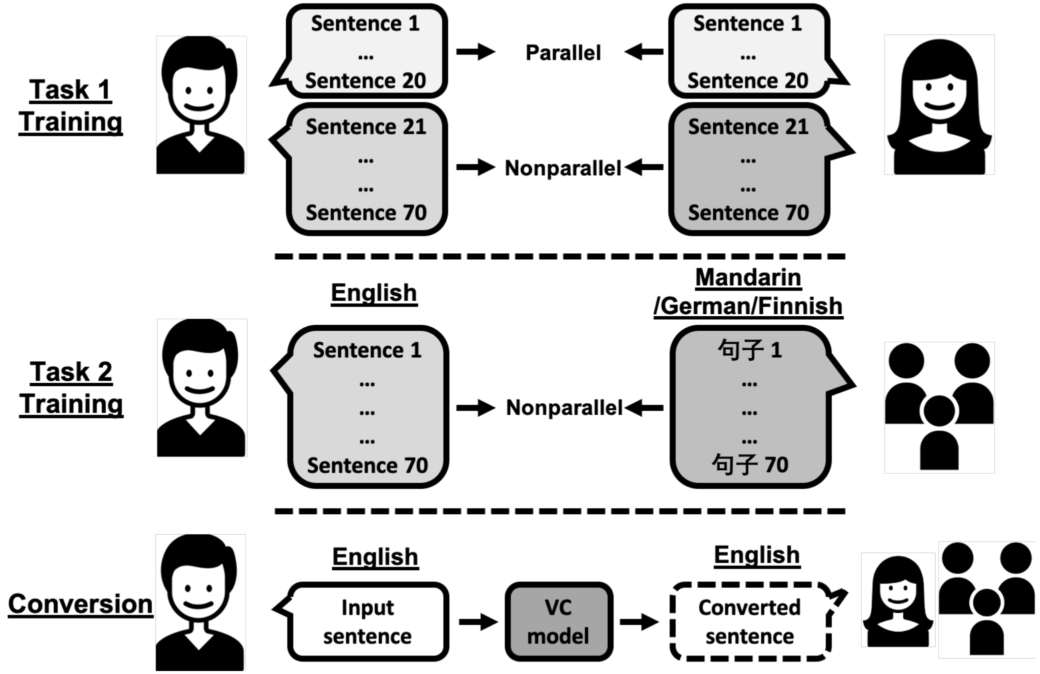
S3PRL
SPEECH TOOLKIT

SUPERB
Speech processing **U**niversal **PER**formance **B**enchmark

- Evaluate S3R-based VC in a unified setting:
  - Dataset – VCC2020
  - Tasks – any-to-one/any, intra-/cross-lingual
  - Implementation – synthesizer model, vocoder
  - Competing systems – top systems in VCC2020
  - Evaluation metrics – both objective and subjective

Dataset:
Voice conversion challenge (VCC) 2020

- Bi-annual event to compare SOTA techniques.
- VCC2020 has two tasks:
  - Task 1: intra-lingual VC; task 2: cross-lingual VC



Tasks

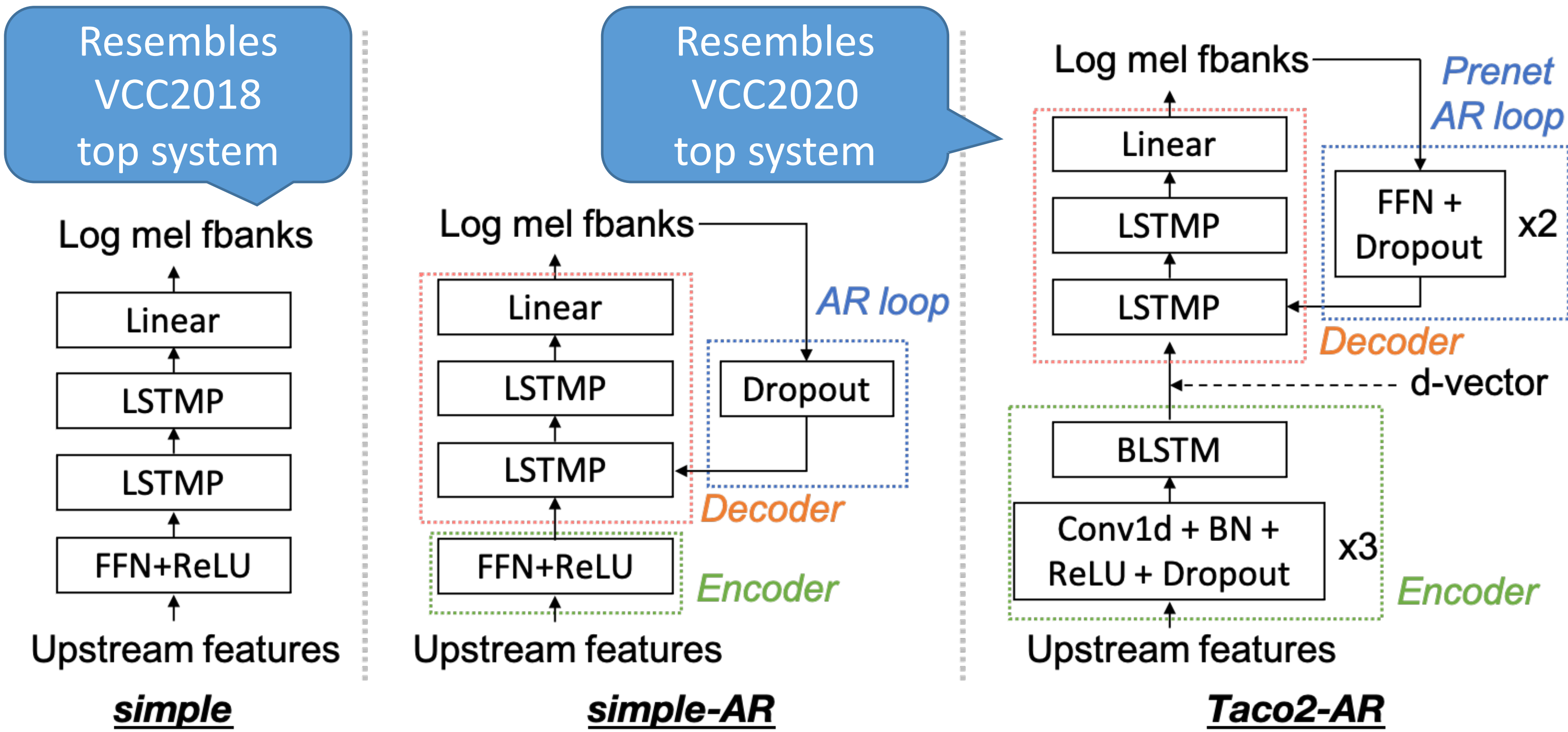
- Any-to-one (A2O) VC
  - Unseen speaker → **seen** speaker
  - Consider intra-lingual and cross-lingual.

$$\boxed{\mathbf{Y} = \mathrm{Synth}(\mathbf{H}), \mathbf{H} = \mathrm{Recog}(\mathbf{X})}$$

- Any-to-any (A2A) VC (a.k.a. zero-shot VC)
  - Unseen speaker → **unseen** speaker
    - Unseen: data is limited (less than 1 min)
  - Only consider the intra-lingual setting.
  - Speaker info injected with pretrained d-vector.

$$\boxed{\mathbf{Y} = \mathrm{Synth}(\mathbf{H}, \mathbf{s}), \mathbf{H} = \mathrm{Recog}(\mathbf{X}), \mathbf{s} = \mathrm{SpkEnc}(\mathbf{D}_{\mathrm{trg}}).}$$

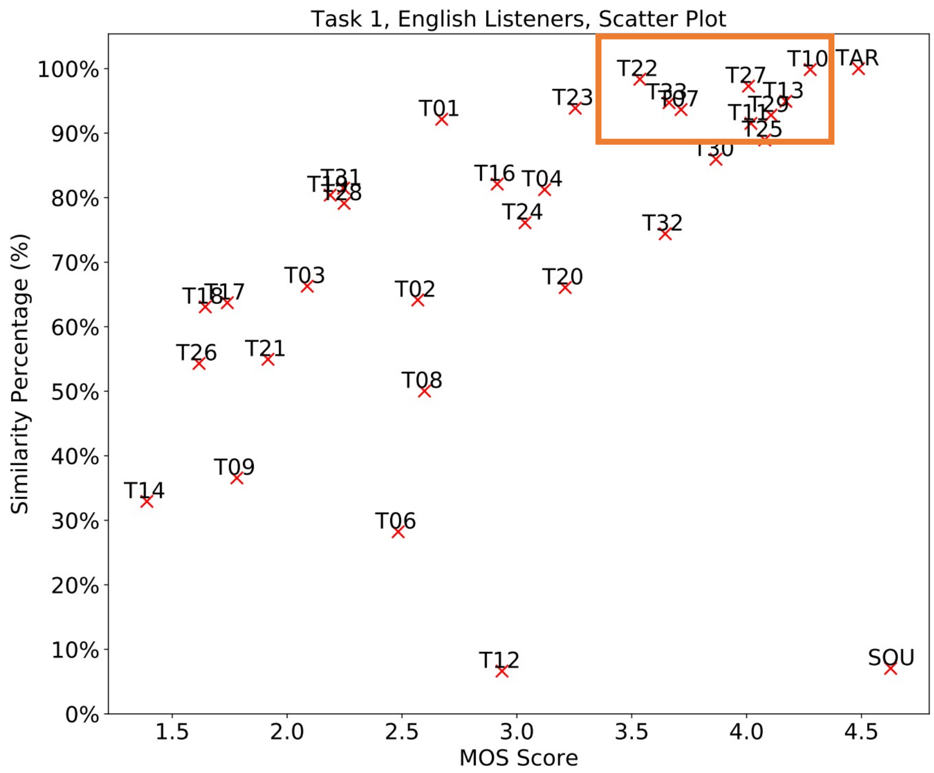
Implementation:
Synthesizer models, vocoder



Vocoder: Hifi-GAN (non-AR vocoder) trained on VCC2020+VCTK (A2O) / VCTK (A2A)

Competing systems

- A collection of SOTA rec-syn based VC systems
  - VCC2020 top systems
    - PPGs
    - pretraining on a multi-speaker dataset
    - AR vocoders (WaveNet)
  - VCC2020 baseline
    - Cascade ASR+TTS (text as representation)
    - Non-AR vocoder (Parallel WaveGAN)



Evaluation metrics

- Objective
  - Mel cepstrum distortion (**MCD↓**): L2-norm based, commonly used in VC.
  - Word error rate (**WER↓**): intelligibility measure from a pretrained wav2vec 2.0 model.
  - Accept rate from **ASV↑**: cosine similarity between d-vectors extracted from converted and ground truth.
- Subjective
  - **Naturalness↑**: mean opinion score (MOS) from 1-5
  - **Similarity↑**: judge whether converted and ground truth are spoken by the same speaker.

Results (1): Comparison of different models

Upstream	Intra-lingual A2O								
	Simple			Simple-AR			Taco2-AR		
	MCD	WER	ASV	MCD	WER	ASV	MCD	WER	ASV
mel	8.41	48.5	59.00	8.92	22.7	49.75	8.47	38.3	77.25
PPG (TIMIT)	7.78	69.0	85.50	7.83	58.9	95.25	7.18	33.6	99.75
PASE+	9.29	5.0	26.75	9.52	5.7	26.00	8.66	30.6	63.20
APC	8.67	8.6	48.00	8.73	7.1	41.75	8.05	27.2	87.25
VQ-APC	8.12	10.8	81.25	8.37	7.4	60.50	7.84	22.4	94.25
NPC	7.74	39.0	92.75	8.15	21.1	76.75	7.86	30.4	94.75
Mockingjay	8.58	31.3	51.00	8.74	9.5	47.00	8.29	35.1	79.75
TERA	8.60	11.4	46.50	8.67	6.0	42.50	8.21	25.1	83.75
Modified CPC	8.71	9.4	40.00	8.87	7.0	30.00	8.41	26.2	71.00
DeCoAR 2.0	8.31	7.4	54.75	8.33	6.4	53.00	7.83	17.1	90.75
wav2vec	7.45	14.0	**95.50**	7.64	4.9	90.50	7.45	10.1	98.25
vq-wav2vec	**7.41**	13.4	91.00	**7.24**	11.6	**98.75**	**7.08**	13.4	**100.00**
wav2vec 2.0 Base	7.80	24.7	92.75	7.77	5.0	86.50	7.50	10.5	98.00
wav2vec 2.0 Large	7.64	12.5	81.75	7.67	9.0	82.75	7.63	15.8	97.25
HuBERT Base	7.70	**5.5**	89.25	7.79	**4.7**	84.25	7.47	**8.0**	98.50
HuBERT Large	7.54	5.6	95.00	7.54	5.6	93.00	7.22	9.0	99.25

Simple → Simple-AR: large improvements in WER

Results (1): Comparison of different models

Upstream	Intra-lingual A2O								
	Simple			Simple-AR			Taco2-AR		
	MCD	WER	ASV	MCD	WER	ASV	MCD	WER	ASV
mel	8.41	48.5	59.00	8.92	22.7	49.75	8.47	38.3	77.25
PPG (TIMIT)	7.78	69.0	85.50	7.83	58.9	95.25	7.18	33.6	99.75
PASE+	9.29	5.0	26.75	9.52	5.7	26.00	8.66	30.6	63.20
APC	8.67	8.6	48.00	8.73	7.1	41.75	8.05	27.2	87.25
VQ-APC	8.12	10.8	81.25	8.37	7.4	60.50	7.84	22.4	94.25
NPC	7.74	39.0	92.75	8.15	21.1	76.75	7.86	30.4	94.75
Mockingjay	8.58	31.3	51.00	8.74	9.5	47.00	8.29	35.1	79.75
TERA	8.60	11.4	46.50	8.67	6.0	42.50	8.21	25.1	83.75
Modified CPC	8.71	9.4	40.00	8.87	7.0	30.00	8.41	26.2	71.00
DeCoAR 2.0	8.31	7.4	54.75	8.33	6.4	53.00	7.83	17.1	90.75
wav2vec	7.45	14.0	**95.50**	7.64	4.9	90.50	7.45	10.1	98.25
vq-wav2vec	**7.41**	13.4	91.00	**7.24**	11.6	**98.75**	**7.08**	13.4	**100.00**
wav2vec 2.0 Base	7.80	24.7	92.75	7.77	5.0	86.50	7.50	10.5	98.00
wav2vec 2.0 Large	7.64	12.5	81.75	7.67	9.0	82.75	7.63	15.8	97.25
HuBERT Base	7.70	**5.5**	89.25	7.79	**4.7**	84.25	7.47	**8.0**	98.50
HuBERT Large	7.54	5.6	95.00	7.54	5.6	93.00	7.22	9.0	99.25

Simple-AR → Taco2-AR: large improvements in ASV,
moderate degradation in WER

Results (1): Comparison of different models

Upstream	Intra-lingual A2O								
	Simple			Simple-AR			Taco2-AR		
	MCD	WER	ASV	MCD	WER	ASV	MCD	WER	ASV
mel	8.41	48.5	59.00	8.92	22.7	49.75	8.47	38.3	77.25
PPG (TIMIT)	7.78	69.0	85.50	7.83	58.9	95.25	7.18	33.6	99.75
PASE+	9.29	5.0	26.75	9.52	5.7	26.00	8.66	30.6	63.20
APC	8.67	8.6	48.00	8.73	7.1	41.75	8.05	27.2	87.25
VQ-APC	8.12	10.8	81.25	8.37	7.4	60.50	7.84	22.4	94.25
NPC	7.74	39.0	92.75	8.15	21.1	76.75	7.86	30.4	94.75
Mockingjay	8.58	31.3	51.00	8.74	9.5	47.00	8.29	35.1	79.75
TERA	8.60	11.4	46.50	8.67	6.0	42.50	8.21	25.1	83.75
Modified CPC	8.71	9.4	40.00	8.87	7.0	30.00	8.41	26.2	71.00
DeCoAR 2.0	8.31	7.4	54.75	8.33	6.4	53.00	7.83	17.1	90.75
wav2vec	7.45	14.0	**95.50**	7.64	4.9	90.50	7.45	10.1	98.25
vq-wav2vec	**7.41**	13.4	91.00	**7.24**	11.6	**98.75**	**7.08**	13.4	**100.00**
wav2vec 2.0 Base	7.80	24.7	92.75	7.77	5.0	86.50	7.50	10.5	98.00
wav2vec 2.0 Large	7.64	12.5	81.75	7.67	9.0	82.75	7.63	15.8	97.25
HuBERT Base	7.70	**5.5**	89.25	7.79	**4.7**	84.25	7.47	**8.0**	98.50
HuBERT Large	7.54	5.6	95.00	7.54	5.6	93.00	7.22	9.0	99.25

Taco2-AR is chosen to be the final model because:
(1) WER is a strict measurement of intelligibility
(2) Yields best MCD scores

Results (2): different tasks

Upstream	Intra-lingual A2O Taco2-AR MCD	Intra-lingual A2O Taco2-AR WER	Intra-lingual A2O Taco2-AR ASV	Cross-lingual A2O Taco2-AR WER	Cross-lingual A2O Taco2-AR ASV	Intra-lingual A2A Taco2-AR MCD	Intra-lingual A2A Taco2-AR WER	Intra-lingual A2A Taco2-AR ASV
mel	8.47	38.3	77.25	39.0	46.67	9.49	4.2	19.50
PPG (TIMIT)	7.18	33.6	99.75	51.0	84.67	**8.31**	12.9	**83.50**
PASE+	8.66	30.6	63.20	36.3	34.67	9.85	4.2	8.00
APC	8.05	27.2	87.25	33.9	52.33	9.57	3.5	23.25
VQ-APC	7.84	22.4	94.25	28.4	68.00	9.43	4.0	22.00
NPC	7.86	30.4	94.75	37.6	59.00	9.39	4.4	21.00
Mockingjay	8.29	35.1	79.75	39.2	46.00	9.43	5.0	25.00
TERA	8.21	25.1	83.75	29.2	49.33	9.31	5.2	18.75
Modified CPC	8.41	26.2	71.00	35.3	32.83	9.61	4.1	10.75
DeCoAR 2.0	7.83	17.1	90.75	26.8	59.33	9.28	4.0	27.00
wav2vec	7.45	10.1	98.25	13.9	75.83	8.77	3.5	40.00
vq-wav2vec	**7.08**	13.4	**100.00**	21.0	**88.83**	8.47	4.2	73.25
wav2vec 2.0 Base	7.50	10.5	98.00	14.9	82.17	9.03	3.2	27.00
wav2vec 2.0 Large	7.63	15.8	97.25	22.7	78.00	8.99	4.1	22.25
HuBERT Base	7.47	**8.0**	98.50	**13.5**	82.33	9.19	3.4	23.25
HuBERT Large	7.22	9.0	99.25	15.9	86.50	9.13	**3.0**	27.75

S3Rs still works in cross-lingual VC even trained on mono-lingual data. However, WER and ASV both degraded.

Results (2): different tasks

Upstream	Intra-lingual A2O			Cross-lingual A2O		Intra-lingual A2A		
	Taco2-AR			Taco2-AR		Taco2-AR		
	MCD	WER	ASV	WER	ASV	MCD	WER	ASV
mel	8.47	38.3	77.25	39.0	46.67	9.49	4.2	19.50
PPG (TIMIT)	7.18	33.6	99.75	51.0	84.67	**8.31**	12.9	**83.50**
PASE+	8.66	30.6	63.20	36.3	34.67	9.85	4.2	8.00
APC	8.05	27.2	87.25	33.9	52.33	9.57	3.5	23.25
VQ-APC	7.84	22.4	94.25	28.4	68.00	9.43	4.0	22.00
NPC	7.86	30.4	94.75	37.6	59.00	9.39	4.4	21.00
Mockingjay	8.29	35.1	79.75	39.2	46.00	9.43	5.0	25.00
TERA	8.21	25.1	83.75	29.2	49.33	9.31	5.2	18.75
Modified CPC	8.41	26.2	71.00	35.3	32.83	9.61	4.1	10.75
DeCoAR 2.0	7.83	17.1	90.75	26.8	59.33	9.28	4.0	27.00
wav2vec	7.45	10.1	98.25	13.9	75.83	8.77	3.5	40.00
vq-wav2vec	**7.08**	13.4	**100.00**	21.0	**88.83**	8.47	4.2	73.25
wav2vec 2.0 Base	7.50	10.5	98.00	14.9	82.17	9.03	3.2	27.00
wav2vec 2.0 Large	7.63	15.8	97.25	22.7	78.00	8.99	4.1	22.25
HuBERT Base	7.47	**8.0**	98.50	**13.5**	82.33	9.19	3.4	23.25
HuBERT Large	7.22	9.0	99.25	15.9	86.50	9.13	**3.0**	27.75

In A2A VC, only vq-wav2vec provided the required disentanglement.

Results (3): compare with SOTA

System	MCD	WER	ASV	Naturalness	Similarity
Intra-lingual A2O					
mel	8.47	38.3	77.25	2.61 ± 0.11	$35\% \pm 3\%$
PPG (TIMIT)	7.18	33.6	99.75	3.32 ± 0.10	$58\% \pm 4\%$
PASE+	8.66	30.6	63.20	2.58 ± 0.12	$31\% \pm 3\%$
APC	8.05	27.2	87.25	2.92 ± 0.11	$43\% \pm 4\%$
VQ-APC	7.84	22.4	94.25	3.08 ± 0.10	$40\% \pm 4\%$
NPC	7.86	30.4	94.75	2.98 ± 0.11	$46\% \pm 3\%$
Mockingjay	8.29	35.1	79.75	2.81 ± 0.12	$42\% \pm 4\%$
TERA	8.21	25.1	83.75	2.91 ± 0.12	$37\% \pm 4\%$
Modified CPC	8.41	26.2	71.00	2.74 ± 0.11	$33\% \pm 3\%$
DeCoAR 2.0	7.83	17.1	90.75	3.04 ± 0.11	$43\% \pm 4\%$
wav2vec	7.45	10.1	98.25	3.40 ± 0.05	$52\% \pm 2\%$
vq-wav2vec	7.08	13.4	100.00	3.59 ± 0.10	$59\% \pm 4\%$
wav2vec 2.0 B.	7.50	10.5	98.00	3.36 ± 0.06	$51\% \pm 2\%$
wav2vec 2.0 L.	7.63	15.8	97.25	3.26 ± 0.10	$50\% \pm 4\%$
HuBERT B.	7.47	8.0	98.50	3.48 ± 0.10	$55\% \pm 4\%$
HuBERT L.	7.22	9.0	99.25	3.47 ± 0.10	$54\% \pm 4\%$
USTC-2018†	–	6.5	99.00	4.20 ± 0.08	$55\% \pm 4\%$
USTC-2020	6.98	5.4	100.00	4.41 ± 0.07	$82\% \pm 3\%$
SRCB	8.90	11.5	92.00	4.16 ± 0.08	$68\% \pm 3\%$
CASIA	7.13	11.0	98.25	4.25 ± 0.08	$61\% \pm 4\%$
ASR+TTS	6.48	8.2	100.00	3.84 ± 0.09	$75\% \pm 3\%$
Target	–	0.7	–	4.57 ± 0.14	–

Cross-lingual A2O					
PPG (TIMIT)	–	51.0	84.67	2.79 ± 0.08	$43\% \pm 3\%$
vq-wav2vec	–	21.0	88.83	3.28 ± 0.08	$44\% \pm 3\%$
HuBERT L.	–	15.9	86.50	3.13 ± 0.08	$41\% \pm 3\%$
USTC-2018	–	5.6	97.67	4.17 ± 0.06	$34\% \pm 3\%$
USTC-2020	–	7.6	96.00	4.27 ± 0.07	$43\% \pm 3\%$
SRCB	–	8.6	78.67	4.34 ± 0.07	$34\% \pm 3\%$
CASIA	–	10.5	91.67	4.11 ± 0.07	$45\% \pm 3\%$
ASR+TTS	–	34.5	67.83	2.51 ± 0.08	$39\% \pm 3\%$
Target	–	–	–	4.48 ± 0.12	–

Intra-lingual A2A					
PPG (TIMIT)	8.32	12.7	84.25	3.41 ± 0.08	$34\% \pm 4\%$
vq-wav2vec	8.47	4.2	73.25	3.58 ± 0.09	$28\% \pm 3\%$
S2VC†	–	12.4	71.50	2.90 ± 0.09	$29\% \pm 3\%$

- Best upstream in A2O: vq-wav2vec
- There is still a gap between vq-wav2vec and SOTA
- vq-wav2vec beats S2VC (SOTA in A2A VC)

Results (4): impact of supervision

System	MCD	WER	ASV	Naturalness	Similarity
Intra-lingual A2O					
mel	8.47	38.3	77.25	2.61 ± 0.11	$35\% \pm 3\%$
PPG (TIMIT)	7.18	33.6	99.75	3.32 ± 0.10	$58\% \pm 4\%$
PASE+	8.66	30.6	63.20	2.58 ± 0.12	$31\% \pm 3\%$
APC	8.05	27.2	87.25	2.92 ± 0.11	$43\% \pm 4\%$
VQ-APC	7.84	22.4	94.25	3.08 ± 0.10	$40\% \pm 4\%$
NPC	7.86	30.4	94.75	2.98 ± 0.11	$46\% \pm 3\%$
Mockingjay	8.29	35.1	79.75	2.81 ± 0.12	$42\% \pm 4\%$
TERA	8.21	25.1	83.75	2.91 ± 0.12	$37\% \pm 4\%$
Modified CPC	8.41	26.2	71.00	2.74 ± 0.11	$33\% \pm 3\%$
DeCoAR 2.0	7.83	17.1	90.75	3.04 ± 0.11	$43\% \pm 4\%$
wav2vec	7.45	10.1	98.25	3.40 ± 0.05	$52\% \pm 2\%$
vq-wav2vec	7.08	13.4	100.00	3.59 ± 0.10	$59\% \pm 4\%$
wav2vec 2.0 B.	7.50	10.5	98.00	3.36 ± 0.06	$51\% \pm 2\%$
wav2vec 2.0 L.	7.63	15.8	97.25	3.26 ± 0.10	$50\% \pm 4\%$
HuBERT B.	7.47	8.0	98.50	3.48 ± 0.10	$55\% \pm 4\%$
HuBERT L.	7.22	9.0	99.25	3.47 ± 0.10	$54\% \pm 4\%$
USTC-2018†	–	6.5	99.00	4.20 ± 0.08	$55\% \pm 4\%$
USTC-2020	6.98	5.4	100.00	4.41 ± 0.07	$82\% \pm 3\%$
SRCB	8.90	11.5	92.00	4.16 ± 0.08	$68\% \pm 3\%$
CASIA	7.13	11.0	98.25	4.25 ± 0.08	$61\% \pm 4\%$
ASR+TTS	6.48	8.2	100.00	3.84 ± 0.09	$75\% \pm 3\%$
Target	–	0.7	–	4.57 ± 0.14	–

Cross-lingual A2O					
PPG (TIMIT)	–	51.0	84.67	2.79 ± 0.08	$43\% \pm 3\%$
vq-wav2vec	–	21.0	88.83	3.28 ± 0.08	$44\% \pm 3\%$
HuBERT L.	–	15.9	86.50	3.13 ± 0.08	$41\% \pm 3\%$
USTC-2018	–	5.6	97.67	4.17 ± 0.06	$34\% \pm 3\%$
USTC-2020	–	7.6	96.00	4.27 ± 0.07	$43\% \pm 3\%$
SRCB	–	8.6	78.67	4.34 ± 0.07	$34\% \pm 3\%$
CASIA	–	10.5	91.67	4.11 ± 0.07	$45\% \pm 3\%$
ASR+TTS	–	34.5	67.83	2.51 ± 0.08	$39\% \pm 3\%$
Target	–	–	–	4.48 ± 0.12	–

Intra-lingual A2A					
PPG (TIMIT)	8.32	12.7	84.25	3.41 ± 0.08	$34\% \pm 4\%$
vq-wav2vec	8.47	4.2	73.25	3.58 ± 0.09	$28\% \pm 3\%$
S2VC†	–	12.4	71.50	2.90 ± 0.09	$29\% \pm 3\%$

- A more fair comparison between PPG and S3Rs in a unified setting.
- PPG (TIMIT): trained with TIMIT (3 hours)
    - **Low quality** proven by high WER &low naturalness
    - **Good speaker disentanglement ability** shown by high ASV & high similarity

Results (5): justify the objective metrics

System	MCD	WER	ASV	Naturalness	Similarity
Intra-lingual A2O					
mel	8.47	38.3	77.25	2.61 ± 0.11	35% ± 3%
PPG (TIMIT)	7.18	33.6	99.75	3.32 ± 0.10	58% ± 4%
PASE+	8.66	30.6	63.20	2.58 ± 0.12	31% ± 3%
APC	8.05	27.2	87.25	2.92 ± 0.11	43% ± 4%
VQ-APC	7.84	22.4	94.25	3.08 ± 0.10	40% ± 4%
NPC	7.86	30.4	94.75	2.98 ± 0.11	46% ± 3%
Mockingjay	8.29	35.1	79.75	2.81 ± 0.12	42% ± 4%
TERA	8.21	25.1	83.75	2.91 ± 0.12	37% ± 4%
Modified CPC	8.41	26.2	71.00	2.74 ± 0.11	33% ± 3%
DeCoAR 2.0	7.83	17.1	90.75	3.04 ± 0.11	43% ± 4%
wav2vec	7.45	10.1	98.25	3.40 ± 0.05	52% ± 2%
vq-wav2vec	7.08	13.4	100.00	3.59 ± 0.10	59% ± 4%
wav2vec 2.0 B.	7.50	10.5	98.00	3.36 ± 0.06	51% ± 2%
wav2vec 2.0 L.	7.63	15.8	97.25	3.26 ± 0.10	50% ± 4%
HuBERT B.	7.47	8.0	98.50	3.48 ± 0.10	55% ± 4%
HuBERT L.	7.22	9.0	99.25	3.47 ± 0.10	54% ± 4%

Metric	MCD	WER	ASV	Nat.	Sim.
MCD	–	0.678	-0.934	-0.968	-0.961
WER	–	–	-0.640	-0.808	-0.587
ASV	–	–	–	0.910	0.911
Nat.	–	–	–	–	0.932
Sim.	–	–	–	–	–

linear corr. coeff. using the intra-lingual A2O results

Motivation: listening tests can be expensive.

Goal: examine if the objective measures align well with human perception.

Finding: **MCD best aligns** with both naturalness and similarity

Samples and codebase

- Demo webpage: https://unilight.github.io/Publication-Demos/publications/s3prl-vc/index.html

- Codebase: https://github.com/s3prl/s3prl/tree/master/s3prl/downstream/a2o-vc-vcc2020

Future research directions

- VC perspective:
  1. Better downstream model design.
     Ex. d-vector in A2A VC → a proper speaker encoder?
  2. Close performance gap between SOTA
     Ex. better vocoder, waveform modeling, etc.

- S3R perspective:
  1. Use VC as a probing task when designing new S3R
  2. Analyze what components are key to VC
     Ex. discretization (quantization) in vq-wav2vec?