Interspeech 2024
1-5 September 2024, Kos, Greece

SRC4VC: Smartphone-Recorded Corpus for Voice Conversion Benchmark

Yuki Saito[1], *Takuto Igarashi*[1], *Kentaro Seki*[1], *Shinnosuke Takamichi*[1,2], *Ryuichi Yamamoto*[3], *Kentaro Tachibana*[3], *Hiroshi Saruwatari*[1]

[1]The University of Tokyo, Japan, [2]Keio University, Japan, [3]LY Corp., Japan.

yuuki_saito@ipc.i.u-tokyo.ac.jp

Abstract

We present *SRC4VC*, a new corpus containing 11 hours of speech recorded on smartphones by 100 Japanese speakers. Although high-quality multi-speaker corpora can advance voice conversion (VC) technologies, they are not always suitable for testing VC when low-quality speech recording is given as the input. To this end, we first asked 100 crowdworkers to record their voice samples using smartphones. Then, we annotated the recorded samples with speaker-wise recording-quality scores and utterance-wise perceived emotion labels. We also benchmark SRC4VC on any-to-any VC, in which we trained a multi-speaker VC model on high-quality speech and used the SRC4VC speakers' voice samples as the source in VC. The results show that the recording quality mismatch between the training and evaluation data significantly degrades the VC performance, which can be improved by applying speech enhancement to the low-quality source speech samples.

Index Terms: speech corpus, smartphone-recorded, crowdsourcing, annotation, voice conversion

1. Introduction

Voice conversion (VC) is a technology for transforming the voice characteristics of source speech into those of target speech while keeping the phonetic content of the source speech unchanged [1]. VC enables expressions beyond the physical constraints of the human voice and enriches speech communication through social applications, such as speech-to-speech translation [2] and customer service [3]. As a result of developments in machine learning techniques [4, 5, 6] and well-designed speech corpora [7, 8], deep neural network (DNN)-based VC [9], which trains a VC model on a speech corpus including multiple speakers' voice samples, has become the mainstream of VC research and improved the quality of converted voices, an essential factor in VC performance evaluation.

The robustness of a trained VC model towards degraded speech input (i.e., degradation robustness [10]) is another crucial factor in real-world VC applications. Namely, model must be capable of high-quality VC even if the source speech, target speech, or both are degraded due to the recording environment and transmission channel. The degradation robustness can be improved by using artificially generated noisy speech when training the VC model or by cascading speech enhancement and VC models in the inference [11]. However, conventional degradation-robust VC work used only artificially degraded speech for VC evaluation. One primary reason is the lack of publicly available speech corpora recorded by devices owned by end-users with adequate annotations, which hinders validating the degradation robustness of state-of-the-art VC technologies.

To facilitate research on degradation-robust VC models, we present *SRC4VC* (**S**martphone-**R**ecorded **C**orpus for (**4**) **V**oice **C**onversion), a new multi-speaker speech corpus for validating the degradation robustness of a VC model. The corpus consists of 11 hours of smartphone-recorded speech uttered by 100 crowdsourced Japanese speakers. Through crowdsourcing, the recorded samples were annotated with speaker-wise recording-quality scores and utterance-wise perceived emotion labels to the recorded samples. This design makes it possible to evaluate the robustness of a DNN-based VC model towards actual degraded speech input. We also benchmark SRC4VC on any-to-any VC, in which a multi-speaker VC model is trained on high-quality speech and used the SRC4VC speakers' voice samples as the source speech in VC. Our contributions are as follows:

- We construct a new, publicly available corpus designed for evaluating the performance of VC from end-users' source speech input. Our corpus is open-sourced for research purposes only from `https://y-saito.sakura.ne.jp/sython/Corpus/SRC4VC/index.html`.
- We analyze the smartphone-recorded speech samples and discuss the results of speech quality assessments using DNNs and crowdsourcing.
- We present the results of an any-to-any VC experiment and show that, for VC using an end-user's voice as the source speech, the recording-quality mismatch can be addressed by simply introducing speech enhancement as preprocessing, rather than training a VC model with data augmentation.

2. Construction of SRC4VC

2.1. Core design

We designed SRC4VC so that it includes diverse voice samples from various actual smartphones and advances various VC tasks, such as emotional VC [12] and singing VC [13]. Specifically, the corpus includes 100 speakers who each recorded 52 voice samples categorized into the following four subsets: 10 read-aloud, 30 expressive, 10 conversational, and two singing samples. To examine non-parallel any-to-any VC, we randomly selected sentences for each speaker to record from the subsets except for singing.

Read-aloud subset: We randomly selected 10 sentences per speaker from the Recitation324 subset of ITA [14], which contains 324 phonetically balanced sentences. This subset aims to record neutral, reading-style samples.

Expressive subset: We used JVNV [15], which covers six emotions (*Angry, Disgust, Fear, Happy, Sadness, Surprise*). We randomly selected five sentences for each emotion category and asked the speakers to read the presented sentences while expressing the corresponding emotion.

Conversational subset: We used STUDIES [16] and CALLS [17] consisting of teacher-student and operator-customer dialogues, respectively. Although these corpora contain human-annotated emotion labels, we did not present the la-

10.21437/Interspeech.2024-388

bels to the speakers because the ground-truth emotion for actual conversation generally cannot be defined. Instead, we allowed the speakers to express their own emotions by interpreting the presented sentences.

Singing subset: We used two Japanese copyright-free songs: "katatsumuri" (child-song) and "Shining star[1]" (J-POP). We asked the speakers to sing only the first chorus of these two songs because singing entire songs without retakes is difficult for non-professionals.

2.2. Voice recording by crowdworkers

We crowdsourced smartphone-recorded voice samples using the above-mentioned four subsets. We collected the samples by macrotask crowdsourcing [18], where crowdworkers are employed by clients taking the workers' skills and experiences into consideration. We describe the recording process in detail below.

Preparing voice recording platform: We created a webpage for voice recording that can be accessed via a smartphone. The webpage contained instructions for recording for each subset, a recording start/stop button, and text with the pronunciation of the words to be spoken. The recording function was implemented using Web Audio API[2].

Recording speakers' voice samples: We recruited speakers through crowdsourcing and required them to understand the purpose of this project (i.e., collecting smartphone-recorded voice samples), have a smartphone, and record their voice in a quiet environment. We asked the speakers to record their voice using smartphones and our web-based recording platform. The speakers recorded 52 samples under the four subsets described in Section 2.1 in their own rooms. They could listen to reference singing voice samples when recording the "Singing" subset. We allowed the speakers to rerecord their voice if they made a mistake in the recording. After the recording process, the speakers submitted their recordings to a web server owned by the authors. The sampling rate was 48 kHz. With our manual validation of recorded data, we asked some of the speakers to rerecord misrecorded samples.

2.3. Annotation

We crowdsourced speaker-wise recording quality scores and utterance-wise perceived emotion labels.

Speaker-wise recording quality scores: We conducted recording quality rating of the collected voice samples and constructed the annotation sets as follows. Allowing sample overlap among sets, we made 400 sets containing 25 read-aloud utterances for each. The final speaker-wise recording quality scores were calculated by taking the average of 100 obtained scores for each speaker.

Utterance-wise perceived emotion labels: We conducted emotion classification tests of the recorded "Expressive" and "Conversational" subsets. The purpose was to investigate how well the speakers recruited by crowdsourcing were able to express the instructed or situation-oriented emotion in their speech. For each of the 30 expressive and 10 conversational samples, five different crowdworkers annotated the perceived emotion from seven options (i.e., the six emotions and *Neutral*).

[1] https://maou.audio/14_shining_star/ (in Japanese)
[2] https://webaudioapi.com/

Table 1: *Crowdsourcing setup: the numbers of recruited crowdworkers and the amounts of paid rewards (per crowdworker). "Rec," "Emo," "expr," and "conv" in this table denote recording, emotion, expressive, and conversational, respectively.*

Task	# workers	Rewards
Voice rec.	100	$33.2
Rec quality MOS test	400	$0.44
Emo labeling (expr)	500	$1.31
Emo labeling (conv)	500	$0.44

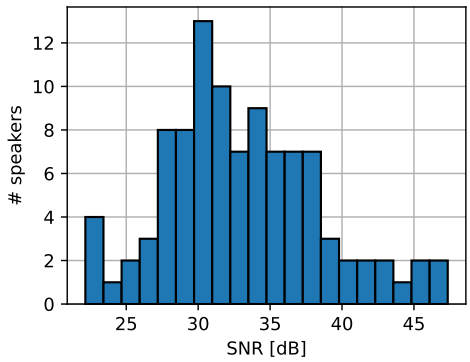

Figure 1: *Histogram of SNRs averaged over speakers.*

3. Corpus Analysis

3.1. Crowdsourcing setting

We used Lancers[3] as the crowdsourcing platform. The overall period of the voice recording ran 2023. Table 1 lists the number of crowdworkers and the amounts of paid rewards per crowdworker. The recording quality ratings were conducted on the basis of five-scale mean opinion score (MOS) test on a scale of 1 (very bad) to 5 (very good).

3.2. Corpus specification

Our SRC4VC consists of 1,000 read-aloud, 3,000 expressive, 1,000 conversational, and 200 singing voice samples recorded by 100 speakers, the total durations of which are 1.46, 7.16, 1.66, and 0.87 hours, respectively.

Figure 1 shows the histogram of averaged priori signal-to-noise ratio (SNR) per speaker. We estimated SNRs based on the Ephraim-Malah algorithm [23] implemented in Essentia [24]. The minimum, median, and maximum values of the estimated SNRs with the speaker IDs are 22.2 dB (044), 32.5 dB (017), and 47.3 dB (064), respectively. These results indicate that our SRC4VC covers wider SNR range (approximately 25 dB) of recorded samples.

3.3. Speaker distribution

We discuss the distributions of SRC4VC speakers regarding gender, age, locale, and used device. The ratio of male to female speakers was 37 to 63, and the histogram of speaker ages is shown in Figure 2. The youngest and oldest ages with the speaker IDs were 21 (085) and 68 (009), respectively. Figure 3 maps the speakers' locales. Our SRC4VC covers 34 of 47 prefectures in Japan as the speaker locales. Sixty-one percent of the speakers used an iPhone as their recording device, with models ranging from SE to 15 Pro. These results demonstrate that our SRC4VC corpus consists of speech recorded by diverse speakers on various smartphones.

[3] https://www.lancers.jp/

Table 2: *List of existing open-sourced multi-speaker speech corpora related to SRC4VC. "Lang." and "Dur." denote covered languages and total durations in hours, respectively.*

Corpus	Speaking styles	Lang.	Dur.	# of speakers	Recording quality
VCTK [7]	Reading w/ various accents	En	44	109	Studio-recorded
DDS [19]	Reading (from VCTK [7] and DAPS [20])	En	2000	48	Device-rerecorded
JVS [21]	Reading, Whisper, Falsetto	Ja	30	100	Studio-recorded
CPJD [22]	Reading w/ dialects	Ja	7	22	Device-recorded
SRC4VC	Reading, Expressive, Conversational, Singing	Ja	11	100	Smartphone-recorded

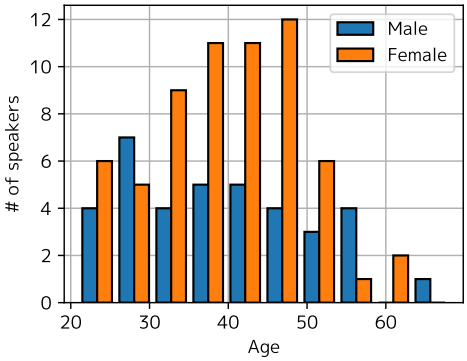

Figure 2: *Histogram of speaker ages.*

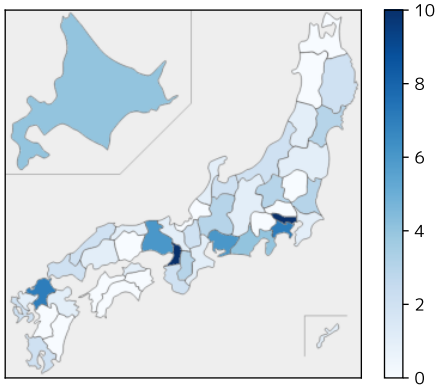

Figure 3: *Number of speakers from each prefecture in Japan. The map chart was drawn using the japanmap library.*

3.4. Comparison with existing corpora

Table 2 lists existing open-sourced multi-speaker speech corpora related to SRC4VC. Although SRC4VC is smaller than existing de-facto standard corpora for developing high-quality multi-speaker VC technologies, such as VCTK [7] and JVS [21], it covers various speech degradations that rarely occur in controlled and supervised studio recording. DDS [19] includes aligned parallel high-quality speech recordings taken from VCTK [7] and DAPS [20] and their rerecorded versions with various conditions by combining diverse acoustic environments and devices. In contrast, SRC4VC contains actual end-users' voice samples recorded in their own acoustic environments using smartphones. The design of SRC4VC was inspired by that of CPJD [22], in which the developers collected parallel speech data of Japanese dialects by crowdsourcing. As discussed in Section 3.3, our SRC4VC covers speakers from a wider area than those of CPJD (21 dialects).

3.5. Annotation results

We analyze the results of the annotations for recording quality and emotion labels.

Table 3: *SRCC between human-annotated recording quality and each NISQA score. "Nat." denotes Naturalness.*

Noisiness	Coloration	Discontinuity	Loudness	Nat.
0.15	**0.67**	**0.62**	0.36	0.54

Table 4: *Percentages of agreed emotional utterances in "Expressive" (expr) and "Conversational" (conv) subsets.*

Subset	Ang	Dis	Fea	Hap	Sad	Sur	Neu
expr	14.6	17.8	14.4	16.7	15.7	17.3	0.35
conv	0.45	0.29	0.08	0.59	0.56	0.28	1.08

Recording quality scores: We computed the Spearman's rank correlation coefficient (SRCC) between the human-annotated recording quality MOS values and NISQA scores [25], which quantify the quality of input speech in terms of *Noisiness, Coloration, Discontinuity, and Loudness* [26, 27], in addition to *Naturalness*. Table 3 lists the results. The "Coloration" and "Discontinuity" scores moderately correlated with the human-annotated MOS values. These results indicate that dominant speech degredation factors in SRC4VC are frequency response and non-linear distortion.

Emotion labels: We investigated the inter-annotator agreement of the perceived emotion labels. Specifically, we calculated the percentages of "agreed emotional utterances." Namely, if more than two annotators assigned the same perceived emotion label to one utterance, we regarded it as the agreed emotional utterance. Table 4 lists the results. The distribution of agreed emotion labels for the "Expressive" subset was close to uniform except for "Neutral." In contrast, the percentage of agreed emotion labels for the "Conversational" subset was generally low, with the highest percentage of 1% for the "Neutral" emotion. These results suggest that the SRC4VC speakers could express and utter the indicated emotion to a certain extent, though they were not necessarily professional actors.

4. Any-to-Any VC Experiment

4.1. Experimental conditions

Our experiment aims to investigate whether the recording-quality mismatch between the training data and evaluation data actually worsens VC performance and what kind of approaches can mitigate the mismatch.

Datasets: We used JVS [21] for building an any-to-any VC model. The training, validation, and test sets included 86 (jvs001–jvs086), 4 (jvs087–jvs090), and 10 (jvs091–jvs100) speakers, respectively. We used the parallel100 subset of JVS including 100 parallel speech utterances for each speaker as the training and validation sets. The test set, which was used as reference speech samples of unseen target speakers in any-to-any VC, was the nonpara30 subset of JVS including 30 non-parallel speech utterances for each speaker. Therefore, the trained VC model was evaluated on the task of converting 100 SRC4VC speakers' samples into those of the 4 JVS test speakers. Although our SRC4VC corpus contains the "Singing" subset, we excluded it from the evaluation data. The reason was that the VC task, i.e., zero-shot singing VC using a model trained

Table 5: *Objective evaluation results (NISQA Naturalness). "DA" and "SE" denote data augmentation and speech enhancement, respectively.*

Methods	SRC4VC Q1	Q2	Q3	Q4	Avg.	JVS
Baseline (B)	2.36	2.50	2.38	2.41	2.41	2.83
B + DA (noise)	2.39	2.53	2.41	2.47	2.45	2.82
B + DA (reverb)	2.33	2.46	2.48	2.50	2.44	2.90
B + DA (band)	2.34	2.51	2.43	2.51	2.45	2.82
B + SE (Demucs)	2.39	2.50	2.42	2.43	2.43	N/A
B + SE (Miipher)	2.49	2.60	2.51	2.50	**2.52**	N/A

on reading-style speech only, was too difficult for the trained model. We downsampled all speech data to 16 kHz.

Backbone VC model: We adopted S2VC [28] as the backbone VC model for our experiment. The S2VC model incorporates self-supervised learning features [29] as the intermediate representations used for VC. Referring to Huang et al.'s degradation-robust VC study [11], we trained the S2VC model to generate target speaker's mel-spectrograms from the source and target speakers' contrastive predictive coding features [30]. We used the implementation of S2VC publicly available in GitHub[4]. We trained S2VC with 100K iterations and used the model that achieved the lowest validation loss for our evaluation. The initial learning rate was set to 5.0×10^{-5} and decayed with warm-up cosine annealing [31]. The optimizer was AdamW [32]. We also trained HiFi-GAN [33] to synthesize a speech waveform from a mel-spectrogram using the same training data for S2VC. The optimization algorithm was Adam [34] with η of 0.0002, β_1 of 0.8, and β_2 of 0.99. We trained HiFi-GAN with 600K iterations. The computational resources were two NVIDIA GeForce 1080 Ti GPUs.

Compared methods: We compared data augmentation and preprocessing techniques for improving the robustness of VC [11]. Specifically, we examined three data augmentation methods: additive noise, reverberation, and band rejection, using the same configurations as those in Huang et al.'s study [11]. We also used two speech enhancement models: Demucs [35][5] and Miipher [36])[6] with their official and unofficial implementation available on GitHub, respectively. For simplicity, we did not investigate combinations of these techniques, i.e., applying two or more data augmentations combined with speech enhancement.

Speaker groups: To investigate the relation between recording quality and VC performance, we first divided the 100 SRC4VC speakers into four groups (Q1–Q4) using the quantile values of their recording-quality MOS. Then, we randomly sampled 250 samples from each group. Finally, we paired the samples with randomly selected JVS test speakers' voice and used them as the source and target in VC.

4.2. Objective evaluation

As the objective evaluation criteria, we used the NISQA *Naturalness* score [25] of the converted speech. We did not evaluate the JVS-to-JVS VC cases when we used speech enhancement as the preprocessing.

Table 5 shows the evaluation results. The predicted *Naturalness* scores of the two VC cases using JVS and SRC4VC as the source speakers were significantly different, indicating that the recording-quality mismatch between the training and evaluation actually worsened the naturalness of converted speech. The introduced data augmentation strategies could mitigate the deterioration in naturalness, but the effectiveness was not so meaningful. We observed that the relatively high-quality recorded speakers (i.e., Q3 and Q4) benefited from the data augmentation, while Q1 and Q2 suffered from the naturalness degradation when inappropriate data augmentations for them (i.e., reverb or band) were adopted. The optimal strategy was using Miipher for speech enhancement. These results suggest that, for VC using an end-user's voice as the source speech, the recording-quality mismatch can be addressed by simply introducing speech enhancement as preprocessing, rather than training a VC model with data augmentation.

[4]https://github.com/cyhuang-tw/robust-vc/tree/main

[5]https://github.com/facebookresearch/demucs

[6]https://github.com/Wataru-Nakata/miipher

Table 6: *Results of naturalness and similarity MOS tests.*

Methods	Naturalness	Similarity
Baseline (B)	2.54	2.17
B + DA (noise)	2.59	2.18
B + DA (reverb)	2.66	**2.22**
B + DA (band)	2.62	2.17
B + SE (Demucs)	2.53	2.17
B + SE (Miipher)	**2.74**	2.21

4.3. Subjective evaluation

We conducted two five-scaled MOS tests regarding the naturalness and speaker similarity of converted speech. In the naturalness test, we presented 30 converted speech samples to listeners in random order. The listeners rated the naturalness of each speech sample with an integer between 1 (very bad) to 5 (very good). In the similarity test, listeners first listened to the target speaker's natural speech for reference. Then, they scored how similar the speaker of the presented voice was to the reference. A total of 30 synthetic speech samples were presented. Two hundred listeners participated in each test using our crowdsourcing evaluation system, for a total of 400 listeners. Unlike the objective evaluation, we did not evaluate JVS-to-JVS converted samples or distinguish the SRC4VC source speakers by the recording-quality scores.

Table 6 shows the results of the subjective evaluation. The "Baseline + SE (Miipher)" achieved the highest naturalness, whose score was statistically significant compared to that of "Baseline" ($p < 0.05$). This further demonstrates the effectiveness of speech enhancement preprocessing. From this result, we decided to publish the restored version of SRC4VC (i.e., *SRC4VC-R*) as well. Focusing on the evaluation results of speaker similarity, there was no statistical significance among the obtained scores, although the score of "Baseline + DA (reverb)" was slightly higher than the others.

5. Conclusion

We presented SRC4VC, a new multi-speaker speech corpus for validating the degradation robustness of a VC model. The results of any-to-any VC experiments demonstrated that the recording-quality mismatch between the training and evaluation data significantly worsened the VC performance, and applying speech enhancement to the low-quality source speech samples improved the performance. In the future, we intend to develop a method to adapt a VC model trained with clean speech to an end-user's degraded speech input. We also continue to enlarge SRC4VC so that it includes speech degradations caused by transmission channels (e.g., live VoIP calls covered by the test subsets of NISQA corpus [25]).

Acknowledgements: This research was conducted as joint research between LY Corporation and Saruwatari-Takamichi Laboratory of The University of Tokyo, Japan. This work was

supported by Research Grant S of the Tateishi Science and Technology Foundation.

6. References

[1] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[2] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 4115–4119.

[3] H. Okano, Y. Okada, N. Wakatsuki, and K. Zempo, "Effect of voice imitation using voice conversion by avatar on customer service in virtual environments," in *Proc. VRST*, Christchurch, New Zealand, Jun. 2023.

[4] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, Rome, Italy, Sep. 2018, pp. 2114–2118.

[5] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, Sep. 2019.

[6] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *Proc. ICLR*, Virtual Conference, Apr. 2022.

[7] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: english multi-speaker corpus for CSTR voice cloning toolkit," https://doi.org/10.7488/ds/2645, 2019.

[8] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1526–1530.

[9] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

[10] T.-h. Huang, J.-h. Lin, and H.-y. Lee, "How far are we from robust voice conversion: a survey," in *Proc. SLT*, Virtual Conference, 2021, pp. 514–521.

[11] C.-Y. Huang, K.-W. Chang, and H.-Y. Lee, "Toward degradation-robust voice conversion," in *Proc. ICASSP*, Singapore, May 2022, pp. 6777–6781.

[12] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[13] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda, "The Singing Voice Conversion Challenge 2023," in *Proc. ASRU*, Taipen, Taiwan, Dec. 2023.

[14] J. Koguchi, I. Kanai, Y. Oda, T. Saito, and M. Morise, "ITA Corpus," https://github.com/mmorise/ita-corpus, 2021.

[15] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "JVNV: A corpus of Japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, Feb. 2024.

[16] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of Japanese empathetic dialogue speech towards friendly voice agent," in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 5155–5159.

[17] Y. Saito, E. Iimori, S. Takamichi, K. Tachibana, and H. Saruwatari, "CALLS: Japanese empathetic dialogue speech corpus of complaint handling and attentive listening in customer center," in *Proc. INTERSPEECH*, Dublin, Ireland, Aug. 2023, pp. 5561–5565.

[18] E. Simperl, "How to use crowdsourcing effectively: Guidelines and examples," *LIBER Quarterly*, vol. 25, no. 1, pp. 18–39, Aug. 2015.

[19] H. Li and J. Yamagishi, "DDS: A new device-degraded speech dataset for speech enhancement," in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 2913–2917.

[20] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, Aug. 2015.

[21] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

[22] S. Takamichi and H. Saruwatari, "CPJD Corpus: Crowdsourced parallel speech corpus of Japanese dialects," in *Proc. LREC*, Miyazaki, Japan, May 2018, pp. 434–437.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[24] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: an open-source library for sound and music analysis," in *Proc. ISMIR*, Curitiba, Brazil, Nov. 2013, pp. 493–498.

[25] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 2127–2131.

[26] M. Waltermann, *Dimension-based Quality Modeling of Transmitted Speech*. Springer, 2012.

[27] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of loudness level on the overall quality of transmitted speech," in *Proc. AES Convention*, New York, U.S.A., Oct. 2007.

[28] J. 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*, Brno, Czech Republic, Sep. 2021, pp. 836–840.

[29] W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, and T. Toda, "S3PRL-VC: Open-source voice conversion framework with self-supervised speech representations," in *Proc. AAAI SAS Workshop*, Virtual Conference, Feb. 2022.

[30] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. abs/1807.03748, 2018.

[31] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, Toulon, France, Apr. 2017.

[32] ——, "Decoupled weight decay regularization," in *Proc. ICLR*, New Orleans, U.S.A., May 2019.

[33] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020.

[34] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[35] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 3291–3295.

[36] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," in *Proc. WASPAA*, New York, U.S.A., Oct. 2023.