

2. RELATED WORKS

2.1. Past voice conversion challenges

VCC2016 [8] was held as a special session at INTERSPEECH 2016, and attracted 17 participants. A parallel VC database consisted of two source and two target native American English speakers (two females and two males), each of whom spoke 162 parallel sentences, was constructed for the only task in VCC2016. It was reported that the best system in VCC 2016 obtained an average naturalness score of 3.0 and a similarity score of 70%¹. However, it was obvious that there was a huge gap between the target natural speech and the converted speech.

VCC2018 [9] was held as a special session of the ISCA Speaker Odyssey Workshop 2018 and attracted 32 participants. The two tasks were based on a newly constructed but smaller parallel VC database and a non-parallel VC database. There were four native speakers of American English (two females and two males) for both the target and source speakers, each of whom uttered 80 sentences. The same evaluation methodology in VCC 2016 was adopted for the 2018 challenge, and we observed significant progress. The best system performed well in both parallel and non-parallel tasks and obtained an average of 4.1 in naturalness and about 80% in similarity. However, it was confirmed that there were still statistically significant differences between the target natural speech and the best converted speech in terms of both naturalness and speaker similarity.

VCC2020 [10] was held in a joint workshop with the Blizzard Challenge [11] and attracted 30 participants. There was a semi-parallel intra-lingual conversion task and a cross-lingual conversion task, with two corresponding datasets newly built. For more details, please refer to [10]. The listening test results first showed that for the intra-lingual semi-parallel task, the speaker similarity scores of several systems were as high as the target speakers, while none of them achieved human-level naturalness. For the relatively harder cross-lingual task, although the overall naturalness and similarity scores were lower, the best systems had naturalness scores higher than 4.0 and similarity scores above 70%.

2.2. Singing voice conversion

The task of SVC aims at converting the singing voice of a source singer to that of a target singer without changing the contents. Mainstream SVC models can be categorized into two classes: 1) parallel spectral feature matching models and 2) information disentanglement based models. Early works on SVC use parametric statistical models, such as Gaussian mixture models (GMMs), to model source-target spectral conversion function leveraging parallel singing data [12, 13]. Parallel approaches based on generative adversarial networks (GANs) have also been proposed to improve conversion performance [14]. Since parallel singing data is expensive to collect on a large scale, especially in multi-singer applications, researchers have investigated the use of non-parallel data for SVC. Both implicit and explicit information disentanglement methods have been studied to decompose voice identity, pitch, and linguistic content from a singing voice. CycleGAN and StarGAN-based SVC models [15, 16] use adversarial training and cycle consistency loss to implicitly disentangle voice identity from other information including linguistic content, pitch information, etc. The encoder-decoder framework is another hot topic in the research of SVC, which explicitly use either domain confusion loss or textual supervision

¹Defined as the percentage of a system’s converted samples that were judged to be the same as the target speakers.

to obtain pitch-invariant and singer-invariant content representation. An auto-encoder-based unsupervised SVC model is studied, which uses speaker confusion techniques to disentangle singer information from the encoder output [17]. Based on this model, PitchNet [18] employs an additional adversarial pitch confusion term to extract pitch-invariant and singer-invariant features from the encoder. Rather than relying on domain confusion losses, various models separately train a content encoder model and an information-fusion decoder model to tackle the task of SVC. The encoder uses text supervision to obtain singer-invariant content features, either through phonetic posteriorgrams (PPGs) [19] or features extracted from some intermediate layer in an ASR acoustic model [20, 21]. To increase the expressiveness of the model, it is likely that the decoder incorporates generative modeling, such as auto-regressive models [19], GANs [20–22], or denoising diffusion probabilistic models (DDPMs) [23].

It is worthwhile to note that the SVC open-source community has been extremely active recently. The most popular project, *so-vits-svc*, has over 15k stars on its Github repository². It is a collective effort of over 30 contributors, providing training scripts on a variety of encoders, acoustic models, and vocoders.

3. TASKS, DATABASES, AND TIMELINE FOR SINGING VOICE CONVERSION CHALLENGE 2023

Similar to the past VCC iterations, the primary objective is to conduct speaker conversion. For SVCC 2023, we separate the challenge into two any-to-one tasks: in-domain SVC and cross-domain SVC. The organizers developed a dedicated challenge dataset for the challenge and released the dataset in a manner that gave the participants around two months to train their models.

Task 1: In-domain SVC: For Task 1, the main task was to convert to a target speaker, by using the target speakers’ singing voices as training data. Compared to speech, the prosody of singing voices mostly follows musical notes rather than that of the spoken language. Although some previous VC methods could be directly applied to singing datasets, the main point of the task was to verify which methods could effectively replicate how the target singer sings the musical notes.

Task 2: Cross-domain SVC: For Task 2, the main task was to convert to a target speaker, by using the target speakers’ speech data. Compared to Task 1, Task 2 is generally considered harder as the model does not see how the target speaker’s singing voice sounds, as the pitch range in the dataset is narrower compared to the one in Task 1. Moreover, a person’s singing style cannot be seen from their speech alone. Although a more challenging task, it is important to note that Task 2 may be a more realistic and generalizable setting, as not all humans have the formal training to control their vocal cords and sing songs in the correct notes or key.

3.1. Dataset construction

The SVCC 2023 database is a subset of the NUS-HLT Speak-Sing (NHSS) dataset [24]. The original database is parallel in the sense that it contains a speaker’s singing and speech data. Each speaker records 10 songs from a selection of 20 songs, making the dataset semi-parallel. For both tasks, we use six songs from each speaker as the training data. For the evaluation data, we used six phrases from each of the remaining four songs. We labeled the target singers

²<https://github.com/svc-develop-team/so-vits-svc>, accessed on 2023.6.18.

Table 1: An overview of the SVCC 2023 dataset.

SVCC 2023 ID	NHSS ID	Minutes	No. of phrases
IDM1	M04	11.84	150
IDF1	F01	12.72	159
CDM1	M03	4.31	161
CDF1	F02	6.75	150
SM1	M02	2.35	24
SF1	F04	2.39	24

Table 2: List of participant affiliations of SVCC 2023 in random order. In addition, five participants did not identify themselves.

Affiliation	Task 1	Task 2
University of Sheffield	Y	Y
RIKEN Guardian Robot Project	Y	Y
Duke Kunshan University	Y	N
WIZ.AI	Y	Y
National Tsing Hua University	Y	N
Huya.Inc	Y	Y
Advanced Micro Devices, Inc.	Y	Y
Samsung Research China-Beijing	Y	Y
TME Lyra Lab, Northwestern Polytechnical University, Xian Jiaotong University	Y	Y
Shanghai Jiao Tong University	Y	Y
Bilibili Inc.	Y	Y
The Chinese University of Hong Kong (Shenzhen)	Y	Y
Northwestern Polytechnical University, TME Lyra Lab	Y	Y
Soochow University	Y	Y
Nagoya University	Y	Y
Parakeet Inc.	Y	Y
Federal university of Goiás (UFG)	Y	Y
Individual 1	Y	Y
Individual 2	Y	Y

for Task 1 with "ID" and the target speakers "CD" for Task 2. On the other hand, the source speakers were labeled with "S". Male speakers were labeled with "M", while female speakers were labeled with "F". An overview of the dataset's details is shown in Table 1. An open-sourced script³ can be used to generate the SVCC 2023 dataset from the NHSS dataset. Aside from the SVCC 2023 dataset, we allowed participants to use other external datasets, provided that these were open-sourced to allow reproducible experiments.

3.2. Timeline

The challenge was first announced and promoted on January 19, 2023. Training data was released on February 17, 2023, while the evaluation data was released on April 21, 2023, giving participants around two months to develop their models. Participants were then asked to submit their converted results on April 28, 2023, along with a brief description of their systems.

4. PARTICIPANTS AND SUBMITTED SYSTEMS

4.1. Challenge participants

Table 2 shows the participant affiliations and in which tasks they participated, listed in random order. In total, we have 24 submissions and 2 baselines systems, ending up with 25 and 24 systems in Tasks 1 and 2, respectively. As in previous VCCs, we anonymized each team with a unique team ID (T01 to T24 for the participants and B01 and B02 for the baseline systems), and informed each team of their own team ID except for five participants who did not submit system descriptions despite repeated warnings from the organizers. The ordering is random and different from that in Table 2.

³https://github.com/lesterphillip/SVCC23_FastSVC/tree/main/egs/generate_dataset

Table 3: Details of participating systems in SVCC 2023.

ID	Content Feature	VAE	Vocoder
B01	PPG	N	HiFi-GAN
B02	HuBERT	N	HN-uSFGAN
T01	PPG	N	HiFi-GAN + BigVGAN*
T02	HuBERT	Y	DSPGAN
T03	HuBERT	Y	HiFi-GAN
T04	Unknown		
T05	PPG	N	HiFi-GAN
T06	ContentVec	Y	N/A (nsf-HiFi-GAN)‡
T07	HuBERT	Y	N/A (HiFi-GAN)‡
T08	Unknown		
T09	Uncertain	Y	nsf-HiFi-GAN
T10	WavLM	N	BigVGAN
T11	PPG	N	HiFi-GAN
T12	HuBERT	N	HiFi-GAN
T13	ContentVec	N	SiFi-GAN
T14	Unknown		
T15	None (Melspec)†	N	nsf-HiFi-GAN
T16	PPG	N	BigVGAN
T17	HuBERT	N	nsf-HiFi-GAN
T18	Unknown		
T19	None (Melspec)†	N	HiFi-GAN
T20	HuBERT	Y	nsf-HiFi-GAN
T21	ContentVec	Y	nsf-HiFi-GAN
T22	PPG+ContentVec	N	BigVGAN
T23	PPG	Y	DSPGAN
T24	Unknown		

"Unknown" implies teams who did not submit their system description.

*: BigVGAN was used as a postfilter.

†: No content feature as only the melspectrogram was used.

‡: No vocoder was used since the decoder outputs waveform.

4.2. Baseline systems

B01 (DiffSVC System): The first baseline system is similar to the system presented in the DiffSVC paper [23], which was considered state-of-the-art as we organized this challenge. The detailed description is presented in Appendix D.

B02 (Decomposed FastSVC System): The second baseline system aims to provide a simple open-sourced baseline⁴ for the challenge. The network is similar to FastSVC [21], but decomposed into an acoustic model and a vocoder to reduce training time. A detailed description of the system is found in Appendix E.

4.3. Description of the submitted systems

4.3.1. Common components

Most systems this year adopt the recognition-synthesis (rec-syn) framework⁵, where several encoders (or recognizer) are first used to extract a set of features, including a content feature which contains compact linguistic or phonetic information from the input, and prosodic related features such as f0, energy, etc. Then, conversion is mostly carried out by a decoder (or synthesizer) to inject target information. The content feature encoder is usually trained to be speaker-independent, thus is assumed to be capable of handling any unseen speaker. The decoder training is often conducted by pre-training on a multi-speaker/singer dataset, and then fine-tuned on the target dataset or directly uses a speaker embedding to control the identity. Exceptions are T15 and T19, both of whom adopted StarGANv2-VC [16] which jointly trains the encoders and decoder. Finally, we noticed that most teams did not develop special techniques for individual tasks.

⁴https://github.com/lesterphillip/SVCC23_FastSVC

⁵Following the definition in [25], any VC system that separately trains the recognizer and synthesizer can be categorized as the rec-syn framework.

4.3.2. Taxonomy

While the VCC 2020 analysis paper [10] analyzed the submitted systems by the feature conversion model and the vocoder, we found that the viewpoint should advance along with the development of VC techniques. This year, we base our analysis on three aspects that give the largest variations among different systems: content feature type, use of variational autoencoders (VAEs), and vocoder type. Note that the goal of this section is not to derive meaningful tendencies or scientific differences, but rather a trend of popular techniques used in the current moment.

Content feature type. The content feature plays an important role in rec-syn based VC. A good content feature should be rich in content but contains little to no speaker information [25]. To facilitate this property, the PPG is a straightforward choice as it is derived from an ASR model which is trained in a supervised fashion to extract linguistic information. A total of 7 teams used PPGs. In recent years, self-supervised learning (SSL) based speech representations are drawing attention in VC as they benefit from large-scale unlabeled corpora and are shown to be able to disentangle speaker information. Among the 12 teams that used SSL speech representations, popular choices included HuBERT [26], WavLM [27] and ContentVec [28].

Use of VAEs. Introducing the VAE probabilistic framework in conditional generative models improves the generalization ability to unseen condition combinations [29, 30], which is essential in low-resource tasks like SVC. This is backed by the fact that, among the 8 teams that adopted VAE, many of them ranked in the top three in Tasks 1 and 2, as we will show in later sections.

Vocoder type. Despite the development of end-to-end SVC systems [17], we observed that most teams still adopt a two-stage framework such that a converted acoustic feature (mostly mel spectrogram) is first generated, and then a vocoder is used to generate the final waveform. Exceptions are T06 and T07, who directly trained their decoders to generate the converted waveform. All vocoders used by this year’s teams are still based on GANs, showing that these are still the most popular choice when it comes to vocoders, despite the progress in other generative frameworks like flow-based models or DDPMs. While 8 teams used the original HiFi-GAN [31], 5 teams used its neural source filter (NSF) extension, which combines NSF [32] to improve the generalization ability. Other popular choices include BiGVGAN [33], SiFi-GAN [34] and DSPGAN [35]. We noticed that both T23 and T02, the top systems for Tasks 1 and 2 in naturalness, respectively, adopted DSPGAN. However, this sample size is too small to conclude that DSPGAN is the ideal choice for SVC.

Other notable observations. Due to the scarcity of singing voice datasets, many teams included speech data to train their models. For example, T07 used more than 1000 hours of speech training data. While 7 teams applied DDPMs, most teams still used classical deep learning frameworks like VAEs or GANs. Finally, 5 teams mentioned that they directly based their system on the *so-vits-svc* project.

5. SUBJECTIVE EVALUATION

As in the previous VCCs, the perceptual study is considered the main evaluation method in SVCC 2023. Here we present, to our knowledge, by far the first large-scale subjective evaluation for SVC. In the following sections, we consider the results of the English subjects’ main results.

5.1. Listening test setup

Two common aspects of VC are evaluated in this challenge, namely naturalness and similarity. The protocol was basically consistent with that in the previous VCCs, where listeners were asked to evaluate the naturalness on a five-point scale, and 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 more details, please refer to the VCC 2020 paper [10].

Crowdsourcing on platforms like Amazon Mechanical Turk has been attractive in recent years thanks to its efficiency; however, it suffers from listener quality variations and trustworthy issues. Considering budget constraints, we followed the same protocol in VCC 2020 and outsourced the recruiting of listeners to two companies. Specifically, English and Japanese listeners were recruited by the Inter Group Corporation and Koto Ltd., respectively. The two sets of perceptual evaluation required a total of more than ¥700,000 Japanese yen. Each evaluation set contained 53 webpages (25 systems for Task 1, 24 systems for Task 2, and source/target for Tasks 1 and 2), each of which contained one naturalness and one similarity question to evaluate the same sample. The numbers of total and average scores per system from the English/Japanese listeners are 12720/38160 and 120/360, respectively.

5.2. Main results on English listeners

5.2.1. Naturalness

Figure 2 shows the boxplot of the naturalness evaluation results of Tasks 1 and 2. First, baseline B01 was outperformed by around half and one-thirds of the teams in Tasks 1 and 2, respectively, showing that the SVC field has made significant progress in naturalness since DiffSVC was proposed. The top system in Task 1 was T23, which ranked second in Task 2. On the other hand, T02, the top system in Task 2, ranked fifth in Task 1. Although no system had a mean score higher than those of the source and target, Figures 3a and 3c show that T23, T07, and T02 are in fact not significantly different from the natural samples, showing that **the top systems have reached human-level naturalness**. Finally, we can see that in Task 2, only 8 teams scored more than 3.0, compared to the 14 teams in Task 1, showing that cross-domain SVC is indeed harder than in-domain SVC.

5.2.2. Similarity

Figure 4 shows the results for the similarity evaluation results of Tasks 1 and 2. The similarity percentage is defined as the sum of the percentages from the “same (not sure)” and “same (sure)” categories, and the averaged scores are also shown. First, different from the naturalness results, baseline B01 ranked the fifth and the first in Tasks 1 and 2, respectively. T14, the top system in Task 1, ranked the second in Task 2. In contrast to naturalness, there is a clear gap (around 0.4 points) between the target samples and the top system in both tasks. This can also be observed from Figures 3b and 3d, which show that the target samples were significantly better than all other systems in terms of similarity. In conclusion, **when it comes to similarity in SVC, there is still a large room for improvement**.

Similar to naturalness, there is a significant similarity degradation in Task 2. To our surprise, even the target samples suffer from such a drop (3.4 v.s. 3.0). As we manually inspected the natural samples, we found that due to the high variation of singing voices, different phrases of the same singer in the same song can sound like

Table 4: Spearman correlation between objective and subjective metrics. Highlights in red indicate the highest correlation with corresponding subjective metrics among the objective metrics. CER metric refers to Conformer-based speech recognition results, while CER+ refers to Whisper results. Significance levels are shown by * (Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$).

Sub. Score	Listener	MCD	FORMSE	F0CORR	CER	CER+	D_{Embed}	UTMOS	SSL-MOS
Task 1 MOS	JPN	-0.28	-0.41**	0.48**	-0.62***	-0.80***	-0.58***	0.77***	0.53***
	ENG	-0.24	-0.28	0.45**	-0.57***	-0.73***	-0.45**	0.72***	0.42
Task 1 SIM	JPN	-0.62***	-0.26	0.37*	-0.42**	-0.40**	-0.83***	0.49**	0.30
	ENG	-0.45**	-0.10	0.21	-0.26	-0.27	-0.63***	0.38*	0.13
Task 2 MOS	JPN	-0.38*	-0.27	0.10	-0.62***	-0.77***	-0.58***	0.60***	0.15
	ENG	-0.29	-0.06	-0.16	-0.60***	-0.73***	-0.45**	0.49**	0.11
Task 2 SIM	JPN	-0.38*	-0.67***	0.03	-0.25	-0.53***	-0.67***	-0.08	-0.27
	ENG	-0.29	-0.22	-0.37*	-0.11	-0.28	-0.41**	-0.20	-0.23



Fig. 4: Similarity results for Tasks 1 and 2. Similarity scores are arranged in accordance with their mean value (red dot). SOU and TAR represents the source and target samples, respectively.

ESPnet+S3PRL [44–46] and the whisper-large ASR model⁷ [47].

Speaker similarity: Previous works in SVC have explored using a singer identification/verification model to estimate singer conversion accuracy [17, 20, 48]. Some other works also estimate the singer similarity with pre-trained singer embedding [21, 22]. In this work, we adopt RawNet-3 based speaker embedding [49] to estimate the singer similarity by calculating their cosine similarity (i.e., D_{Embed}).

Neural MOS predictor: As a previous work also use Neural MOS predictor [22], we also examine the performance of pretrained neural MOS predictor with the baseline system (SSL-MOS) and best system (UTMOS) in VoiceMOS Challenge 2022 [50, 51].

⁷<https://github.com/openai/whisper>

6.2. Analysis with subjective evaluation

In order to examine the relationship between subjective and objective evaluation metrics, we computed the Spearman correlation coefficients for each metric. The detailed results can be found in Table 4. (1) In most cases, metrics related to spectrogram and fundamental frequencies do not exhibit a significant correlation with subjective evaluation, which diverges from the findings of previous studies on VC in speech. (2) Speech recognition measures, both for the Conformer-based recognizer and Whisper, demonstrate a noteworthy correlation with the subjective MOS. (3) Currently, it is challenging to accurately assess singer similarity using objective metrics. The singer embedding cosine distance performs the best among the metrics, showing statistical significance for both Task 1 and Task 2 evaluations among Japanese and English speakers. However, even this metric yields insignificant results when assessing the similarity of Task 2 subjective measures with native speakers. (4) Despite being trained on speech corpora, the existing MOS predictor, UTMOS, exhibits a moderate correlation with subjective measures of naturalness, indicating its generalization capability.

7. CONCLUSION

The singing voice conversion challenge 2023 is the fourth edition of the voice conversion challenge series, held to compare and understand different VC systems built on a common dataset. We introduced two tasks, namely any-to-one in-domain SVC and any-to-one cross-domain SVC, and curated a database which is essentially a subset of the NHSS dataset. After giving participants two and a half months to train their SVC systems, we received a total of 26 submissions, including 2 baselines. As the first large-scale listening test for SVC, we observed that the top SVC systems in both tasks have achieved human-level naturalness. However, we also confirmed that there is a significantly large gap between the similarity scores of the target and all submitted systems. In addition, we confirmed that the cross-domain task is indeed a more difficult task, as the overall naturalness and similarity scores were lower. Finally, we showed that as few objective evaluation metrics can moderately correlate with the subjective scores, even the metric that best correlates with the similarity scores only yields a weak correlation, showing that objective assessment for SVC still has a lot to improve.

8. ACKNOWLEDGMENTS

This work was partly supported by JSPS KAKENHI Grant Number 21J20920, and JST CREST Grant Number JPMJCR19A3. We

thank Yusuke Yasuda from Nagoya University for his assistance in conducting the listening test.

9. REFERENCES

- [1] S. H. Mohammadi and A. Kain, “An overview of voice conversion systems,” *Speech Communication*, vol. 88, pp. 65–82, 2017.
- [2] B. Sisman, J. Yamagishi, S. King, and H. Li, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM TASLP*, vol. 29, pp. 132–157, 2021.
- [3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech,” *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.
- [4] T. Toda, “Augmented speech production based on real-time statistical voice conversion,” in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 592–596.
- [5] K. Probst, Y. Ke, and M. Eskenazi, “Enhancing foreign language tutors – in search of the golden speaker,” *Speech Communication*, vol. 37, no. 3, pp. 161–173, 2002.
- [6] J. Latorre, V. Wan, and K. Yanagisawa, “Voice expression conversion with factorised hmm-tts models,” in *Proc. Interspeech*, 2014, pp. 1514–1518.
- [7] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE/ACM TASLP*, vol. 20, no. 9, pp. 2505–2517, 2012.
- [8] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The voice conversion challenge 2016,” in *Proc. Interspeech*, 2016, pp. 1632–1636.
- [9] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” in *Proc. Odyssey The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.
- [10] 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.
- [11] X. Zhou, Z.-H. Ling, and S. King, “The Blizzard Challenge 2020,” in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 1–18.
- [12] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Statistical singing voice conversion with direct waveform modification based on the spectrum differential,” in *Proc. Interspeech*, 2014.
- [13] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Statistical singing voice conversion based on direct waveform modification with global variance,” in *Proc. Interspeech*, 2015.
- [14] B. Sisman, K. Vijayan, M. Dong, and H. Li, “Singan: Singing voice conversion with generative adversarial networks,” in *Proc. APSIPA ASC*, 2019, pp. 112–118.
- [15] B. Sisman and H. Li, “Generative Adversarial Networks for Singing Voice Conversion with and without Parallel Data,” in *Proc. Odyssey The Speaker and Language Recognition Workshop*, 2020, pp. 238–244.
- [16] Y. A. Li, A. Zare, and N. Mesgarani, “StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion,” in *Proc. Interspeech*, 2021, pp. 1349–1353.
- [17] E. Nachmani and L. Wolf, “Unsupervised singing voice conversion,” in *Proc. Interspeech*, 2019.
- [18] C. Deng, C. Yu, H. Lu, C. Weng, and D. Yu, “Pitchnet: Unsupervised singing voice conversion with pitch adversarial network,” in *Proc. ICASSP*, 2020, pp. 7749–7753.
- [19] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, “Ppg-based singing voice conversion with adversarial representation learning,” in *Proc. ICASSP*, 2021, pp. 7073–7077.
- [20] A. Polyak, L. Wolf, Y. Adi, and Y. Taigman, “Unsupervised cross-domain singing voice conversion,” in *Proc. Interspeech*, 2020.
- [21] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, “FastSVC: Fast cross-domain singing voice conversion with feature-wise linear modulation,” in *Proc. ICME*, 2021, pp. 1–6.
- [22] Y. Zhou and X. Lu, “HiFi-SVC: Fast High Fidelity Cross-Domain Singing Voice Conversion,” in *Proc. ICASSP*, 2022, pp. 6667–6671.
- [23] S. Liu, Y. Cao, D. Su, and H. Meng, “Diffsvc: A diffusion probabilistic model for singing voice conversion,” in *Proc. ASRU*, 2021, pp. 741–748.
- [24] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, “NHSS: A speech and singing parallel database,” *Speech Communication*, vol. 133, pp. 9–22, 2021.
- [25] W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, “A comparative study of self-supervised speech representation based voice conversion,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1308–1318, 2022.
- [26] W.-N. Hsu, Y.-H. Hubert Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, “HuBERT: How Much Can a Bad Teacher Benefit ASR Pre-Training?,” in *Proc. ICASSP*, 2021, pp. 6533–6537.
- [27] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [28] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, “ContentVec: An improved self-supervised speech representation by disentangling speakers,” in *Proc. ICML*, Jul 2022, vol. 162, pp. 18003–18017.
- [29] Y. Ren, J. Liu, and Z. Zhao, “PortaSpeech: Portable and High-Quality Generative Text-to-Speech,” in *Proc. NeruIPS*, 2021, vol. 34, pp. 13963–13974.
- [30] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” in *Proc. ICML*, Jul 2021, vol. 139, pp. 5530–5540.

- [31] 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.
- [32] X. Wang, S. Takaki, and J. Yamagishi, “Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis,” *IEEE/ACM TASLP*, vol. 28, pp. 402–415, 2020.
- [33] S. g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “BigVGAN: A universal neural vocoder with large-scale training,” in *Proc. ICLR*, 2023.
- [34] R. Yoneyama, Y.-C. Wu, and T. Toda, “Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [35] K. Song, Y. Zhang, Y. Lei, J. Cong, H. Li, L. Xie, G. He, and J. Bai, “DSPGAN: A Gan-Based Universal Vocoder for High-Fidelity TTS by Time-Frequency Domain Supervision from DSP,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [36] F. Villavicencio and J. Bonada, “Applying voice conversion to concatenative singing-voice synthesis,” in *Proc. Interspeech*, 2010.
- [37] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, “Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system,” in *Proc. APSIPA ASC*, 2012, pp. 1–6.
- [38] J. Lu, K. Zhou, B. Sisman, and H. Li, “VAW-GAN for singing voice conversion with non-parallel training data,” in *Proc. APSIPA ASC*, 2020, pp. 514–519.
- [39] J. Shi, S. Guo, T. Qian, N. Huo, T. Hayashi, Y. Wu, F. Xu, X. Chang, H. Li, P. Wu, S. Watanabe, and Q. Jin, “Muskits: an end-to-end music processing toolkit for singing voice synthesis,” in *Proc. Interspeech*, 2022, pp. 4277–4281.
- [40] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, “Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining,” *Proc. Interspeech*, pp. 4676–4680, 2020.
- [41] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, “Pretraining techniques for sequence-to-sequence voice conversion,” *IEEE/ACM TASLP*, vol. 29, pp. 745–755, 2021.
- [42] D. G. Rajpura, J. Shah, M. Patel, H. Malaviya, K. Phatnani, and H. A. Patil, “Effectiveness of transfer learning on singing voice conversion in the presence of background music,” in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, 2020, pp. 1–5.
- [43] G. Roa Dabike and J. Barker, “Automatic lyric transcription from karaoke vocal tracks: Resources and a baseline system,” in *Proc. Interspeech*, 2019.
- [44] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplín, J. Heymann, M. Wiesner, N. Chen, et al., “ESPhet: End-to-end speech processing toolkit,” *Proc. Interspeech*, pp. 2207–2211, 2018.
- [45] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, et al., “Recent developments on espnet toolkit boosted by conformer,” in *Proc. ICASSP*, 2021, pp. 5874–5878.
- [46] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhota, 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.
- [47] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *arXiv preprint arXiv:2212.04356*, 2022.
- [48] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, “Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders,” in *Proc. ICASSP*, 2020, pp. 3277–3281.
- [49] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, “Pushing the limits of raw waveform speaker recognition,” *Proc. Interspeech*, 2022.
- [50] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, “Generalization ability of mos prediction networks,” in *Proc. ICASSP*, 2022, pp. 8442–8446.
- [51] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “UTMOS: UTokyo-SaruLab System for Voice-MOS Challenge 2022,” in *Proc. Interspeech*, 2022, pp. 4521–4525.
- [52] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, et al., “Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio,” *arXiv preprint arXiv:2106.06909*, 2021.
- [53] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, “OpenCPop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis,” in *Proc. Interspeech*, 2022, pp. 4242–4246.
- [54] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, “Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus,” in *Proc. ACM MM*, 2021, pp. 3945–3954.
- [55] C. Veaux, J. Yamagishi, K. MacDonald, et al., “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.
- [56] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, “The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech,” in *Proc. APSIPA ASC*, 2013, pp. 1–9.
- [57] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen, and Z. Zhao, “M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus,” in *Proc. NeurIPS: Datasets and Benchmarks Track*, 2022.
- [58] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [59] R. Yoneyama, Y.-C. Wu, and T. Toda, “Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation,” in *Proc. Interspeech*, 2022, pp. 848–852.
- [60] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, “A comparison of discrete and soft speech units for improved voice conversion,” in *Proc. ICASSP*, 2022.
- [61] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, “HiFiSinger: Towards high-fidelity neural singing voice synthesis,” *arXiv preprint arXiv:2009.01776*, 2020.

A. LISTENER DETAILS

We recruited 40 unique English listeners (17 female, 22 male, and 1 unknown), and Figure 5 shows the accent and age distributions of the English and Japanese listeners. Half of the English participants were in their 30s or 40s, and most of them had an American accent. For Japanese listeners, we had a total of 319 unique valid listeners (162 male and 157 female). Figure 5 also shows that most of the Japanese listeners were in their 30s or 40s.

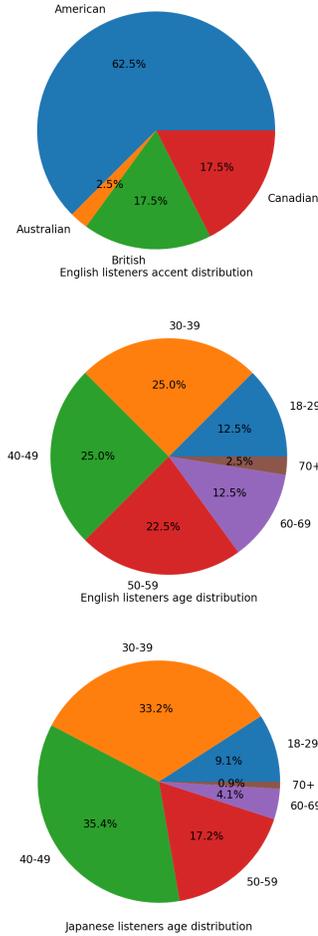


Fig. 5: Age and accent distribution of English and Japanese listeners.

B. EVALUATION RESULTS FROM JAPANESE LISTENERS

B.1. Naturalness

Figure 6 shows the boxplot of the naturalness evaluation results of Tasks 1 and 2 from the Japanese listeners. In both tasks, baseline B01 was outperformed by around half of the teams. T23 was the top system in both tasks. Surprisingly, different from the finding from the English listener results that no team was on average better than the natural samples (TAR, SOU), three teams (T23, T07, T02) and one team (T23) received a naturalness score higher than the natural samples in Tasks 1 and 2, respectively. Furthermore, the pairwise

significance test results in Figures 9a and 9c show that the natural samples are not significantly different with six teams (T23, T07, T02, T06, T14, T20) and two teams (T23, T02) in Tasks 1 and 2, respectively. These findings are in line with our finding that Japanese listeners tend to give higher scores than those given by English listeners. Finally, we can also observe that the scores received in Task 2 are generally lower than those received in Task 1, again showing that cross-domain SVC is indeed harder than in-domain SVC.

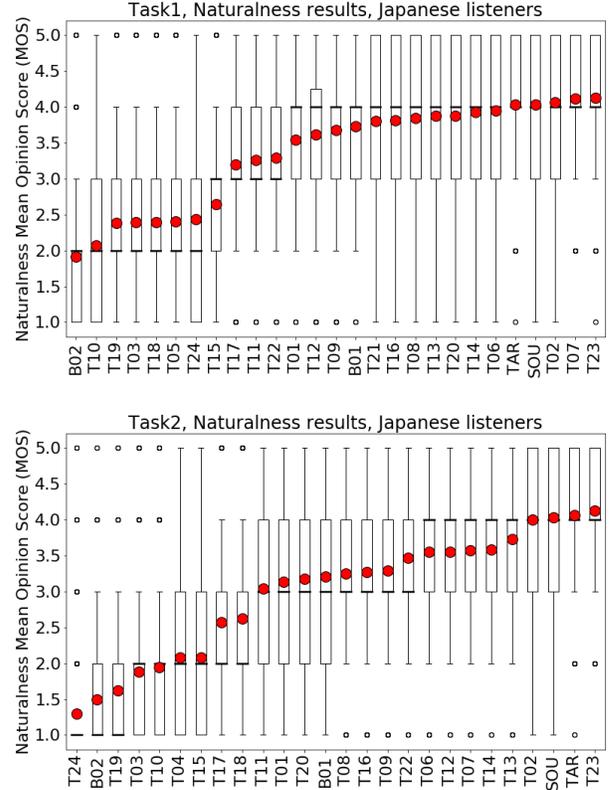


Fig. 6: Japanese listeners’ naturalness results for Tasks 1 and 2. MOS scores are arranged in accordance with their mean (red dot). SOU and TAR represent the source and target samples, respectively.

B.2. Similarity

Figure 7 shows the results for the similarity evaluation results of Tasks 1 and 2 from the Japanese listeners. Again, the similarity percentage is defined as the sum of the percentages from the “same (not sure)” and “same (sure)” categories, and the averaged scores are also shown. The baseline B01 received a stronger ranking from Japanese listeners, ranking second in both tasks. The top system in Task 1, T14, ranked third in Task 2, while the top system in Task 2, T02, ranked fourth in Task 1. Similar to English listeners’ results, there is also a clear gap (around 0.4 points) between the target samples and the top system in both tasks. This can also be observed from Figures 9b and 9d, which show that the target samples were significantly better than all other systems in terms of similarity. The conclusion is therefore similar to that of the English listeners: there is still a lot to work on for similarity.

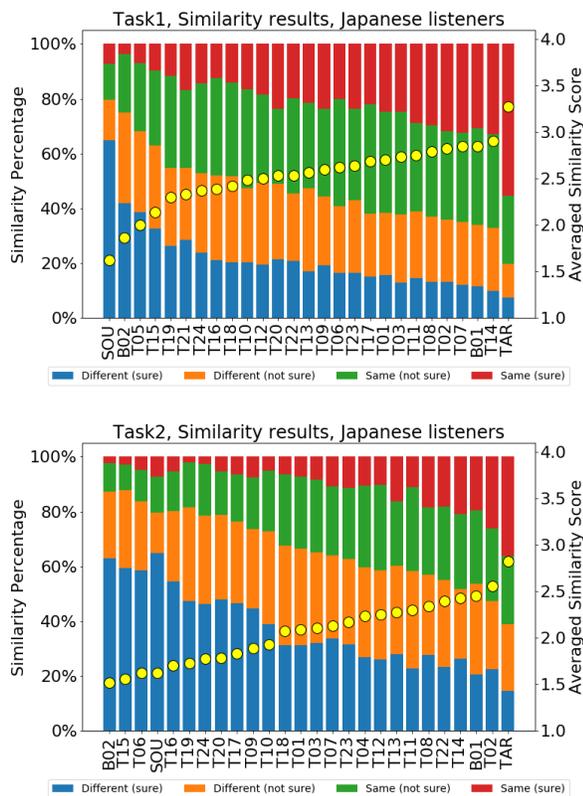


Fig. 7: Japanese listeners similarity results for Tasks 1 and 2. Similarity scores are arranged in accordance with their mean value (red dot). SOU and TAR represents the source and target samples, respectively.

Figure 8 shows the scatter plots of naturalness and similarity percentage of both tasks from Japanese listeners. Similar to English listeners results, there is a trade-off between naturalness and similarity for most systems, i.e. no team is dominant in both naturalness and similarity. Again, all teams need to improve either similarity or naturalness.

C. COMPARISON BETWEEN ENGLISH AND JAPANESE LISTENERS

Figure 10 shows the scatter plots from Japanese and English listeners, and it can be observed that Japanese listeners tend to give higher scores in naturalness, and English listeners tend to give higher scores in similarity.

We made a hypothesis in Section 5.3 that the larger the number of scores, the easier it is to observe statistically significant differences between systems, which means fewer red dots should be observed in Figures Figure 9 and 3. However, by comparing Figure 9 (Japanese listeners) and Figure 3 (English listeners), we observed that while this hypothesis somehow stands for similarity, it is surprising to see that for naturalness, English listeners can reach a similar level of confidence with only one-thirds of scores.

We further examine whether the above-mentioned hypothesis implies the following statement: the larger the number of scores, the

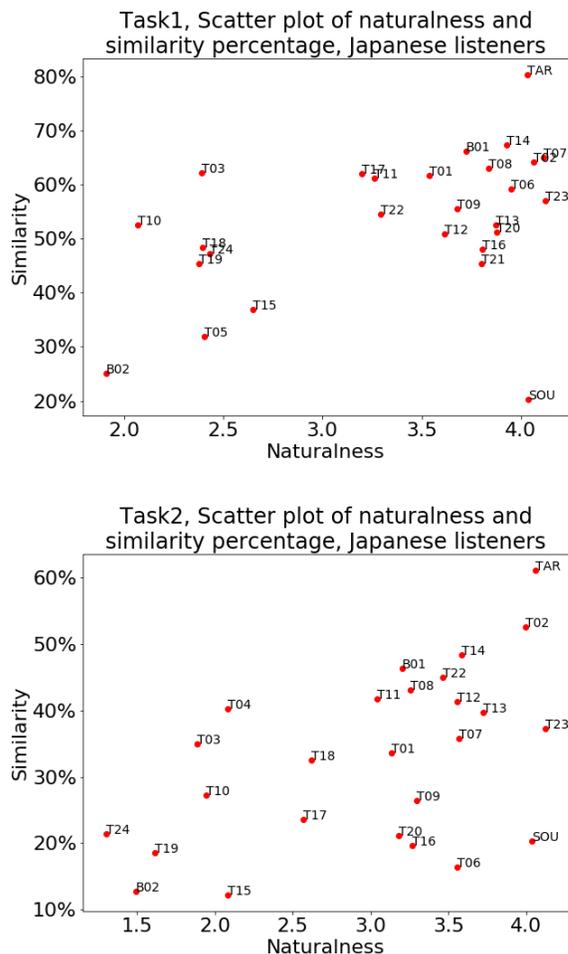


Fig. 8: Scatter plots of naturalness and similarity percentage for task 1 (in-domain) and task 2 (cross-domain), from Japanese listeners.

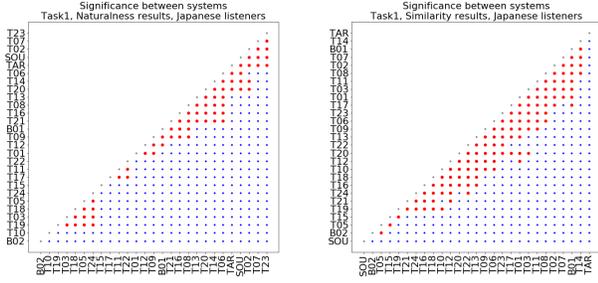
smaller the system-level variance. We therefore plotted the system-level variance from Japanese and English listeners in Figure 11. However, we did not observe any obvious tendency, thus the above-mentioned statement was not implied in this challenge.

D. DETAILS OF THE B01 DIFFSVC BASELINE SYSTEM

The first baseline system is much similar to the system presented in the DiffSVC paper [23]. We use a different PPG model, which is a Conformer-based phoneme recognizer containing 7 conformer blocks. The encoder dimension is 256. In total, the PPG model contains 31.2 million trainable parameters. The training data is a combination of a random half from the WenetSpeech dataset (Mandarin Chinese)⁸ and a random half from the GigaSpeech dataset (English) [52], which in total has 10k hours speech data. We take the feature from the last hidden layer as the content feature.

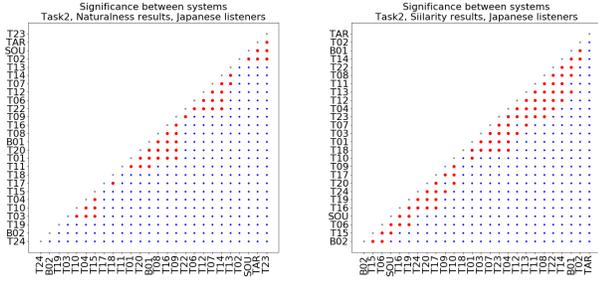
The PPG-to-Mel-spectrogram model has the same network structure as that presented in [23]. We extend the model to sup-

⁸<https://wenet.org.cn/WenetSpeech>



(a) Task 1, naturalness

(b) Task 1, similarity



(c) Task 2, naturalness

(d) Task 2, similarity

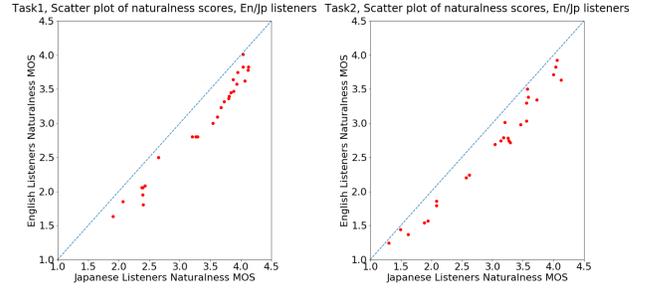
Fig. 9: Japanese listeners pairwise significance between systems, calculated with Wilcoxon signed-rank tests. Blue dots: significantly different; Red dots: no significant difference.

port multi-singer generation by adding a speaker/singer embedding vector to every residual block. The training set is a mixture of the SVCC 2023 dataset, OpenCPOP dataset [53], MultiSinger [54], VCTK [55], NUS-48N [56] and M4Singer [57]. In total, the training set contains 116 hours of speech and singing data from 221 speakers or singers. We do not conduct any finetuning procedure for the target singers and use the multi-singer model directly for evaluation. We use a HiFi-GAN V1 [31] to convert the generated Mel spectrogram to a waveform, which is trained with the same training set.

During conversion, for the task of in-domain SVC (i.e., Task 1), we shift the source pitch by multiplying a ratio, which is computed as the ratio of the median of the target pitch and the median of a source phrase. For the task of cross-domain SVC (i.e., Task 2), we shift the source pitch down by an octave in female-to-male conversion and shift the source pitch up by an octave in male-to-female conversion, respectively.

E. DETAILS OF THE B02 DECOMPOSED FASTSVC BASELINE SYSTEM

For the acoustic model, we use Tacotron 2 [58] encoder, along with an autoregressive decoder due to its success in [10]. The Tacotron 2 encoder consists of two stacks of one-dimension convolutional layers and a bidirectional long short-term memory (BLSTM) layer. On the other hand, the decoder is an autoregressive network, due to its proven ability in the previous challenge [10]. To implement the autoregressive loop, the previous output is consumed by the first long short-term memory with projection (LSTMP) layer at each time step. The acoustic model predicts the concatenated mel-cepstral coefficients (mcep) and band-aperiodicity (bap), which are used as inputs of the hn-uSFGAN vocoder. For the vocoder, we use HN-



(a) Naturalness

(b) Similarity

Fig. 10: Scatter plots of scores from Japanese listeners and English listeners.

uSFGAN [59] as is due to its ability to synthesize waveforms outside the training pitch range.

The network is trained with the SVCC 2023 dataset, along with the large-scale speech dataset VCTK [55], and large-scale singing datasets M4Singer [57], MultiSinger [54], OpenCPOP [53], and NUS-48E [56]. To handle the multilingual datasets, we replace the PPG encoder with HuBERT soft features due to its proven ability in cross-lingual VC [60]. To optimize the acoustic model, we use two loss functions: 1) an L2 reconstruction loss and 2) a sub-frequency discriminator, which was introduced in [61], to improve the predicted mcep/bap features. To shift the pitch, we use linear transformation by using the mean-variance transformation.

F. BREAKDOWN USING DIFFERENT TECHNIQUES

Although we mentioned in Sec. 4.3.2 that the goal of the taxonomy analysis is not to derive meaningful tendencies or scientific differences, we still tried to find certain techniques that contribute to a high performance. In light of this, we created variants of the scatter plot in Figure 1 by coloring each system with the technique used in that system. The results are shown in Figure 12.

We did not find particular trends for the content feature, vocoder, and the use of *so-vits-svc*. We would like to emphasize that, despite the seeming success of *so-vits-svc*, SVC systems based on that toolkit did not necessarily perform better. On the other hand, many of the VAE-based systems had high rankings in Task 1, which somehow shows that VAE can be a promising framework for SVC.

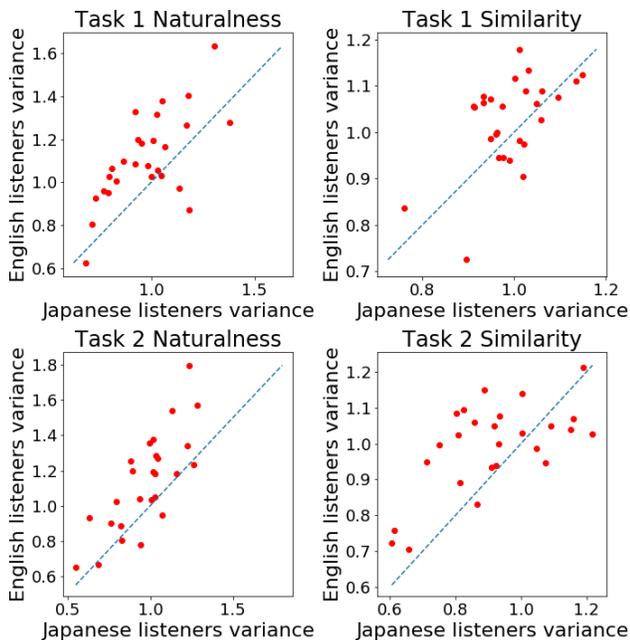


Fig. 11: Scatter plots of system-level variance from Japanese and English listeners.

G. DETAILED OBJECTIVE EVALUATION RESULTS AND ANALYSIS

Detailed results of the objective evaluation for each team can be found in Table 5 and Table 6. In general, it is challenging to identify a universally accepted objective measure that correlates strongly with subjective evaluation. This observation is consistent with the findings presented in Table 4 and Table 7. To further assess performance, we conducted linear regression modeling to examine the impact of different objective measures.

For Task 1, we excluded variables that exhibited high collinearity based on variance inflation factors (VIF). However, for Task 2, we included all variables since no strong collinearity was observed among the different factors. Nonetheless, as shown in Table 8 and Table 9, although the R-squared values are relatively high, there is limited consensus across different metrics and listening types. One possible reason for the unsuccessful regression could be the limited sample size used in the analysis. But at the same time, these results emphasize the need for further research to identify effective objective evaluation metrics for the challenging SVC task.

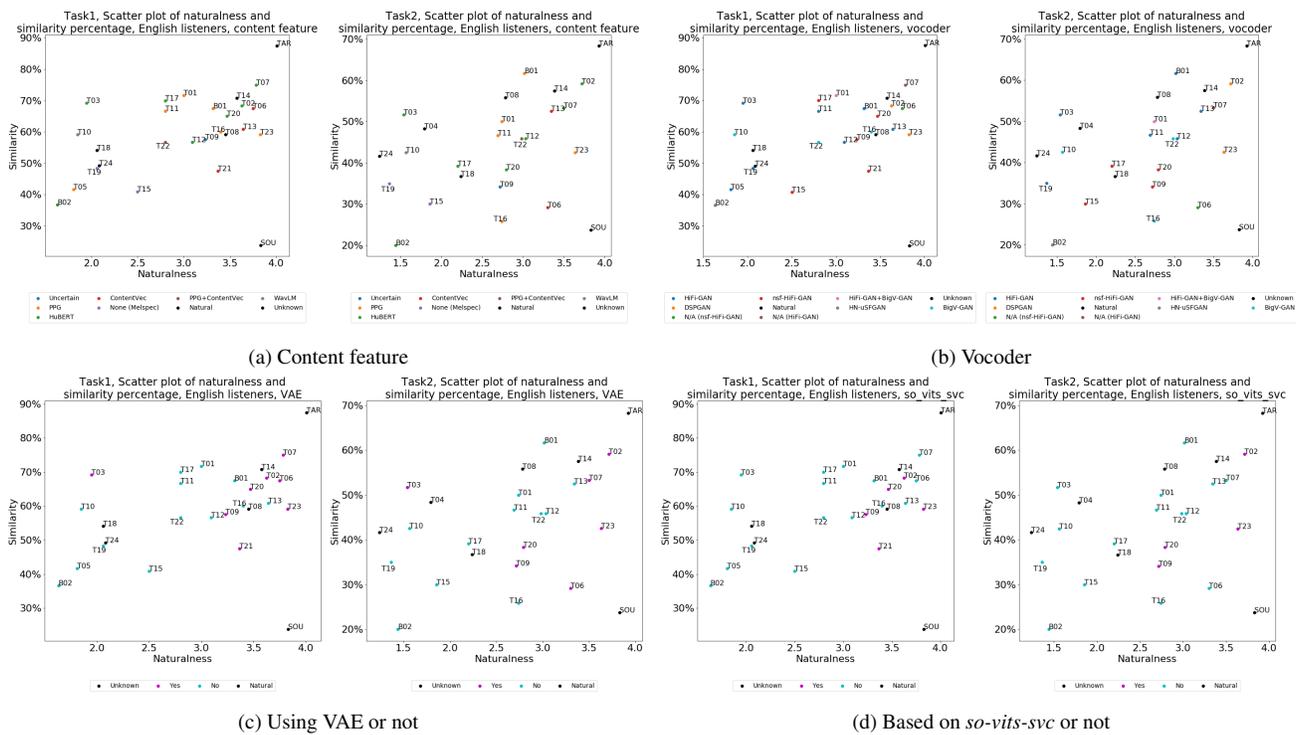


Fig. 12: Scatter plots of naturalness and similarity in tasks 1 and 2 from English listeners, colored on basis of different techniques.

Table 5: Detailed objective evaluation results for each team in Task 1.

Team ID	MCD(↓)	F0RMSE(↓)	F0CORR(↑)	CER(Conformer)(↓)	CER(Whisper)(↓)	D_{Embed} (↓)	UTMOS(↑)	SSL-MOS(↑)
B01	11.856	70.018	0.655	34.8	23.6	0.734	1.595	0.784
B02	10.409	59.028	0.664	31.3	18.5	0.454	2.028	1.072
T01	11.880	59.165	0.633	27.1	16.2	0.355	2.399	1.162
T02	8.524	56.187	0.722	29.8	16.1	0.439	2.345	1.112
T03	9.287	78.748	0.599	38.4	33.3	0.419	1.893	1.080
T05	15.867	66.427	0.671	53.4	62.0	0.647	1.917	1.275
T06	11.206	54.839	0.676	29.0	18.1	0.379	2.526	1.173
T07	9.427	60.757	0.680	27.7	15.0	0.357	2.420	1.193
T08	9.333	57.048	0.716	28.5	18.4	0.446	2.070	1.081
T09	12.155	73.416	0.587	35.0	21.4	0.477	2.598	1.145
T10	12.307	65.586	0.679	36.0	30.4	0.560	1.573	0.976
T11	9.160	67.867	0.671	28.8	19.1	0.428	2.246	1.113
T12	12.622	61.391	0.667	28.5	18.3	0.494	2.101	1.149
T13	10.111	68.034	0.692	30.0	18.4	0.464	2.228	1.183
T14	9.762	75.901	0.686	28.7	15.6	0.448	2.057	0.999
T15	10.941	61.391	0.601	34.2	25.3	0.624	1.631	0.963
T16	10.299	55.229	0.707	24.7	12.6	0.537	2.657	1.201
T17	9.454	62.937	0.657	33.3	23.1	0.427	2.082	1.109
T18	14.136	75.119	0.616	30.2	18.9	0.515	2.061	1.040
T19	10.606	75.214	0.673	33.2	23.8	0.544	1.758	0.979
T20	14.229	99.252	0.643	29.9	16.2	0.484	2.313	0.971
T21	12.361	98.740	0.611	31.2	16.0	0.469	2.174	1.116
T22	11.324	57.878	0.692	26.8	15.4	0.523	2.037	1.048
T23	11.536	57.784	0.678	26.8	14.5	0.423	2.717	1.323
T24	11.730	94.873	0.531	29.7	19.2	0.508	1.578	0.870

Table 6: Detailed objective evaluation results for each team in Task 2.

Team ID	MCD(↓)	F0RMSE(↓)	F0CORR(↑)	CER(Conformer)(↓)	CER(Whisper)(↓)	D_{Embed} (↓)	UTMOS(↑)	SSL-MOS(↑)
B01	12.495	85.693	0.243	36.3	25.0	0.761	1.679	1.023
B02	11.835	51.866	0.478	33.9	22.7	0.552	1.893	1.092
T01	12.218	52.841	0.391	25.9	26.7	0.501	2.468	1.405
T02	10.278	64.436	0.254	30.5	15.5	0.551	2.415	1.223
T03	10.608	64.821	0.205	39.6	30.1	0.560	1.971	1.242
T04	12.317	63.237	0.302	31.3	20.2	0.630	1.811	0.979
T06	10.651	82.362	0.356	28.1	21.4	0.577	2.870	1.792
T07	11.034	58.759	0.277	27.8	14.6	0.525	2.383	1.220
T08	10.188	66.720	0.355	29.3	23.4	0.592	2.199	1.379
T09	12.448	69.065	0.322	35.5	24.5	0.584	2.715	0.969
T10	14.236	63.071	0.336	36.6	28.3	0.651	1.840	1.282
T11	10.642	53.160	0.373	29.6	20.1	0.544	2.159	1.093
T12	13.281	84.585	0.351	28.5	16.4	0.576	2.084	1.178
T13	11.498	66.228	0.358	32.5	21.2	0.578	2.424	1.518
T14	11.863	65.375	0.248	27.6	15.2	0.508	2.076	1.041
T15	13.331	84.585	0.284	38.8	29.0	0.776	1.986	1.476
T16	10.267	78.880	0.368	25.8	15.9	0.652	2.964	1.640
T17	10.654	81.299	0.281	30.2	28.4	0.617	2.540	1.645
T18	13.983	70.474	0.260	29.1	19.8	0.566	2.416	1.200
T19	13.349	69.981	0.386	36.6	31.4	0.694	1.785	1.315
T20	14.213	71.723	0.288	29.0	21.8	0.594	2.576	1.152
T22	12.314	60.867	0.252	27.1	15.4	0.549	1.966	1.020
T23	12.365	63.348	0.264	27.2	15.3	0.534	2.675	1.331
T24	13.236	90.349	-0.013	34.9	28.0	0.585	1.457	1.037

Table 7: Pearson correlation between objective and subjective metrics. Red highlights indicate the highest correlation with corresponding subjective metrics among the objective metrics. CER metric refers to Conformer-based speech recognition results, while CER+ refers to Whisper results. Significance levels are shown by *.

Sub. Score	Listener	MCD	F0RMSE	F0CORR	CER	CER+	D_{Embed}	UTMOS	SSL-MOS
Task 1 MOS	JPN	-0.33	-0.23	0.41**	-0.57***	-0.58***	-0.68***	0.82***	0.58***
	ENG	-0.28	-0.15	0.39*	-0.55***	-0.57***	-0.36*	0.66***	0.30
Task 1 SIM	JPN	-0.59***	-0.19	0.27	-0.51**	-0.47**	-0.89***	0.51**	0.35*
	ENG	-0.43**	-0.16	0.27	-0.41**	-0.40**	-0.46**	0.37*	0.04
Task 2 MOS	JPN	-0.39*	-0.37*	0.37*	-0.71***	-0.77***	-0.64***	0.69***	0.18
	ENG	-0.37*	-0.10	0.10	-0.65***	-0.75***	-0.38*	0.59***	0.14
Task 2 SIM	JPN	-0.30	-0.69***	0.17	-0.35*	-0.53***	-0.71***	-0.06	-0.36*
	ENG	-0.23	-0.20	-0.26	-0.14	-0.25	-0.27	-0.19	-0.32

Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 8: Task 1 linear regression models over subjective metrics with objective metrics as inputs. Highlights in orange are coefficients with statistical significance. CER metric refers to Conformer-based speech recognition results.

Sub. Score Listener	Task 1 MOS		Task 1 SIM	
	JPN	ENG	JPN	ENG
Intercept	3.211***	-3.227	-1.143	2.478*
MCD	-0.266	-0.094	-0.024	-0.034
FORMSE	0.002	0.016	0.010	0.002
F0CORR	1.196	4.122	4.435*	1.004
CER	-1.757e-4	-0.037	-0.021	-0.005
D_{Embed}	-2.422***	2.637	-1.758	-0.540
UTMOS	-0.054	1.531***	1.241***	0.104
R^2	0.855	0.634	0.794	0.332
Adjust R^2	0.807	0.511	0.726	0.109
F Significance	<1e-3	0.003	<1e-3	0.237

Significance Levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 9: Task 2 linear regression models over subjective metrics with objective metrics as inputs. Highlights in orange are coefficients with statistical significance. CER metric refers to Conformer-based speech recognition results, while CER+ refers to Whisper results.

Sub. Score Listener	Task 1 MOS		Task 1 SIM	
	JPN	ENG	JPN	ENG
Intercept	3.376**	3.702	4.930***	4.771***
MCD	0.138	-0.071	-0.033	-0.068
FORMSE	0.006	0.002	-0.006	-0.005
F0CORR	2.392	-0.571	0.652	-0.884
CER	0.049	-0.006	0.018	-0.014
CER+	-0.099***	-0.085**	-0.030**	-0.002
D_{Embed}	-5.179**	0.690	-2.517***	0.230
UTMOS	0.609*	0.683	-0.363**	-0.154
SSL-MOS	0.501	-0.068	0.086	-0.235
R^2	0.887	0.678	0.828	0.343
Adjust R^2	0.827	0.506	0.736	-0.007
F Significance	<1e-3	0.011	<1e-3	0.488

Significance Levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$