VOICE CONVERSION WITH CYCLIC RECURRENT NEURAL NETWORK AND FINE-TUNED WAVENET VOCODER

Patrick Lumban Tobing[⋆] *Yi-Chiao Wu*[⋆] *Tomoki Hayashi*[⋆]
Kazuhiro Kobayashi[†] *Tomoki Toda*[†]

[⋆] Graduate School of Information Science, Nagoya University, Japan
[†] Information Technology Center, Nagoya University, Japan

ABSTRACT

This paper presents a novel framework for providing high-quality parallel voice conversion (VC) using a cyclic recurrent neural network (RNN) and a finely tuned WaveNet vocoder. Using the proposed system, we are tackling the quality degradation issue faced by WaveNet when it is fed with estimated (oversmoothed) speech features, such as mel-cepstrum parameters predicted from a statistical model. In VC, providing predicted features to fine-tune a pretrained WaveNet model is not straightforward owing to the difference in time-sequence alignment. To overcome this problem, we propose the use of a cyclic spectral conversion network that is capable of performing a conversion flow, i.e., source-to-target, and a cyclic flow, i.e., generate self-predicted target speaker features, and is trained by using both the conversion and cyclic losses. The experimental results demonstrate that, overall, the proposed system significantly improves the converted speech, resulting in a mean opinion score of 3.79 and a speaker similarity score of 73.86%.

Index Terms— voice conversion, cyclic recurrent neural network, WaveNet fine-tuning, oversmoothed parameters

1. INTRODUCTION

In a voice conversion (VC) framework [1], a conversion procedure is performed to transform the voice timbre, such as the conversion of vocal-tract spectra and the prosody, such as the transformation of fundamental frequency (F0) values. The conversion of spectral features, such as mel-cepstrum (spectral envelope) parameters [2], can be conveniently performed through the use of a statistical data-driven feature mapping model. Indeed, the development of statistical VC has been proceeding rapidly, as shown by various related works, such as through the use of Gaussian mixture model (GMM)-based methods [1, 3] and neural-network-based methods [4, 5].

In this work, we focus on the use of a recurrent neural network (RNN) for VC with parallel (paired) data. In recent years, VC with nonparallel (unpaired) data has become one of the main points of interest. Nevertheless, in many situations, it is still viable to acquire a small amount of parallel data for the development of a high-quality VC system. Therefore, it is worthwhile to improve the framework of parallel VC systems.

In addition to the spectral conversion, in VC, another essential aspect is the waveform generation step. It is well known that the conventional vocoder-based method for speech synthesizers has an inherent degradation problem [6], especially when using speech parameters predicted by a statistical model, even when using a superior vocoder system [7]. In a recent work [8], an alternative waveform generation method based on a deep convolutional neural network (CNN), the so-called WaveNet, was proposed. The WaveNet vocoder [9, 10], conditioned on extracted speech parameters such as spectral and excitation features, has been proven to be capable of producing humanlike speech and has become the state-of-the-art system. However, in VC [11], it still suffers from quality degradation when using predicted (oversmoothed) speech parameters.

In a text-to-speech (TTS) system [12], the degradation problem of the WaveNet vocoder can be overcome by using predicted features when developing the WaveNet model. In VC, it is not straightforward to achieve this as there is a mismatch of the time-sequence alignment. The method in [13] ingeniously addresses this issue through the use of phonetic-based (linguistic) intermediate features. In [14], the use of data-driven linguistic-free intermediate features for VC with a variational autoencoder was proposed. However, in practice, a large amount of data is required for training. In [15], a parallel VC method that can be trained using a relatively small amount of training data was proposed; such a method uses a system of concatenated spectral conversions, i.e., of target-to-source and source-to-target networks, to generate self-predicted target features for finely tuning a pretrained WaveNet model. The latter method, although simple and promising, can be further improved, for example by reducing the workload in training two separate models and by bridging the connection gap between the two disjointly trained networks.

In this paper, inspired by the CycleGAN architecture [5], we present a novel approach to improving a parallel VC system, especially for a WaveNet fine-tuning framework [15], by

proposing a VC with a cyclic spectral conversion model. The proposed cyclic network is capable of performing two spectral mapping flows, namely, a conversion flow, i.e., source-to-target, and a cyclic flow, i.e., generate self-predicted target features, and is trained using both the conversion and cyclic losses. A pretrained WaveNet model is then fine-tuned in accordance with the self-predicted target features generated from the cyclic flow. The experimental results demonstrate that, overall, the proposed system significantly improves both the quality and accuracy of the converted speech, compared with the use of conventional WaveNet fine-tuning, either with natural features or with the features predicted from disjoint networks.

2. BASELINE RNN-BASED VC FRAMEWORK WITH WAVENET VOCODER

2.1. Spectral conversion network with RNN and nonlinear autoregressive (AR) output

In [15], long short-term memory (LSTM)-based trajectory estimation [16] was proposed for spectral conversion by using multiple linear autoregressive (AR) layers. In this paper, we further improve the conversion network by using gated recurrent unit (GRU) [17] architecture, which is also capable of modeling long-term context dependences, albeit with fewer trainable model parameters, through the use of a nonlinear AR output. The flow of the RNN-AR-based spectral conversion is shown in Fig. 1.

Let $\boldsymbol{x}_t = [x_t(1), \ldots, x_t(D)]^\top$ and $\boldsymbol{y}_t = [y_t(1), \ldots, y_t(D)]^\top$ be the D-dimensional spectral feature vector of the input speaker and that of the target speaker, at frame t, respectively. Given a sequence of input spectral feature vectors $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_T^\top]^\top$, a sequence of preprocessed feature vectors $f(\boldsymbol{x}) = [f(\boldsymbol{x})_1^\top, \ldots, f(\boldsymbol{x})_T^\top]^\top$ is generated through convolutional layers and then fed into a set of hidden-GRU and AR-GRU blocks to compute both the hidden state and the output $\hat{\boldsymbol{y}}_t$ for each frame t.

2.2. WaveNet vocoder

WaveNet is a deep AR-CNN used for modeling speech waveform samples [8]. In [9], the so-called WaveNet vocoder was proposed, where each waveform sample is conditioned not only by previous samples but also by auxiliary speech features, such as spectral and excitation parameters. Given a sequence of auxiliary features $\boldsymbol{h} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T]^\top$, the likelihood of a sequence of waveform samples $\boldsymbol{s} = [s_1, \ldots, s_T]^\top$ is defined as

$$P(\boldsymbol{s}|\boldsymbol{h}) = \prod_{t=1}^{T} P(s_t|s_1, s_2, \ldots, s_{t-1}, \boldsymbol{h}_t). \quad (1)$$

By using a statistical VC model, such as the RNN-AR network, to estimate spectral parameters, and then feeding them to the WaveNet vocoder, we could generate a converted speech waveform from a source speaker into a particular target speaker.

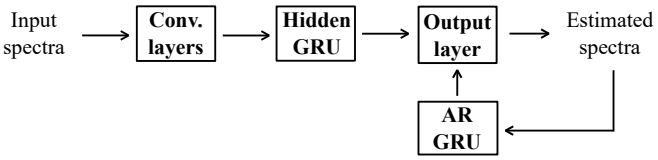

Fig. 1. Conversion network with hidden and AR GRU blocks

In a WaveNet model, a stack of dilated convolutional layers is used to efficiently increase the number of receptive fields. Each layer has a residual block comprising a 2×1 dilated causal convolution with a gated activation function and two 1×1 convolutions connected to either the next residual block or skip connection. All skip connections are summed and then fed to the output layer using the softmax function. The gated activation function is

$$\tanh(\boldsymbol{U}_{f,k} * \boldsymbol{s} + \boldsymbol{V}_{f,k} * \boldsymbol{h}') \odot \sigma(\boldsymbol{U}_{g,k} * \boldsymbol{s} + \boldsymbol{V}_{g,k} * \boldsymbol{h}'), \quad (2)$$

where $\boldsymbol{U} * \boldsymbol{s}$ denotes a dilated causal convolution, $\boldsymbol{V} * \boldsymbol{h}'$ denotes a 1×1 convolution, k is the layer index, f and g denote "filter" and "gate", respectively, and $\boldsymbol{h}'$ denotes an upsampled auxiliary feature vector sequence from an upsampling layer.

3. CYCLIC CONVERSION NETWORK FOR FINE-TUNING OF WAVENET VOCODER

3.1. Problems of WaveNet fine-tuning with predicted features

Conditioned on extracted speech parameters, a WaveNet vocoder is capable of generating speech waveforms with natural quality [10]. However, by introducing oversmoothed features, such as mel-cepstrum parameters estimated from a statistical model, the quality of the converted waveform will be significantly degraded [11]. This is because of the mismatches between the natural parameters used in the training and the oversmoothed (estimated) parameters used in the generation time.

In VC, providing oversmoothed parameters to develop a WaveNet model is not straightforward owing to the difference in time-sequence alignment. On the basis of [11], in [15], a network structure capable of generating self-predicted target features that can be used to fine-tune a pretrained WaveNet model was proposed. The system is a simple concatenation of two separately trained (disjoint) conversion networks, namely, the target-to-source (TSmap) and source-to-target (STmap) mappings, as shown in Fig. 2. This fine-tuning procedure is capable of improving the converted speech quality compared with the fine-tuning of WaveNet with natural features. However, the accuracy is still limited owing to the connection gap between the two disjointly trained networks. Note that in the conversion stage, only the STmap is used.

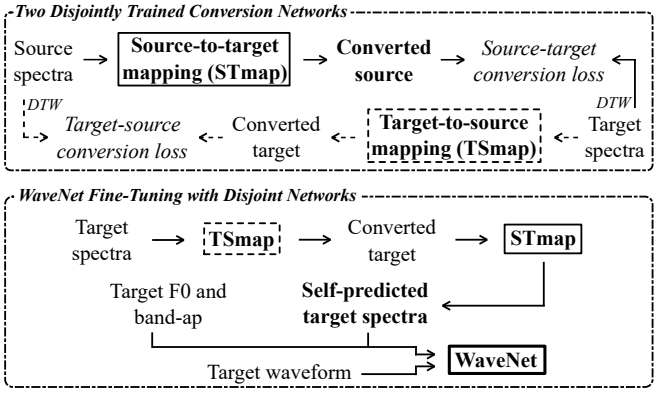

Fig. 2. WaveNet fine-tuning with disjoint networks.

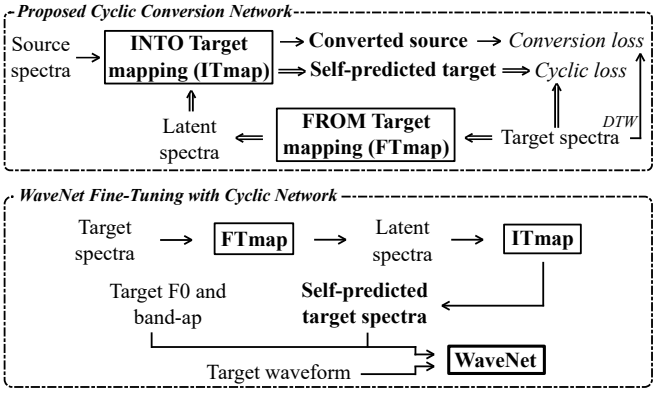

Fig. 3. WaveNet fine-tuning with proposed cyclic network.

3.2. Proposed spectral conversion network training for WaveNet fine-tuning

We propose an improvement to the network model used to generate the self-predicted target features for the fine-tuning of a WaveNet vocoder. Instead of developing two disjoint conversion networks, which was the case in [15], the two modules are trained within an integrated cyclic model by using two losses, i.e., the conversion (source-target) and the cyclic (self-predicted target), as shown in Fig. 3. To generate the self-predicted target features, both the FROM target mapping (FTmap), fed with the original target features, and the INTO target mapping (ITmap) are used. In the conversion stage, only the ITmap, fed with the source features, is used. Note that, in the training, it is very important to use a weighting constant for the cyclic loss owing to its strength compared with the conversion loss so that the estimation of self-predicted target features does not become too good compared with the estimation of converted source-target features.

This cyclic conversion network addresses two problems of the disjoint networks [15], i.e., it reduces the workload in training two separate networks and addresses the unsynchronized connection between the two separately trained networks. Moreover, the proposed model can also be regarded as a multitask learning framework. This is because both the FTmap and ITmap are optimized to improve the conversion flow while maintaining a reasonable accuracy level for the cyclic flow, i.e., through the use of the conversion and weighted cyclic losses. Therefore, within this type of framework, it is crucial to monitor the accuracy difference between the converted features, i.e., from the conversion flow, and the self-predicted features, i.e., those from cyclic flow. Note that compared with the cyclic architecture with adversarial networks (CycleGAN) [5], our proposed cyclic network is more suitable for handling parallel data. This is because our main objective is to improve the parallel VC system through the use of a cyclic structure that can be easily optimized as it does not have a discriminator, as in CycleGAN.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

To evaluate the proposed system, we used the speech dataset provided in VCC 2018 [18], which consists of six female and six male speakers, as well as an additional speech dataset from the ARCTIC database, i.e., "bdl" (male) and "slt" (female), to train a multispeaker WaveNet model [10]. For the spectral conversion networks, we used the "SF1" and "SM1" data for the source speakers and "TF1" and "TM1" for the target speakers, where "F" means female and "M" means male. The number of utterances for the VCC 2018 dataset was 81, whereas that for the ARCTIC dataset was 1132. We used the first 992 utterances of the ARCTIC dataset, as well as the last 71 of the VCC 2018 dataset for the training data. The remaining 140 utterances from the ARCTIC dataset and 10 utterances from the VCC 2018 dataset were used for the validation data. The number of utterances in the evaluation set used in the subjective evaluation was 35.

We used 35-dimensional mel-cepstrum coefficients including the 0th power as the spectral envelope parameters, which were extracted from the WORLD [19, 20] spectrum of the speech signal. As the excitation and aperiodicity features, we used framewise F0 values and two-band aperiodicity coding parameters, respectively, which were also extracted using WORLD. To perform pitch conversion, we performed a linear F0 transform based on the statistics of the speaker data. The speech signal sampling rate was 22,050 Hz. The frame shift was set to 5 ms.

The architecture and the set of auxiliary features for the WaveNet model were exactly the same as that in [15]. A multispeaker WaveNet model was trained and then fine-tuned in accordance with a particular target speaker, either conventionally, i.e., with natural spectral features, or with oversmoothed features. On the other hand, For the spectral conversion models, we used one hidden GRU layer and one GRU output layer. The number of GRU units was 1024. The convolutional input layers were designed to capture the contexts of 4 preced-

Table 1. MCD [dB] and LGD from the cyclic flow (Pred), using training data, and from the conversion (Conv), using validation, by the disjoint and the proposed cyclic networks.

	DisjPred	DisjConv	CyclePred	CycleConv
MCD	5.47	6.23	4.97	6.21
LGD	1.82	1.59	1.24	1.52

ing and 4 succeeding frames. A set of time-warping functions for computing the conversion loss were computed beforehand with dynamic time warping (DTW) algorithm after the removal of silent frames. A global variance (GV) [3] postfilter was not used for the WaveNet fine-tuning with the predicted features, and from our preliminary experiment, it was found to degrade the performance of the fine-tuning. The weight of the cyclic loss was set to 10^{-7}.

4.2. Objective evaluation

We computed the average values of the mel-cepstral distortion (MCD) and the log-GV distance (LGD) between the estimated and natural mel-cepstrum parameters. We compared the disjoint networks (Disj), which are described using the TSmap and STmap notations in Fig. 2, and the proposed cyclic network (Cycle), which is described in Fig. 3 using the FTmap and ITmap notations.

The results are shown in Table 1. The measurements of self-predicted target features of training data, i.e., for WaveNet fine-tuning, are denoted by Pred, whereas those of the converted source-to-target features using validation data are denoted by Conv. These results show that the proposed cyclic network improves the accuracy of the self-predicted target features within the training set (CyclePred) while still maintaining the distance between CyclePred and the conversion accuracy of the validation set (CycleConv). Ideally, in the future, we would like to find the accuracy range of Pred features compared to the Conv features that can be regarded as not too good and not too bad.

4.3. Subjective evaluation

We also evaluated the converted speech waveforms [1] according to their naturalness and their similarity to the natural speech of the intended target. A five-scaled mean opinion score (MOS) test was performed to assess the naturalness, i.e., from 1 (completely unnatural) to 5 (completely natural). For the speaker similarity test, listeners were given a pair of audio stimuli consisting of a natural speech of a target speaker and a converted speech of a source speaker, and asked to determine whether they can be produced by the same speaker, with the confidence of their decision, i.e., sure or not sure. We compared the combination of conventional vocoder with direct waveform modification [6, 21] and GV [3] (DiffGV), our VC system in the VCC 2018 [22] (WNDiffGV), which was fine-tuned with natural features, the disjoint conversion networks for WaveNet fine-tuning (WNFT) [15], and the proposed cyclic network (WNCycFT), where the latter two were fine-tuned with oversmoothed features. The number of listeners was 11, none of which were native English speakers. The two-tailed Mann–Whitney U test with $\alpha < 0.05$ was used to determine the statistical significance of the best system in each conversion category.

[1] Speech samples are available at http://bit.ly/2WTsbmR

Table 2. Result of MOS test for speech naturalness. $\pm$ denotes the 95% confidence interval of the sample mean. [·] denotes a sytem with a statistically significant lower value than the highest value in each conversion category.

	DiffGV	WNDiffGV	WNFT	WNCycFT
All	[2.53±0.14]	[3.06±0.13]	[3.38±0.14]	**3.79±0.14**
F-F	[2.68±0.28]	[3.18±0.27]	3.39±0.30	**3.57±0.31**
F-M	[2.05±0.25]	[2.84±0.30]	[3.23±0.31]	**3.82±0.27**
M-F	[2.66±0.25]	[2.98±0.22]	[3.14±0.26]	**3.71±0.26**
M-M	[2.75±0.34]	[3.25±0.26]	3.75±0.27	**4.07±0.27**

Table 3. Result of speaker similarity scores aggregated from "same_sure" and "same_not-sure" decisions. [·] denotes a system with a statistically significant lower value than the highest value in each conversion category.

	DiffGV	WNDiffGV	WNFT	WNCycFT
All	[37.50%]	[57.39%]	[63.64%]	**73.86%**
F-F	[38.64%]	[68.18%]	72.73%	**79.55%**
F-M	[29.55%]	[43.18%]	[45.46%]	**65.91%**
M-F	[34.09%]	[50.00%]	63.64%	**70.46%**
M-M	[47.73%]	[68.18%]	72.73%	**79.55%**

The evaluation results for naturalness and speaker similarity are given in Tables 2 and 3, respectively. These results show that by fine-tuning the WaveNet vocoder using the self-predicted target features generated from the proposed cyclic conversion network (WNCycFT), the overall quality and accuracy of the converted speech are significantly improved.

5. CONCLUSION

We have proposed a parallel VC framework using a cyclic RNN-based spectral conversion to generate self-predicted target features used in the fine-tuning of a WaveNet vocoder. The experimental results demonstrate that the proposed system significantly improves both the quality and accuracy of the converted speech, resulting in a mean opinion score of 3.79 and a speaker similarity score of 73.86%. In future work, we will extend the proposed concept to a nonparallel VC.

6. ACKNOWLEDGMENT

This work was partly supported by JST, PRESTO Grant Number JPMJPR1657, and JSPS KAKENHI Grant Number JP17H06101.

7. REFERENCES

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

[2] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, pp. 1043–1046.

[3] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[4] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, South Brisbane, Australia, Apr. 2015, pp. 4869–4873.

[5] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," Published in EUSIPCO 2018.

[6] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Commun.*, vol. 99, pp. 211–220, 2018.

[7] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[9] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1118–1122.

[10] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 712–718.

[11] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1138–1142.

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[13] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1983–1987.

[14] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3364–3368.

[15] P. L. Tobing, T. Hayashi, Y.-C. Wu, K. Kobayashi, and T. Toda, "An evaluation of deep spectral mappings and WaveNet vocoder for voice conversion," in *Proc. SLT*, Athens, Greece, Dec. 2018, pp. 297–303.

[16] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, South Brisbane, Australia, Apr. 2015, pp. 4470–4474.

[17] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[18] 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," *arXiv preprint arXiv:1804.04262*, 2018.

[19] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[20] C.-C. Hsu. [Online]. Available: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[21] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software," in *Proc. Odyssey*, Les Sables d'Olonne, France, Jun. 2018, pp. 203–210.

[22] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "NU voice conversion system for the Voice Conversion Challenge 2018," in *Proc. Odyssey*, Les Sables d'Olonne, France, Jun. 2018, pp. 219–226.