CycleSiFiNF-VC: Controllable Non-Parallel Voice Conversion by Neural Formant Manipulation with Improved Cycle-Consistency Loss

Sumiharu Kobayashi*, Takashi Nose* and Akinori Ito*
* Tohoku University, Japan
E-mail: {kobayashi.sumiharu.r4@dc, takashi.nose.b7@, akinori.ito.a2@}.tohoku.ac.jp

***Abstract*—Recent voice conversion (VC) methods based on self-supervised learning (SSL) achieve high-quality and high speaker similarity conversions but face challenges in controlling speech attributes, such as prosody and brightness. A CycleGAN-based voice conversion using mel-cepstral coefficients has a kind of speech controllability by adjusting the frequency-warping parameter of the mel-cepstral coefficients. However, unlike speech parameters such as fundamental frequency or formants, which have a one-to-one correspondence with the physical properties of speech, it is difficult to intuitively adjust speech using mel-cepstral coefficients. Therefore, this paper proposes CycleSiFiNF-VC, a non-parallel voice conversion method that transforms speech parameters using cycle-consistency learning. Our proposed method simultaneously predicts and transforms speech parameters from mel-spectrograms, enabling individual manipulation of speech parameters during the voice conversion process. Experiments demonstrate that the proposed method achieves higher synthesis quality and comparable or superior speaker similarity compared to CycleGAN-VC2, and enables control over speech parameters, which is difficult to achieve with conventional methods.**

I. Introduction

Recently, high-quality voice conversion has been achieved using GANs and flow-based models [1–3]. Furthermore, voice conversion methods utilizing self-supervised learning (SSL) models such as HuBERT [4], WavLM [5] and ContentVec [6], as well as large-language models (LLMs) exemplified by Transformer [7], have been actively researched [8–12]. These methods allow for voice conversion without the need for parallel data, achieving high speaker similarity even with limited data (e.g., one-shot or zero-shot learning) by utilizing latent features from flow-based models, as well as SSL and LLM features. However, a drawback is the complexity of these features, which makes it difficult to intuitively control specific speech parameters, such as voice brightness or prosody. Although some voice conversion methods offer more controllability, such as CycleGAN-VC2 [13], which uses mel-cepstral coefficients with frequency-warping parameters [14] as speech features, intuitive control over speech characteristics remains a challenge. This is because the frequency-warping parameters do not directly correspond to the physical properties of speech.

In the neural vocoders, neural formant synthesis [15] has been proposed as a controllable vocoder system using speech parameters that directly correspond to the physical characteristics of speech. This method demonstrates that speech synthesis and control are achievable by predicting a mel-spectrogram from speech parameters using a feature-mapping network and then feeding it as an input to a neural vocoder. To enhance the synthesis quality and controllability, we proposed end-to-end neural formant synthesis [16] as an extension of [15]. This approach showed that, through the integration of end-to-end learning and a source-filter structure, it is possible to directly synthesize speech quality comparable to HiFi-GAN [17] from nine speech parameters. These methods allow the modification of specific speech attributes, by individually controlling independent speech parameters.

This study extends [16] to voice conversion by proposing CycleSiFiNF-VC, a non-parallel voice conversion method that achieves both voice conversion and speech controllability using a small set of speech parameters. CycleSiFiNF-VC is a model that simultaneously performs speech parameter prediction and conversion using a non-causal gated convolutional generator based on WaveNet [18] and cycle-consistency learning. By using the speech parameters predicted and converted from mel-spectrograms and then synthesizing speech waveforms with E2E-SiFi-NF [16], both speech control and voice conversion are achieved.

II. Related Work

A. CycleGAN-Based Voice Conversion

CycleGAN-based voice conversion is a non-parallel voice conversion method that applies CycleGAN [19], an image transfer method, to learn the correspondence of mel-cepstral coefficients between two speakers. CycleGAN-VC [20] introduced an encoder-decoder structure with 1D CNNs that extend in the temporal direction to capture relationships in the feature dimension while preserving the temporal structure. However, 1D CNNs can affect the frequency-domain structure that should be preserved during conversion. Therefore, CycleGAN-VC2 introduced a 2-1-2D CNN architecture, using 2D CNNs in the encoder-decoder parts and 1D CNNs in the residual blocks. Additionally, a two-step adversarial loss, which applies adversarial loss to the cycle-converted mel-cepstral coefficients, was introduced to suppress the excessive smoothing caused by cycle-consistency learning. These modifications enabled CycleGAN-VC2 to achieve higher-quality

979-8-3315-7206-8/25/$31.00

conversions than CycleGAN-VC. Furthermore, this method allows for speech control by manipulating the frequency-warping parameter α, which represents the phase characteristics of an all-pass filter that approximates the mel scale. However, because α does not have a one-to-one correspondence with physical speech characteristics, such as formants or spectral tilt, controlling specific speech properties remains challenging.

B. Neural Formant Synthesis

Many recently proposed neural vocoders synthesize speech waveforms from mel-spectrograms, which makes it challenging to control physical speech properties, such as pitch and formants, individually. To address the limitation, neural formant synthesis [15] was proposed, which synthesizes speech waveforms from a small set of speech parameters. Neural formant synthesis consists of a feature-mapping network that predicts mel-spectrograms from speech parameters and a neural vocoder that synthesizes speech from these mel-spectrograms. It utilizes nine speech parameters: a voiced/unvoiced flag, log fundamental frequency (F0), formants (F1-F4), spectral tilt, spectral centroid, and energy. Experiments have shown that this model is speaker-independent and that it is possible to reproduce natural speech from given speech parameters and to manipulate each speech parameter independently.

C. End-to-End Neural Formant Synthesis

While neural formant synthesis enables speech control through speech parameters, it suffers from degraded synthesis quality owing to its stepwise speech generation process from these parameters. To address the problem, we proposed end-to-end neural formant synthesis [16], which directly synthesizes speech waveforms from the nine speech parameters in Section II-B. This method demonstrated that end-to-end speech synthesis and control can be achieved simultaneously by leveraging the structure of existing neural vocoders. Furthermore, to enhance both the synthesis quality and controllability of the speech parameters, we introduced end-to-end source-filter neural formant synthesis (E2E-SiFi-NF) [16]. This model applies a neural architecture based on the source-filter structure to directly synthesize the speech waveforms from the speech parameters. The integration of the source-filter model allowed for end-to-end synthesized speech from the nine speech parameters to achieve a synthesis quality that surpassed that of HiFi-GAN. Moreover, it led to improved synthesis quality and fidelity during speech parameter manipulation compared to the conventional neural formant synthesis.

III. CycleSiFiNF-VC

A. CycleGAN-Based Voice Conversion and Manipulation

The most straightforward way to achieve both voice conversion and manipulation by speech parameters is to convert these parameters within a CycleGAN-based voice conversion method (as discussed in Section II-A), replacing mel-cepstral coefficients with the speech parameters of neural formant synthesis, and then synthesize the speech waveform using E2E-SiFi-NF as the vocoder. One prominent CycleGAN-based voice conversion method, CycleGAN-VC2 [13], uses 2D CNNs because the mel-cepstral coefficients possess a time-frequency structure. However, speech parameters are independent time-series data and lack this time-frequency structure. Therefore, to preserve the temporal structure while capturing the interrelationships among speech parameters, we adopted an encoder-decoder architecture composed of 1D CNNs, similar to CycleGAN-VC [20]. In our proposed method, for both the input and output, we used seven speech parameters for conversion, excluding the F0 and voiced/unvoiced (V/UV) flag from the nine speech parameters mentioned in Section II-B. This choice ensured consistency with CycleGAN-VC2, as these seven parameters were related to the vocal tract, similar to how mel-cepstral coefficients are features pertaining to the vocal tract filter.

B. WaveNet-Style Feature Conversion Network

The method proposed in Section III-A suffers from a problem in which the phonetic content of the source speech is not well preserved, particularly in cross-gender conversions. This issue likely arises because each speech parameter is an independent one-dimensional time-series data, and the receptive field of the encoder-decoder of CycleGAN-VC, which performs temporal stretching and compression, is too narrow to capture long-term dependencies. To effectively capture these long-term dependencies, we introduced a feature conversion network into the generator. This network is a non-causal WaveNet-style gated convolutional architecture inspired by the feature-mapping network [15]. It can efficiently capture long-term temporal dependencies using dilated convolutions and residual and skip connections. This network features six residual blocks with a dilation pattern of (1, 2, 4, 1, 2, 3), a kernel size of 3, 256 channels, and a postnet. Unlike the original feature-mapping network, this postnet uses a gated linear unit (GLU) as its activation function, and instance normalization [21] is applied to some convolutional layers.

C. Conversion from Mel-Spectrogram to Speech Parameters

Our preliminary experiments revealed that by adjusting the input layer dimension of the feature conversion network, we could predict speech parameters from mel-spectrograms, which is the inverse of the approach in [15]. This suggests that by using mel-spectrograms as inputs to the feature conversion network, we can simultaneously perform speech parameter prediction and conversion. However, calculating the cycle-consistency loss requires that the input and output features be of the same type. Therefore, when using mel-spectrograms as inputs, as shown in Fig. 2, the cycle-consistency loss is calculated by predicting speech parameters from mel-spectrograms using a feature-mapping network. To capture the temporal structure of continuous frames in the mel spectrogram, the input layer and postnet of the feature conversion network were replaced by a decoder consisting of a stride=2 CNN and a CNN and a pixel shuffler[22], as proposed by CycleGAN-VC.

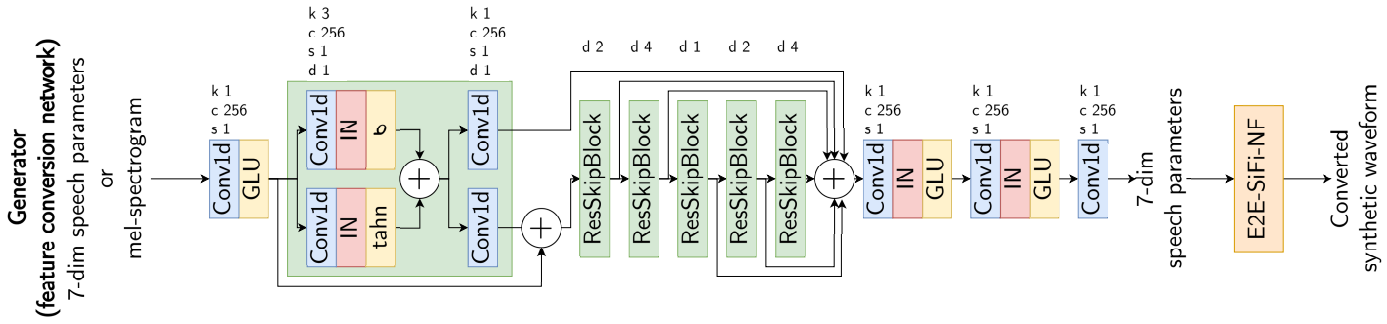

Fig. 1. Network architecture of feature conversion network for CycleSiFiNF-VC generator. where k, c, d, and s denote the kernel size, channel size, dilation size, and stride, respectively. IN and GLU denote instance normalization and gated linear unit, respectively.

D. Cycle-Consistency and Identity-Mapping Speech Parameter Losses

When mel-spectrograms are used as input to CycleSiFiNF-VC, the temporal structure of the speech parameters is indirectly guaranteed by the cycle-consistency loss between the input and mel-spectrograms predicted from the speech parameters by the feature-mapping network. Consequently, the stability of the early training phase depends on the prediction accuracy of the mel-spectrograms by the feature mapping network. Among the speech parameters, formants are particularly crucial because they are closely related to phoneme content, and improving their prediction accuracy is vital for enhancing conversion quality. Therefore, to boost the prediction accuracy of the speech parameters by the feature conversion network, our proposed method incorporates an additional cycle-consistency speech parameter loss, $L_{\text{cyc\_sp}}$, which directly ensures the temporal structure of the speech parameters.

$$\begin{aligned} L_{\text{cyc\_sp}} = \mathbb{E}_{x\sim P_X}\left[\|G_{Y\to X}(F(G_{X\to Y}(x_{\text{mel}}))) - x_{\text{sp}}\|_1\right] \\ + \mathbb{E}_{y\sim P_Y}\left[\|G_{X\to Y}(F(G_{Y\to X}(y_{\text{mel}}))) - y_{\text{sp}}\|_1\right] \end{aligned} \quad (1)$$

Here, $G_{X\to Y}$ and $G_{Y\to X}$ are generators for converting X to Y and Y to X, respectively, while the feature-mapping network F generates mel-spectrograms from speech parameters. x_{sp} and y_{sp} represent the pre-extracted speech parameters of the source and target speakers. $L_{\text{cyc\_sp}}$ is the L1 loss between the speech parameters obtained from the cycle conversion and the original source speaker's speech parameters, which serves to improve the prediction accuracy of the speech parameters. Furthermore, to stabilize the prediction of the speech parameters during the initial stages of training, we introduce the following identity-mapping speech parameter loss, $L_{\text{id\_sp}}$.

$$\begin{aligned} L_{\text{id\_sp}} = \mathbb{E}_{y\sim P_Y}\left[\|G_{X\to Y}(y_{\text{mel}}) - y_{\text{sp}}\|_1\right] \\ + \mathbb{E}_{x\sim P_X}\left[\|G_{Y\to X}(x_{\text{mel}}) - x_{\text{sp}}\|_1\right] \end{aligned} \quad (2)$$

The addition of these losses, which are related to speech parameters, is expected to improve the conversion quality, particularly in terms of phoneme preservation. The final loss function used in CycleSiFiNF-VC is as follows:

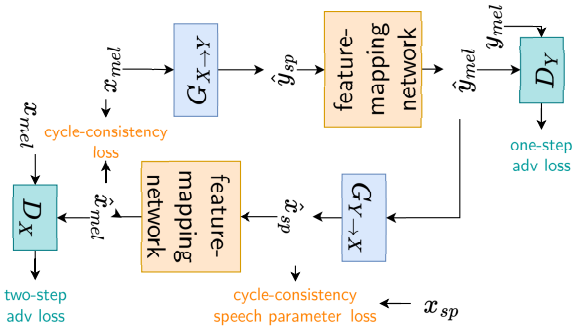

Fig. 2. Training flow of CycleSiFiNF-VC

$$\begin{aligned} L_{\text{total}} = L_{\text{adv}}(G_{X\to Y}, D_Y) + L_{\text{adv}}(G_{Y\to X}, D_X) \\ + L_{\text{adv2}}(G_{X\to Y}, G_{Y\to X}, D'_X) \\ + L_{\text{adv2}}(G_{Y\to X}, G_{X\to Y}, D'_Y) \\ + \lambda_{\text{cyc}} L_{\text{cyc}} + \lambda_{\text{id}} L_{\text{id}} \\ + \lambda_{\text{cyc\_sp}} L_{\text{cyc\_sp}} + \lambda_{\text{id\_sp}} L_{\text{id\_sp}} \end{aligned} \quad (3)$$

Here, L_{adv} denotes the one-step adversarial loss, L_{adv2} the two-step adversarial loss, L_{cyc} the cycle-consistency loss, and L_{id} the identity-mapping loss, all of which are defined in [13].

IV. Experiments

A. Experimental Conditions

The dataset utilized 100 utterances from the parallel100 subset of the JVS corpus [23], which comprised 100 Japanese males and females recorded at 24 kHz. The data were split into training and evaluation sets at a 90:10 ratio for each of the 100 utterances. Male speaker jvs001 and female speaker jvs002 were used as source speakers, whereas male speaker jvs003 and female speaker jvs004 were used as target speakers. All combinations of these source and target speakers were evaluated. speech parameters, specifically 40-dimensional mel-cepstral coefficients and 80-dimensional mel-spectrograms, were extracted using the same parameters as in [16].

For comparison, we evaluated CycleGAN-VC2 with mel-cepstral features as the conventional CycleGAN-based voice conversion. CycleGAN-VC2 and CycleSiFiNF-VC were trained for 200k iterations using a batch size of 1 and the Adam optimizer [24] with $\beta_1 = 0.5$ and $\beta_2 = 0.9$. The generator and discriminator learning rates were set to 0.0002 and 0.0001, respectively. The lengths of the input frames to CycleSiFiNF-VC and CycleGAN-VC2 were 64 and 128, respectively. The loss weights were set as $\lambda_{\mathrm{cyc}} = 10$, $\lambda_{\mathrm{id}} = 5$, $\lambda_{\mathrm{cyc\_sp}} = 10$, and $\lambda_{\mathrm{id\_sp}} = 5$. These training parameters were based on CycleGAN-VC2[13]. Note that L_{id} and $L_{\mathrm{id\_sp}}$ were employed only for the initial 10k iterations. The discriminator for both models adopted the PatchGAN [25] architecture, similar to that of CycleGAN-VC2. However, for CycleSiFiNF-VC (Baseline) and CycleSiFiNF-VC (FCN), the 2D CNN in PatchGAN was replaced with a 1D CNN to learn the conversion from source speaker speech parameters to target speaker speech parameters.

The feature-mapping network described in Section II-B was used as the pretrained model. This network predicts 80-dim mel-spectrograms from seven speech parameters. It was trained for 99k iterations with a batch size of 128 using the Adam optimizer and an input frame length of 46 frames. E2E-SiFi-NF was employed as the vocoder for CycleSiFiNF-VC. For CycleGAN-VC2, SiFi-GAN [26] was used as the vocoder, taking 40-dimensional mel-cepstral coefficients and 3-dimensional band aperiodicity (bap) as the input features. Each vocoder was trained for 500k iterations using speech data from 96 speakers (JVS005-JVS100) from the JVS corpus. During conversion, F0 was linearly transformed in the logarithmic domain, and the source speaker's bap was directly utilized.

In the following experiments, the proposed method was evaluated under the following four conditions.

- **Baseline**: This model is essentially the CycleGAN-VC described in Section III-A with the input features changed from mel-cepstral coefficients to 7-dimensional speech parameters.
- **FCN**: This model introduces the feature conversion network (FCN), detailed in Section III-B, into the Baseline model.
- **FCN+Mel**: This model builds upon the FCN model by incorporating the conversion from mel-spectrogram to speech parameters, as described in Section III-C.
- **FCN+Mel+L_{sp}**: This model adds $L_{\mathrm{cyc\_sp}}$ and $L_{\mathrm{id\_sp}}$, which are explained in Section III-D, to the FCN+Mel model. When referred to simply as CycleSiFiNF-VC, this model is implied.

B. Objective Evaluation

To objectively assess the quality of the converted speech, the predicted speech quality score (UTMOS) [27] was calculated using the MOS prediction system. As objective evaluation metrics, the cosine similarity of speaker embeddings extracted by Resemblyzer[1] was calculated to assess the speaker similarity between the converted speech and the target speaker. The character error rate (CER) was calculated to evaluate whether the utterance content was appropriately preserved.

[1]https://github.com/resemble-ai/Resemblyzer

TABLE I
COMPARISON OF PREDICTED MOS. THE BEST AND THE 2ND BEST SCORES ARE INDICATED IN BOLD AND WITH AN UNDERLINE, RESPECTIVELY.

Model	F2F	M2M	M2F	F2M
CycleGAN-VC2 [13]	1.42	1.46	1.42	1.40
CycleSiFiNF-VC (Baseline)	1.47	1.84	1.25	1.26
CycleSiFiNF-VC (FCN)	<u>1.71</u>	<u>2.05</u>	<u>1.91</u>	<u>1.86</u>
CycleSiFiNF-VC (FCN+Mel)	1.33	1.27	1.37	1.26
CycleSiFiNF-VC (FCN+Mel+L_{sp})	**2.43**	**2.58**	**2.30**	**2.19**

TABLE II
COMPARISON OF SPEAKER COSINE SIMILARITY

Model	F2F	M2M	M2F	F2M
CycleGAN-VC2	0.843	**0.901**	0.824	**0.865**
CycleSiFiNF-VC (Baseline)	0.779	0.825	0.748	0.664
CycleSiFiNF-VC (FCN)	<u>0.861</u>	0.870	**0.844**	0.827
CycleSiFiNF-VC (FCN+Mel)	0.808	0.767	0.807	0.720
CycleSiFiNF-VC (FCN+Mel+L_{sp})	**0.872**	<u>0.889</u>	<u>0.832</u>	<u>0.855</u>

The results for UTMOS, speaker cosine similarity, and CER, are presented in Tables I, II, and III, respectively. From Table I, we found that CycleSiFiNF-VC obtained the highest scores in UTMOS. Regarding the speaker cosine similarity in Table II, CycleSiFiNF-VC (FCN) is comparable to CycleGAN-VC2, indicating that the feature conversion network is feasible through cycle-consistency learning and the non-causal WaveNet-style gated convolutional structure, which is capable of capturing long-term temporal structure. CycleSiFiNF-VC (FCN+Mel+L_{sp}) achieved speaker cosine similarity scores equal to or better than CycleGAN-VC2 for both intra-gender and cross-gender conversions, despite a degradation in CER (Table III). CycleSiFiNF-VC (Baseline) showed a significant deterioration in CER for cross-gender conversions. When L_{sp} was not introduced in CycleSiFiNF-VC (FCN+Mel), all evaluation metrics deteriorated. This demonstrates that when speech parameter prediction and conversion are performed simultaneously with the intent of manipulating speech parameters during the conversion process, the speech parameter loss, which directly guarantees the temporal structure of the speech parameters, proves to be effective. Unlike mel-spectrograms, speech parameters contain less detailed representations of speech, despite being indicative of specific physical characteristics. Consequently, the use of mel-spectrograms enabled the acquisition of representations not obtainable through direct conversion from acoustic parameters, leading to an improvement in conversion quality.

C. Subjective Evaluation

For the subjective evaluation, the conventional CycleGAN-VC2 and CycleSiFiNF-VC (FCN+Mel+L_{sp}), which is the highest performance proposed model, were compared. Six utterances were randomly selected for evaluation for all conversion pairs. A mean opinion score (MOS) test (5: excellent

TABLE III
COMPARISON OF CER

Model	F2F	M2M	M2F	F2M
CycleGAN-VC2	**5.02**	7.53	**5.65**	**5.44**
CycleSiFiNF-VC (Baseline)	11.8	**4.60**	58.2	73.0
CycleSiFiNF-VC (FCN)	15.4	9.79	12.5	12.9
CycleSiFiNF-VC (FCN+Mel)	12.3	13.7	11.3	20.3
CycleSiFiNF-VC (FCN+Mel+L_{sp})	6.70	6.96	9.94	10.1

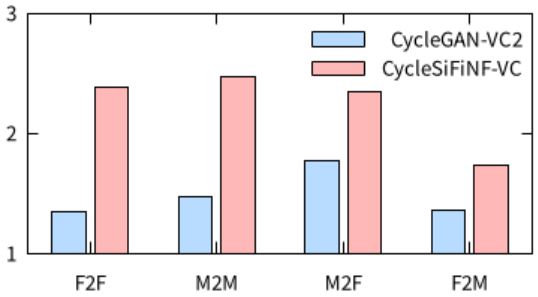

Fig. 3. MOS for conversion quality

to 1: bad) was conducted to assess the quality of the converted speech. An XAB test was performed to measure speaker similarity. In this test, "A" and "B" represented speech converted by CycleGAN-VC2 and CycleSiFiNF-VC, respectively, while "X" is speech of the target speaker. Six pairs of sentences were randomly chosen from the evaluation set, and the pairs were arranged randomly to eliminate any combination bias. For each sentence pair, listeners were asked to select their preference ("A" or "B") or choose "Fair". A total of 13 native Japanese speakers (males and females) participated in the listening tests.

The results of the MOS and XAB tests are presented in Figs. 3 and 4, respectively. The overall low score for the MOS test could be a result of including natural speech in the evaluation. In the MOS test, CycleSiFiNF-VC achieved higher scores than CycleGAN-VC2 for all the conversion pairs. In the XAB test, CycleSiFiNF-VC showed higher similarity for intra-gender conversions than CycleGAN-VC2. However, among the cross-gender conversions, CycleSiFiNF-VC slightly outperformed CycleGAN-VC2 in M2F conversions, whereas CycleGAN-VC2 showed higher scores for F2M conversions.

D. Parameter-Manipulated Conversion

We evaluated speech synthesized by manipulating formants predicted and converted from mel-spectrograms in CycleSiFiNF-VC. To investigate the fidelity to the shift scale, we calculated the root mean square error (RMSE) between the manipulated input F1, F2 and the F1, F2 extracted from the synthesized speech. The results are shown in Figure 5. $\times$n for each item indicates the scale of the manipulated speech parameter. As a baseline, we present the results of parameter-manipulated synthesis using E2E-SiFi-NF for the evaluation data. While the error grows with a larger scale, this observation is explained in [15, 16]. Specific speech properties can be controlled by manipulating individual speech parameters

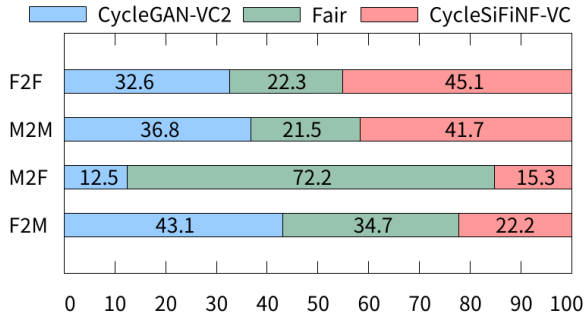

Fig. 4. Average preference score (%) of XAB test for speaker similarity

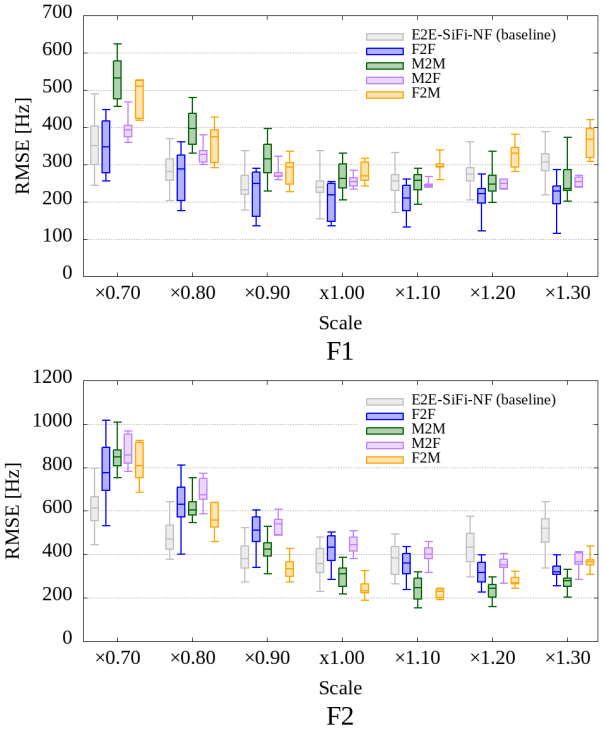

Fig. 5. Plot of RMSE for manipulated F1 and F2

predicted and converted by the feature conversion network. These results suggest that CycleSiFiNF-VC can achieve both speech parameter prediction from mel-spectrograms and high-quality conversion simultaneously. However, the varying error magnitudes in speech parameter prediction across different conversion pairs suggest a potential dependency on the specific conversion pair.

V. CONCLUSIONS

This paper proposed CycleSiFiNF-VC, a CycleGAN-based non-parallel voice conversion method that utilizes a small set of speech parameters to achieve both voice conversion and speech controllability. Our proposed method introduced a feature conversion network and speech parameter losses that effectively handles the temporal structure of speech parameters. Evaluation results showed that our method can simultaneously perform speech parameter prediction and conversion,

achieving quality superior to and speaker similarity comparable to or better than CycleGAN-VC2. Speech parameter prediction enables controllability in the voice conversion process through parameter manipulation. In the future work, we aim to improve conversion quality by enhancing the accuracy of speech parameter prediction.

Acknowledgment

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research JP23K21945, JP25K00471, JP23K20725.

References

[1] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Maskcyclegan-VC: Learning Non-Parallel Voice Conversion with Filling in Frames," in *Proc. ICASSP*, 2021, pp. 5919–5923.

[2] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. ICML*, PMLR, 2021, pp. 5530–5540.

[3] 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.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] S. Chen et al., "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.

[6] K. Qian et al., "ContentVec: An Improved Self-Supervised Speech Representation by Disentangling Speakers," in *Proc. ICML*, 2022.

[7] A. Vaswani et al., "Attention is All you Need," in *Proc. NeurIPS*, vol. 30, Curran Associates, Inc., 2017.

[8] 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.

[9] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *Proc. ICML*, 2022.

[10] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," in *Proc. ICASSP*, 2023, pp. 1–5.

[11] J. Yao et al., "PromptVC: Flexible Stylistic Voice Conversion in Latent Space Driven by Natural Language Prompts," in *Proc. ICASSP*, 2024, pp. 10 571–10 575.

[12] J. Hai, K. Thakkar, H. Wang, Z. Qin, and M. Elhilali, "DreamVoice: Text-guided voice conversion," in *Proc. INTERSPEECH*, 2024, pp. 4373–4377.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," in *Proc. ICASSP*, 2019, pp. 6820–6824.

[14] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[15] P. Pérez Zarazaga, Z. Malisz, G. E. Henter, and L. Juvela, "Speaker-independent neural formant synthesis," in *Proc. INTERSPEECH*, 2023, pp. 5556–5560.

[16] S. Kobayashi, T. Kosaka, and T. Nose, "End-to-End Neural Formant Synthesis Using Low-Dimensional Acoustic Parameters," in *Proc. GCCE*, 2024, pp. 820–823.

[17] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, vol. 33, Curran Associates, Inc., 2020, pp. 17 022–17 033.

[18] A. van den Oord et al., *WaveNet: A Generative Model for Raw Audio*, 2016.

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017.

[20] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, 2018, pp. 2100–2104.

[21] D. Ulyanov, A. Vedaldi, and V. Lempitsky, *Instance Normalization: The Missing Ingredient for Fast Stylization*, 2017.

[22] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[23] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, *JVS corpus: Free Japanese multi-speaker voice corpus*, 2019.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[25] U. Demir and G. Unal, *Patch-Based Image Inpainting with Generative Adversarial Networks*, 2018.

[26] 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.

[27] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. INTERSPEECH*, 2022, pp. 4521–4525.