S3PRL-VC: OPEN-SOURCE VOICE CONVERSION FRAMEWORK WITH SELF-SUPERVISED SPEECH REPRESENTATIONS

Wen-Chin Huang[1], *Shu-wen Yang*[2], *Tomoki Hayashi*[1], *Hung-yi Lee*[2], *Shinji Watanabe*[3], *Tomoki Toda*[1]

[1]Nagoya University, Japan
[2]National Taiwan University, Taiwan
[3]Carnegie Mellon University, USA

ABSTRACT

This paper introduces S3PRL-VC, an open-source voice conversion (VC) framework based on the S3PRL toolkit. In the context of recognition-synthesis VC, self-supervised speech representation (S3R) is valuable in its potential to replace the expensive supervised representation adopted by state-of-the-art VC systems. Moreover, we claim that VC is a good probing task for S3R analysis. In this work, we provide a series of in-depth analyses by benchmarking on the two tasks in VCC2020, namely intra-/cross-lingual any-to-one (A2O) VC, as well as an any-to-any (A2A) setting. We also provide comparisons between not only different S3Rs but also top systems in VCC2020 with supervised representations. Systematic objective and subjective evaluation were conducted, and we show that S3R is comparable with VCC2020 top systems in the A2O setting in terms of similarity, and achieves state-of-the-art in S3R-based A2A VC. We believe the extensive analysis, as well as the toolkit itself, contribute to not only the S3R community but also the VC community. The codebase is now open-sourced[1].

***Index Terms*—** voice conversion, open-source, self-supervised learning, self-supervised speech representation

1. INTRODUCTION

Voice conversion (VC) refers to a technique that converts a certain aspect of speech from a source to that of a target without changing the linguistic content [1, 2]. In this work, we focus on speaker conversion, which is the most widely investigated type of VC. From an information perspective, VC can be performed by first extracting the spoken contents from the source speech, and then synthesizing the converted speech from the extracted contents with the identity of the target speaker. Such a paradigm is sometimes referred to as recognition-synthesis (rec-syn) based VC, as depicted in Figure 1. Formally, starting from the source speech $\mathbf{X}$, a recognizer first extracts the spoken contents, $\mathbf{H}$, which is then consumed by the synthesizer to generate the converted speech, $\mathbf{Y}$:

$$\mathbf{Y} = \text{Synth}(\mathbf{H}), \mathbf{H} = \text{Recog}(\mathbf{X}). \quad (1)$$

In the latest voice conversion challenge 2020 (VCC2020) [3], one of the baselines directly concatenated an automatic speech recognition (ASR) model and a text-to-speech (TTS) model [4]. In addition, several top performing systems also implemented such a framework [5], showing state-of-the-art performance in terms of both naturalness and similarity.

[1]https://github.com/s3prl/s3prl/tree/master/s3prl/downstream/a2o-vc-vcc2020

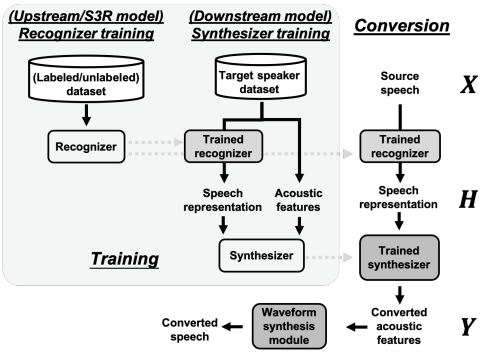

Fig. 1: The training and conversion procedures in any-to-one recognition-synthesis based VC.

In rec-syn based VC, an ASR model trained on a labeled dataset is often used to extract the *supervised* spoken content representation, such as text [4] or phonetic posteriorgram (PPG) [6]. The collection of labeled datasets is often costly, especially in a low-resource setting, such as the cross-lingual VC scenario [3]. Therefore, researchers have resorted to unsupervised or the so-called self-supervised speech representation (S3R) learning paradigm, where a large-scale unlabeled data is used to learn rich, compact speech representations. S3Rs have been applied to any-to-one VC [7], many-to-many VC [8], any-to-any VC [9, 10] and cross-lingual VC [11].

In addition to its label-free property, S3R based VC is also attractive in it being a good probing task for S3R analysis. A recently published SUPERB benchmark [12] dedicates to compare different S3Rs across a range of *discriminative* speech processing tasks, while it remains unclear what representations are optimal for *generation* tasks like VC. For instance, wav2vec 2.0 [13] has been shown to be powerful in not only ASR but also speaker and language recognition [14], implying that it encodes rich content, speaker and language information. Based on the discussion on the information perspective of VC, we may hypothesize that a good $\mathbf{H}$ in Eq. 1 should be compact in content but contains little to none speaker information. Based on such an assumption, wav2vec 2.0 may not be an optimal representation for VC.

In this paper, we describe S3PRL-VC, an extension of the S3PRL toolkit and SUPERB. Our main focus was any-to-one (A2O) VC, where the synthesizer is trained in a target-speaker-dependent fashion. We used the VCC2020 dataset, which allows us to test intra-lingual and cross-lingual settings. We also provide an any-to-any (A2A) extension by using an off-the-shelf d-vector [15] model to encode the unseen speaker information. We implemented models resembling the top systems in VCC2018 [16] and VCC2020 [17],

which allows us to focus on the comparison. We conducted a large-scale evaluation, both objectively and subjectively, to compare the performance between not only different S3Rs but also state-of-the-art systems. S3PRL-VC is a competitive system by yielding (1) a comparable performance with VCC2020 top systems in the A2O setting in terms of similarity, and (2) state-of-the-art performance in S3R-based A2A VC. Our main contributions are:

- Inheriting the property of SUPERB, our S3PRL-VC implementation ensures fast benchmarking but also state-of-the-art performance. Such a fast, easy-to-use property benefits not only S3R researchers but also the VC community.
- We present a large-scale comparison of S3Rs from the VC point-of-view, providing new insights and perspectives to analyze the representations. We also compared with top systems in VCC2020 that used PPGs, showing the limitation and competitiveness of S3Rs.

2. TASKS

2.1. General description of VCC2020

All experiments in this work are benchmarked on the VCC2020 dataset [3]. There are two tasks in VCC2020, with intra-lingual VC being task 1 and cross-lingual VC being task 2. The two tasks share the same two English male and female source speakers. The target speakers include two male and two female English speakers for task 1, and one male and one female speaker each of Finnish, German, and Mandarin for task 2. For each speaker, 70 utterances (roughly five minutes) in their respective languages and contents are provided, and there are 25 test sentences for evaluation. During conversion, $\mathbf{X}$ (which is in English) is converted as if it was uttered by the target speaker while keeping the linguistic contents unchanged.

2.2. Intra-lingual and cross-lingual any-to-one VC

We first consider the two tasks in VCC2020 under the A2O setting. Any-to-one VC aims to convert from any arbitrary speech into that of a predefined target speaker. The training and conversion processes are depicted in Figure 1. The ability to encode $\mathbf{H}$ from any unseen speaker is ensured by the common practice of training S3Rs on a multi-speaker dataset. Using the target speaker dataset, $\mathbf{D}_{\text{trg}}$, the synthesizer is trained to reconstruct the acoustic feature from $\mathbf{H}$. In the conversion phase, the converted features, $\mathbf{Y}$, are generated following Eq. 1. Finally, a waveform synthesizer (ex. neural vocoder) generates the converted waveform.

Any-to-one VC is a good probing task to investigate several characteristics of an upstream S3R model. First, a fundamental requirement of VC is the linguistic consistency, so there is a positive correlation between the VC performance of an S3R model and its ability to faithfully encode $\mathbf{H}$. Second, if an S3R model encodes rich speaker information, then the source speaker information in $\mathbf{X}$ will conflict with the target speaker attributes injected by the synthesizer, which hurts the VC performance. Finally, during the synthesizer training in cross-lingual VC, the S3R model may fail to generalize to $\mathbf{X}$ from a non-English target speaker since most existing S3R models are trained with English datasets only. It is worthwhile to examine the ability of mono-lingual S3R models to transfer to different languages.

2.3. Intra-lingual any-to-any VC

We then provide an extension for the A2A scenario, also known as zero-shot VC. A2A VC attempts to convert to a target speaker where

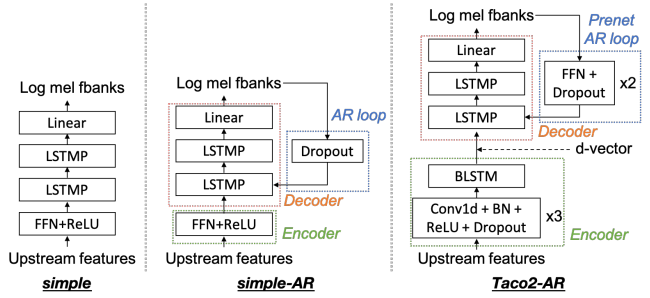

Fig. 2: The models implemented in this work. Left: the simple model. Middle: the simple model with an AR loop. Right: the Tacotron2 model, with extension to an any-to-any model by accepting a d-vector as the speaker embedding.

$\mathbf{D}_{\text{trg}}$ is so limited (less than one minute) such that fine-tuning in infeasible. A2A VC models are usually trained on a multi-speaker dataset. Instead of recovering the target speaker information by the synthesizer as in A2O VC, we use speaker embeddings, s, extracted by an off-the-shelf speaker encoder, which is pretrained on an automatic speaker verification (ASV) dataset and objective. Such a paradigm is also used in zero-shot TTS [18]. In training, the speaker embedding extracted from the target waveform is used. During conversion, given $\mathbf{D}_{\text{trg}}$, $\mathbf{s}$ is formed as an average of each embedding from each utterance. We may then rewrite Eq. 1 as:

$$\mathbf{Y} = \text{Synth}(\mathbf{H}, \mathbf{s}), \mathbf{H} = \text{Recog}(\mathbf{X}), \mathbf{s} = \text{SpkEnc}(\mathbf{D}_{\text{trg}}). \quad (2)$$

3. IMPLEMENTATION

3.1. Recognizer (upstream models)

Table 1 depicts the list of S3Rs we compared in this work, which are the upstream models supported in S3PRL at the date of publication. For a complete list of information (architecture, objective, etc.), refer to [12]. All upstreams are trained with English data (mostly Librispeech). In addition to the S3Rs, two extra upstreams were included: (1) mel-spectrogram, "mel", and (2) "PPG (TIMIT)", which is trained supervisedly on the TIMIT dataset.

3.2. Synthesizer model design

Mel-spectrogram was selected as the target acoustic feature. We implemented several models to resemble top systems of past VCCs, as illustrated in Figure 2. We avoid expensive model components like attention [19] for fast benchmarking.

Simple: We start from the model used by the top system in VCC2018 [16]. The simple model consists of a single layer feed-forward network (FFN), two long short-term memory layers with projection (LSTMP), and a linear projection layer.

Simple-AR: As autoregressive (AR) modeling has been shown to be effective in speech synthesis [20], we added an AR loop to the simple model. At each time step, the previous output is consumed by the first LSTMP layer. Dropout is essential in the AR loop to avoid exposure bias brought by teacher-forcing [21, 22].

Taco2-AR: We increase the model complexity by using a model architecture similar to that of Tacotron 2 [23], which resembles the model used by the top system in VCC2020 [17]. Different from Tacotron 2, the attention module was not used as it was reported to be useless in [17].

Table 1: Objective evaluation results on different VC settings over various S3Rs. For MCD and WER, the smaller the better; for ASV, the higher the better.

Upstream	Intra-lingual A2O									Cross-lingual A2O		Intra-lingual A2A		
	Simple			Simple-AR			Taco2-AR			Taco2-AR		Taco2-AR		
	MCD	WER	ASV	MCD	WER	ASV	MCD	WER	ASV	WER	ASV	MCD	WER	ASV
mel	8.41	48.5	59.00	8.92	22.7	49.75	8.47	38.3	77.25	39.0	46.67	9.49	4.2	19.50
PPG (TIMIT)	7.78	69.0	85.50	7.83	58.9	95.25	7.18	33.6	99.75	51.0	84.67	**8.31**	12.9	**83.50**
PASE+	9.29	5.0	26.75	9.52	5.7	26.00	8.66	30.6	63.20	36.3	34.67	9.85	4.2	8.00
APC	8.67	8.6	48.00	8.73	7.1	41.75	8.05	27.2	87.25	33.9	52.33	9.57	3.5	23.25
VQ-APC	8.12	10.8	81.25	8.37	7.4	60.50	7.84	22.4	94.25	28.4	68.00	9.43	4.0	22.00
NPC	7.74	39.0	92.75	8.15	21.1	76.75	7.86	30.4	94.75	37.6	59.00	9.39	4.4	21.00
Mockingjay	8.58	31.3	51.00	8.74	9.5	47.00	8.29	35.1	79.75	39.2	46.00	9.43	5.0	25.00
TERA	8.60	11.4	46.50	8.67	6.0	42.50	8.21	25.1	83.75	29.2	49.33	9.31	5.2	18.75
Modified CPC	8.71	9.4	40.00	8.87	7.0	30.00	8.41	26.2	71.00	35.3	32.83	9.61	4.1	10.75
DeCoAR 2.0	8.31	7.4	54.75	8.33	6.4	53.00	7.83	17.1	90.75	26.8	59.33	9.28	4.0	27.00
wav2vec	7.45	14.0	**95.50**	7.64	4.9	90.50	7.45	10.1	98.25	13.9	75.83	8.77	3.5	40.00
vq-wav2vec	**7.41**	13.4	91.00	**7.24**	11.6	**98.75**	**7.08**	13.4	**100.00**	21.0	**88.83**	8.47	4.2	73.25
wav2vec 2.0 Base	7.80	24.7	92.75	7.77	5.0	86.50	7.50	10.5	98.00	14.9	82.17	9.03	3.2	27.00
wav2vec 2.0 Large	7.64	12.5	81.75	7.67	9.0	82.75	7.63	15.8	97.25	22.7	78.00	8.99	4.1	22.25
HuBERT Base	7.70	**5.5**	89.25	7.79	**4.7**	84.25	7.47	**8.0**	98.50	**13.5**	82.33	9.19	3.4	23.25
HuBERT Large	7.54	5.6	95.00	7.54	5.6	93.00	7.22	9.0	99.25	15.9	86.50	9.13	**3.0**	27.75

3.3. Other setups

Any-to-any settings. The dataset used to train the A2A VC model is the VCTK dataset [24]. For the speaker encoder, we used the d-vector model [15] trained on a mix of datasets, including LibriSpeech, VoxCeleb 1 and 2.

Waveform synthesizer. We used the HiFi-GAN [25], a state-of-the-art parallel real-time neural vocoder. For the A2O setup, we mixed the data of all 14 speakers in VCC2020 with the VCTK dataset, while for the A2A setup we used only the VCTK dataset.

4. EVALUATION METRICS AND PROTOCOLS

4.1. Objective evaluation

We chose three objective evaluation metrics, all of which measure different aspects of a VC system. Mel cepstrum distortion (MCD) is an intrusive, L2-norm based metric which measures the general performance. Word error rate (WER) measures the intelligibility and the linguistic consistency, and in this work we used a pretrained wav2vec 2.0 model. The accept rate from a pretrained ASV model measures the speaker similarity by calculating the cosine similarity using speaker embeddings. For scenarios like the cross-lingual A2O task where the reference speech is not accessible, we report WER and ASV only since they are non-intrusive.

4.2. Subjective evaluation

For the subjective test, we asked listening participants to evaluate two common aspects in VC: naturalness and similarity. Listeners were asked to evaluate the naturalness on a five-point scale. For conversion similarity, a natural target speech and a converted speech were presented, and listeners were asked to judge whether the two samples were produced by the same speaker on a four-point scale.

For each system, a total of 80 utterances (5 random $\times$ 16 conversion pairs) were evaluated. Recordings of the target speakers were also included in the naturalness test and served as the upper bound. We used an open-source toolkit [26] that implemented the ITU-T Recommendation P.808 [27] to screen unreliable ratings obtained through the Amazon Mechanical Turk (Mturk). We recruited more than 280 listeners from the United States and had each sample rated by five different participants on average. Audio samples are available online[2].

5. EVALUATION RESULTS AND DISCUSSIONS

5.1. Comparison of different models

We first investigate the impact of using different synthesizer models described in Section 3.2 in the intra-lingual A2O setting, as shown in Table 1. First, only by adding the AR loop to the Simple model, most S3Rs benefit from large improvements in WER. With Taco-AR, all S3Rs except PASE+ and modified CPC achieved an ASV accept rate higher 80%, while all S3Rs suffered from a degradation in WER. This shows that increasing the model capacity can significantly improve the speaker similarity, while sacrificing the intelligibility. However, we would like to emphasize that WER is a strict measurement of intelligibility, and human can actually recognize better than machine. On the other hand, the Taco2-AR model yields the best MCD scores, which, as we will show later, correlates better with subjective naturalness and similarity. Also, we empirically found the training time of the three models similar. Based on these reasons, we decided to use the taco2-AR model for the succeeding tasks and comparisons.

5.2. Results on different tasks

Next, we compare the results of using S3Rs for different tasks. Looking again at Table 1, we first find S3Rs trained on a monolingual corpus can still work well in the cross-lingual setting, demonstrating the ability to transfer across languages. However, compared with the intra-lingual A2O task, it could be clearly observed that all S3Rs degraded in terms of both the WER and ASV accept rate, which is similar to the findings in [28]. Finally, in the intra-lingual A2A setting, all S3Rs yielded WERs much lower than those in the A2O setting, while the MCD values and ASV accept rates were significantly worse. Even the best upstream, vq-wav2vec, yielded only an accept rate of 73.25. One possible reason is that in the A2A VC setting, modern S3Rs still fail to disentangle content, such that the synthesizer preserves too much speaker information. Another reason may be that a jointly trained speaker encoder [10] is essential for S3R-based VC.

[2] https://bit.ly/3oydaY2

Table 2: Comparison with state-of-the-art systems. All upstreams use the Taco2-AR model.

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	$29\% \pm 3\%$

†: Systems generate 16kHz, so MCD is not calculable and direct score comparison should be made with caution.

5.3. Comparing with top systems using subjective evaluation

We then compared S3R-based VC models with state-of-the-art systems. **USTC-2018** [16], **USTC-2020** [5, 17][3], **SRCB** [29], **CASIA** [30] were top systems in VCC2020, all of which adopted PPGs, synthesizer pretraining on a multi-speaker dataset, and AR vocoders. Notably, they used thousands of hours of internal data for training. **ASR+TTS** [4] was the seq2seq+non-AR vocoder baseline in VCC2020. **S2VC** [10] is the SOTA system for A2A VC. The results are shown in Table 2. We summarize our observations as follows:

- vq-wav2vec outperformed all other upstreams in the subjective test, with a 3.59 naturalness and 59% similarity in the intra-lingual A2O setting.
- In the A2O settings, there was still a naturalness gap between vq-wav2vec and other VCC2020 top systems (3.59 v.s. 4.16-4.25, 3.28 v.s. 4.11-4.34). As for similarity, vq-wav2vec was on par with USTC-2018 and CASIA in the intra-lingual A2O setting, and achieved top in the cross-lingual setting.
- In the A2A setting, vq-wav2vec was on par with S2VC in similarity, while being significantly better in naturalness. Our system is therefore the new state-of-the-art in S3R-based A2A VC.

[3]USTC's systems used text and PPG for the intra-lingual and cross-lingual tasks, respectively.

Table 3: Linear correlation coefficients between different metrics.

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.	–	–	–	–	–

5.4. Impact of supervision

Although top systems using PPG greatly outperformed vq-wav2vec in naturalness, they used AR vocoders and the system was trained on large internal datasets, so the impact of supervision is not yet clear. To this end, we compared vq-wav2vec result with "PPG (TIMIT)" and the same vocoder. The high WERs and low naturalness scores showed that the PPG was indeed of low quality. Nonetheless, in all three settings, "PPG (TIMIT)" can achieve similar or higher similarity scores than vq-wav2vec. This shows that supervision greatly contributes to similarity, especially in difficult settings like A2A VC. This also shows that the ability of current S3Rs to disentangle speaker information is still limited when compared to PPG, and can be further improved in the future.

5.5. Justify the objective metrics with correlation analysis

Conducting a subjective test whenever a new S3R is developed cannot meet the fast benchmark requirement of SUPERB. Therefore, we examine if the objective measures align well with human perception. Using the intra-lingual A2O results over different upstreams, we calculated pairwise linear correlation coefficients. Results in Table 3 suggested that MCD best aligned with both naturalness and similarity. Note that in this correlation analysis, we considered systems that used the same decoder and neural vocoder. Since the correlation result is strongly affected by the pool of methods evaluated in a listening test, this good correlation could be observed only in such a homogeneous condition. Nonetheless, this result is still very useful for the benchmarking requirement of SUPERB.

6. CONCLUSIONS AND FUTURE WORK

We presented S3PRL-VC, an extension of the S3PRL toolkit that applied S3R to VC. We described the model design choice, and covered a variety of tasks. Extensive experiments, both objective and subjective, evaluated the capability of various S3Rs when applied to different VC scenarios. By comparing S3Rs with supervised presentations like PPG, we showed the competitiveness of S3Rs in certain settings, meanwhile shedding light on improving directions.

We suggest different future directions for readers from different communities. From the VC perspective, it is worthwhile to continue investigating better downstream model design. For instance, in A2A VC, a proper speaker encoder should be used instead of fixed d-vector. Meanwhile, we encourage to use VC as a probing task when designing a new S3R model, considering the challenges to overcome brought by all aspects required in VC.

Acknowledgements We would like to thank the S3PRL/SUPERB team for the fruitful discussions. This work was partly supported by JSPS KAKENHI Grant Number 21J20920, JST CREST Grant Number JPMJCR19A3, and a project, JPNP20006, commissioned by NEDO, Japan.

7. REFERENCES

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE TSAP*, vol. 6, no. 2, pp. 131–142, 1998.

[2] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE TASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice Conversion Challenge 2020 - Intra-lingual semi-parallel and cross-lingual voice conversion -," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 80–98.

[4] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 160–164.

[5] J.-X. Zhang, L.-J. Liu, Y.-N. Chen, Y.-J. Hu, Y. J., Z.-H. Ling, and L.-R. Dai, "Voice Conversion by Cascading Automatic Speech Recognition and Text-to-Speech Synthesis with Prosody Transfer," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 121–125.

[6] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, 2016, pp. 1–6.

[7] W.-C. Huang, Y.-C. Wu, T. Hayashi, and T. Toda, "Any-to-One Sequence-to-Sequence Voice Conversion using Self-Supervised Discrete Speech Representations," in *Proc. ICASSP*, 2021, pp. 5944–5948.

[8] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech*, 2021, pp. 3615–3619.

[9] Y. Y Lin, C.-M. Chien, J.-H. Lin, H.-Y. Lee, and L.-S. Lee, "FragmentVC: Any-to-Any Voice Conversion by End-to-End Extracting and Fusing Fine-Grained Voice Fragments With Attention," in *Proc. ICASSP*, 2021, pp. 5939–5943.

[10] J.-H. Lin, Y. Y Lin, C.-M. Chien, and H.-Y. Lee, "S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations," in *Proc. Interspeech*, 2021, pp. 836–840.

[11] W.-C. Huang, T. Hayashi, X. Li, S. Watanabe, and T. Toda, "On Prosody Modeling for ASR+ TTS based Voice Conversion," in *Proc. ASRU*, 2021.

[12] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Yi Lee, "SUPERB: Speech processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeruIPS*, 2020.

[14] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, 2021, pp. 1509–1513.

[15] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.

[16] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet Vocoder with Limited Training Data for Voice Conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987.

[17] L.-J. Liu, Y.-N. Chen, J.-X. Zhang, Y. Jiang, Y.-J. Hu, Z.-H. Ling, and L.-R. Dai, "Non-Parallel Voice Conversion with Autoregressive Conversion Model and Duration Adjustment," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 126–130.

[18] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," in *Proc. NIPS*, 2018, pp. 4480–4490.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[20] X. Wang, S. Takaki, and J. Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4895–4899.

[21] X. Wang, S. Takaki, and J. Yamagishi, "Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis," *IEEE/ACM TASLP*, vol. 26, no. 8, pp. 1406–1419, 2018.

[22] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[23] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[24] C. Veaux, J. Yamagishi, and Kirsten MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017.

[25] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, 2020, vol. 33, pp. 17022–17033.

[26] B. Naderi and R. Cutler, "An Open Source Implementation of ITU-T Recommendation P.808 with Validation," in *Proc. Interspeech*, 2020, pp. 2862–2866.

[27] ITU-T Recommendation P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," 2018.

[28] R. Kumar Das, T. Kinnunen, W.-C. Huang, Z.-H. Ling, J. Yamagishi, Z. Yi, X. Tian, and T. Toda, "Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 99–120.

[29] Q. Ma, R. Liu, X. Wen, C. Lu, and X. Chen, "Submission from SRCB for Voice Conversion Challenge 2020," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 131–135.

[30] L. Zheng, J. Tao, Z. Wen, and R. Zhong, "CASIA Voice Conversion System for the Voice Conversion Challenge 2020," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 136–139.