Refined WaveNet Vocoder for Variational Autoencoder Based Voice Conversion

Wen-Chin Huang*, Yi-Chiao Wu†, Hsin-Te Hwang*, Patrick Lumban Tobing†, Tomoki Hayashi†,
Kazuhiro Kobayashi†, Tomoki Toda†, Yu Tsao*, Hsin-Min Wang*

* Academia Sinica, Taiwan
†Nagoya University, Japan

Abstract—This paper presents a refinement framework of WaveNet vocoders for variational autoencoder (VAE) based voice conversion (VC), which reduces the quality distortion caused by the mismatch between the training data and testing data. Conventional WaveNet vocoders are trained with natural acoustic features but conditioned on the converted features in the conversion stage for VC, and such a mismatch often causes significant quality and similarity degradation. In this work, we take advantage of the particular structure of VAEs to refine WaveNet vocoders with the self-reconstructed features generated by VAE, which are of similar characteristics with the converted features while having the same temporal structure with the target natural features. We analyze these features and show that the self-reconstructed features are similar to the converted features. Objective and subjective experimental results demonstrate the effectiveness of our proposed framework.

Index Terms—voice conversion, variational autoencoder, WaveNet vocoder, speaker adaptation

Fig. 1: A general mismatch between the training and conversion features of the WaveNet vocoder, which is highlighted in the orange box.

I. INTRODUCTION

Voice conversion (VC) aims to convert the speech from a source to that of a target without changing the linguistic content. While numerous approaches have been proposed [1]–[7], in this work we study variational autoencoder [8] based VC (VAE-VC) [9], where the conversion framework consists of an encoder-decoder pair. First, the source features are encoded into a latent content code by the encoder, and the decoder generates the converted features by conditioning on the latent code and the target speaker code. The whole network is trained in an unsupervised manner, minimizing the reconstruction error and a Kullback-Leibler (KL)-divergence loss, which regularizes the distribution of the latent variable. Because of the unsupervised learning nature of VAE, VAE-VC does not need parallel training data.

The waveform generation process plays an important role in a VC system. Conventional VC frameworks employ parametric vocoders as their synthesis module, which impose many overly simplified assumptions that discard the phase information and result in unnatural excitation signals, and thus cause a significant degradation in the quality of the converted speech [10]–[12]. In recent years, the WaveNet vocoder [13], [14], built upon one of the most promising neural speech generation models, WaveNet [15], has been proposed. It is capable of reconstructing the phase and excitation information, and thus generates more natural sounding speech.

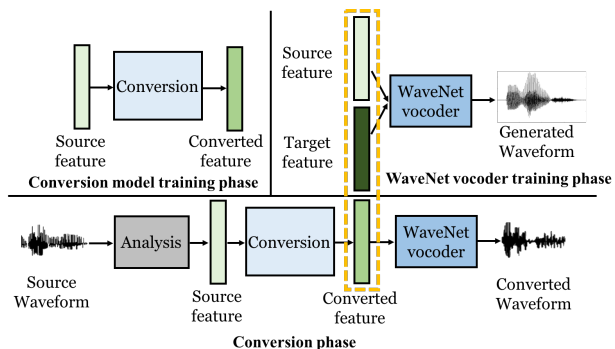In [14], it was confirmed that compared with the speaker independent (SI) WaveNet vocoder, which is trained with a multi-speaker speech corpus, the speaker dependent (SD) variant [13] is superior in terms of speech quality. However, it is impractical to collect a sufficient amount of data (usually more than 1 hour) required to build a target-dependent WaveNet vocoder in a VC framework. To alleviate this problem, many researchers have applied a fine-tuning technique [16]–[21] and its effectiveness has been confirmed in [18].

However, the above mentioned fine-tuning technique suffers from the mismatch between the training and testing input features of the WaveNet vocoder, as shown in Fig. 1. The features used to train the WaveNet vocoder are extracted from either the source's or the target's natural speech, which we will refer to as the *natural* features, and are different from the *converted* features generated by the conversion model. One way to reduce this mismatch is to use the converted features to fine-tune the WaveNet vocoder, similar to [22]. However, this cannot be used directly in VC, since the time alignment between the converted features and the target speech waveform will be required, and an imperfect alignment will introduce an additional mismatch.

To generate alignment-free data, [23] first used intra-speaker conversion, but the performance was limited since there was a gap between such intra-converted features and the actual converted features. Furthermore, [21] used a cyclic conversion model to ensure the self-predicted features were indeed similar to the converted features. Although these methods seem to be promising in parallel VC, the use of such techniques in non-parallel VC has not been properly investigated. For instance, it remains unknown how to effectively apply such techniques
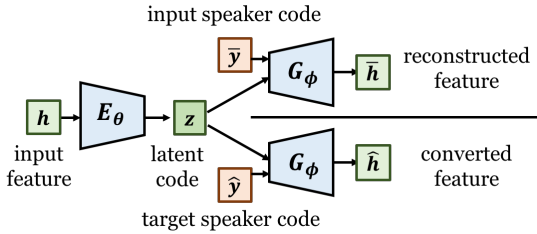
Fig. 2: Illustration of the two types of spectral frames generated by feeding different speaker codes in the forward process of VAE-VC.
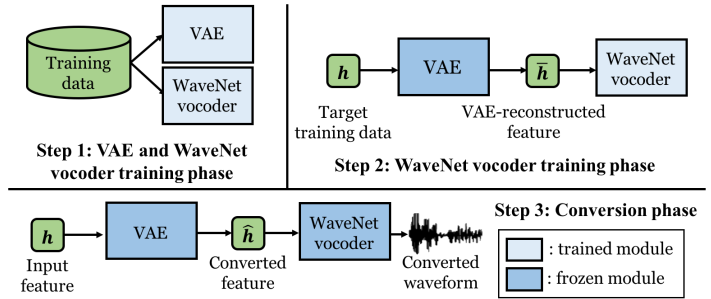

Fig. 3: The proposed WaveNet vocoder fine-tuning framework using *VAE-reconstructed* features. The analysis module is omitted.

to VAE-VC.

In light of this, in this work, we propose a novel WaveNet vocoder fine-tuning framework for VAE-VC. Specifically, the self-reconstructed features of the trained VAE network are used to fine-tune the WaveNet vocoder. Our contributions are:

- The self-reconstructed features of VAE are suitable for fine-tuning. Since the forward processes of the VAE-VC are almost identical except that they are conditioned on different speaker codes, we hypothesize that the characteristics of the self-reconstructed features are similar to that of the converted features. We analyzed the properties of these features and verified this hypothesis with objective measures. To our knowledge, this is the first attempt to examine the "suitableness" of the features used for fine-tuning.
- The self-reconstructed features of VAE are innately time-alignment free, since the temporal structure of self-reconstructed and natural target features is identical. Therefore, we don't need to add additional networks or losses to solve the alignment issue, as in [21], [23].

II. RELATED TECHNIQUES

A. VAE-VC

In VAE-VC [9], the conversion function is formulated as an encoder-decoder network. Specifically, given an observed (source or target) spectral frame $\boldsymbol{h}$, a speaker-independent encoder E_θ with parameter set θ encodes $\boldsymbol{h}$ into a latent code: $\boldsymbol{z} = E_\theta(\boldsymbol{h})$. The (target) speaker code $\hat{\boldsymbol{y}}$ is then concatenated with the latent code, and passed to a conditional decoder G_ϕ with parameter set ϕ to generate the *converted* features. Thus, the conversion function f of VAE-VC can be expressed as: $\hat{\boldsymbol{h}} = f(\boldsymbol{h}) = G_\phi(\boldsymbol{z}, \hat{\boldsymbol{y}})$. Fig. 2 shows the two types of spectral frames used in training and conversion. In the training phase, the forward pass is done by encoding $\boldsymbol{h}$ of an arbitrary speaker into a latent code $\boldsymbol{z}$, and feeding it into the decoder along with the speaker code $\bar{\boldsymbol{y}}$ to generate the reconstructed frame $\bar{\boldsymbol{h}}$, which we will refer to as the *VAE-reconstructed* feature.

During training, there is no need for parallel training data. The model parameters can be obtained by maximizing the variational lower bound, $\mathcal{L}$:

$$\mathcal{L}(\theta, \phi; \boldsymbol{h}, \boldsymbol{y}) = \mathcal{L}_{recon}(\boldsymbol{h}, \boldsymbol{y}) + \mathcal{L}_{lat}(\boldsymbol{h}), \quad (1)$$

$$\mathcal{L}_{recon}(\boldsymbol{h}, \boldsymbol{y}) = \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{h})}\big[\log p_\phi(\boldsymbol{h}|\boldsymbol{z}, \boldsymbol{y})\big], \quad (2)$$

$$\mathcal{L}_{lat}(\boldsymbol{h}) = -D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{h})\|p(\boldsymbol{z})), \quad (3)$$

where $q_\theta(\boldsymbol{z}|\boldsymbol{h})$ is the approximate posterior, $p_\phi(\boldsymbol{h}|\boldsymbol{z}, \boldsymbol{y})$ is the data likelihood, and $p(\boldsymbol{z})$ is the prior distribution of the latent space. $\mathcal{L}_{recon}$ is simply a reconstruction term as in any vanilla autoencoder, whereas $\mathcal{L}_{lat}$ regularizes the encoder to align the approximate posterior with the prior distribution.

B. WaveNet vocoder

WaveNet [15] is a deep autoregressive network, which generates high-fidelity speech waveforms sample-by-sample, using the following conditional probability equation:

$$P(\boldsymbol{X}|\boldsymbol{h}) = \prod_{n=1}^{N} P(x_n|x_{n-r}, \ldots, x_{n-1}, \boldsymbol{h}), \quad (4)$$

where x_n is the current sample point, r is the size of the receptive field, and $\boldsymbol{h}$ is the auxiliary feature vector. The WaveNet vocoder [13], [14] reconstructs time-domain waveforms by conditioning on acoustic features, including spectral features, fundamental frequencies (f_0) and aperiodicity signals (APs). A WaveNet vocoder is composed of several stacked residual blocks skip-connected to the final output, with each residual block containing a 2×1 dilated causal convolution layer, a gated activation function, and two 1×1 convolution layers. In short, given a data sequence pair $(\boldsymbol{h}, \boldsymbol{x})$, the WaveNet vocoder is trained in a supervised fashion, learning to map the acoustic features $\boldsymbol{h}$ to the time-domain signals $\boldsymbol{x}$.

In recent years, fine-tuning a WaveNet vocoder to the target speaker has become a popular technique to improve speech quality when applying the WaveNet vocoder to VC [16]–[21]. Specifically, an initial SI WaveNet vocoder is first trained with a multi-speaker dataset. Then, the target speaker's data is used to fine-tune all or partial model parameters. Therefore, suitable fine-tuning data that matches the converted features is the key to better speech quality.

C. VAE-VC with WaveNet vocoder

A combination of VAE-VC and WaveNet vocoder has been proposed in [19]. First, a VAE conversion model is trained. A WaveNet vocoder is then constructed by training with natural features of a multi-speaker dataset first, followed by fine-tuning using the natural features of the target speaker. In the conversion phase, the VAE model first performs spectral
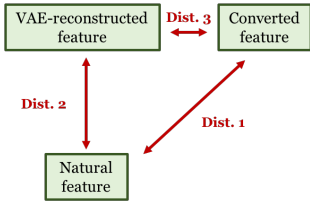
Fig. 4: An illustration of the distances between the three types of features.
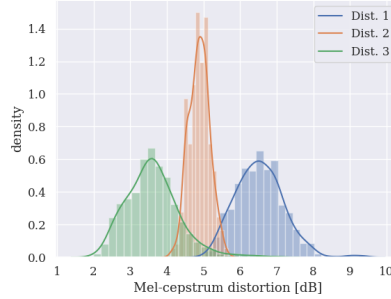

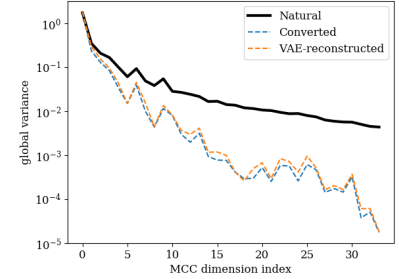Fig. 5: Density plot of the mel-cepstral distortion of the three distances.

Fig. 6: Global variance values of the three types of features from speaker TF1.

feature conversion, and by conditioning on the converted spectral features, source APs and transformed f_0, the WaveNet vocoder generates the converted speech waveform. Here we once again point out that the WaveNet vocoder conditions on features of different characteristics in the training and conversion phases, i.e. there is a mismatch between the input features in the training and conversion phases. In the following section, we will introduce our proposed method for reducing this mismatch.

III. PROPOSED METHOD FOR WAVENET VOCODER FINE-TUNING IN VAE-VC

Using the converted features to fine-tune the WaveNet vocoder is an effective approach to reduce the mismatch between training and conversion in VC. However, this is only possible when a parallel corpus is available, which conflicts with the capability of VAE-VC under non-parallel conditions. Even with parallel data, since conventional VC models perform conversion frame-by-frame, the converted feature and the source speech have the same temporal structure, which is different from that of the target. As a result, we need an elegant fine-tuning method for a non-parallel VC method like VAE-VC that solves both the time-alignment and mismatch issues.

We propose to fine-tune WaveNet vocoders with VAE-reconstructed features by taking advantage of the particular network structure of VAEs, as depicted in Fig. 3. Specifically, a VAE model is first trained with the whole training corpus. Then, the VAE-reconstructed features $\bar{h}$ could be obtained from the target speaker's training data through the reconstruction process described in Section II-A. We can thereafter use data pairs $(\bar{h}, x)$ to fine-tune the SI WaveNet vocoder instead of using the original (h, x) as described in Section II-B, where h and x denote the natural target feature and the target waveform, respectively.

There are two main advantages of using the VAE-reconstructed features to fine-tune the SI WaveNet vocoder. First, since h and $\bar{h}$ have the same temporal structure, time-alignment is no longer required. More importantly, the mismatch between the VAE-reconstructed features and the converted features $\hat{h}$ is small. Recall that in VAE-VC, $\bar{h}$ and $\hat{h}$ can be obtained by feeding the input speaker code $\bar{y}$ and the target speaker code $\hat{y}$, respectively. Since $\bar{h}$ and $\hat{h}$ only differ in speaker codes, we hypothesize that they share similar properties. Experimental results confirm the above hypothesis, as we will show in Section IV-B.

IV. EXPERIMENTAL EVALUATIONS

A. Experimental settings

All experiments were conducted using the Voice Conversion Challenge 2018 (VCC2018) corpus [24], which included recordings of 12 professional US English speakers with a sampling rate of 22050 Hz and a sample resolution of 16 bits. The training set consisted of 81 utterances and the testing set consisted of 35 utterances of each speaker. The total number of training utterances was 972 and the training data length was roughly 54 minutes. The WORLD vocoder [12] was adopted to extract acoustic features including 513-dimensional spectral envelopes (SPs), 513-dimensional APs and f_0. The SPs were normalized to unit-sum, and the normalizing factor was taken out and thus not modified. 35-dimensional mel-cepstral coefficients (MCCs) were extracted from the SPs.

For the spectral conversion model, we used the CDVAE variant [25] as it was capable of modeling MCCs while the original VAE [9] failed to do so [25]. The CDVAE model was trained with the whole training data of the 12 speakers from VCC2018, with the detailed network architecture and training hyper-parameters consistent with [25].

For the WaveNet vocoder, we followed the official implementation[1] with the default settings, including the noise shaping technique [26]. All training data of VCC2018 were used to train the multi-speaker WaveNet vocoder. Although some works used discriminate speaker embeddings for training, we did not adopt such technique since simply mixing all speakers gave better performance as reported in [14]. The default 200000 iterations of model training led to approximate 2.0 training loss (in terms of negative log-likelihood). Fine-tuning of the WaveNet vocoder was performed by updating the whole network until the training loss reached around 1.0. At conversion, the input features included the converted MCCs from the VAE conversion model followed by energy compensation, the linear mean-variance transformed log-f_0 as well as the unmodified source APs.

---

[1] https://github.com/kan-bayashi/PytorchWaveNetVocoder

TABLE I: The compared methods. GV stands for the global variance postfilter [27].

| Name	Vocoder	GV	Training feature	Adapting feature
Baseline 1	WORLD	No	None	None
Baseline 2	WORLD	Yes	None	None
Baseline 3	WaveNet	No	*Natural*	*Natural*
Baseline 4	WaveNet	Yes	*Natural*	*Natural*
Proposed 1	WaveNet	No	*Natural*	*VAE-reconstructed*
Proposed 2	WaveNet	Yes	*Natural*	*VAE-reconstructed*+GV
Upper bound	WaveNet	–	*Natural*	*Natural*

B. Analysis of Mismatch

In order to prove that the VAE-reconstructed features are more similar to the converted features than the natural target features, we analyzed their spectral properties based on the mean mel-cepstral distortion (MCD) and global variance (GV) measurements on the training set. First, we trained a VAE and obtained the three types of features using the training data of the VCC2018 corpus. Then, we calculated the distances (Dist. 1-3) by calculating the mean MCD of each sentence and then comparing the distributions, as illustrated in Fig. 4.

Fig. 5 shows the three distances calculated from conversion pairs of a subset of speakers (SF1, SM1 to TF1, TF2, TM1, TM2). First, we observe that Dist. 2 is rather large, showing that the natural features are somehow distorted through the imperfect reconstruction process of VAE. We further observe that Dist. 3 is smaller than Dist. 1, demonstrating that conventional approaches suffer from the mismatch given by Dist. 1, and our proposed method can alleviate the issue because Dist. 3 is small. Note that Dist. 1 and Dist. 3 are affected by time alignment while Dist. 2 is not, so the former have larger variation than the latter.

Fig. 6 illustrates the GVs of the three types of features. The GVs of the natural features are higher than that of the other two types of features, since the natural features contain the most spectral details and structures while the VAE-reconstructed and converted features suffer from the over-smoothing effect. These results imply that the VAE-reconstructed features are more suitable to fine-tune the WaveNet vocoder as they well simulate the properties of the converted features.

## C. Evaluation of the proposed methods

We conducted objective and subjective evaluations of seven methods, as shown in Table I. The baseline methods (Baseline 1-4) are the VAE-VC systems with the WORLD and the fine-tuned WaveNet vocoders using the natural features [19], [25]. The proposed methods (Proposed 1-2) are the VAE-VC systems with the WaveNet vocoder fine-tuned with the VAE-reconstructed and GV post-filtered VAE-reconstructed features, respectively. The upper bound system is the WaveNet vocoder fine-tuned with the target natural features, i.e. without the conversion process. Note that the GV column indicates whether the GV post-filter is applied to the converted spectral features before sending them into the WaveNet vocoder.
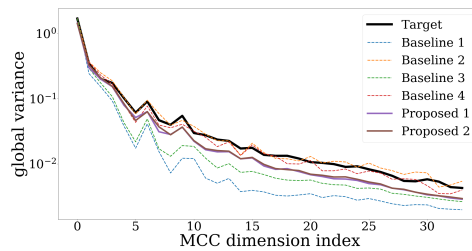
(a) Global variance values of mel-cepstrum coefficients extracted from speech converted to speaker TF1 and the natural TF1 speech.
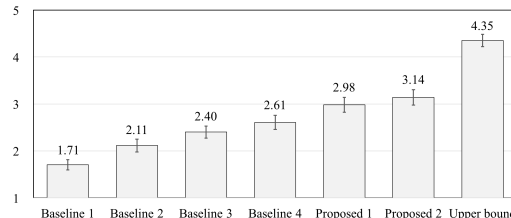
(b) Mean opinion score on naturalness over all speaker pairs. Error bars indicate the 95% confidence intervals.
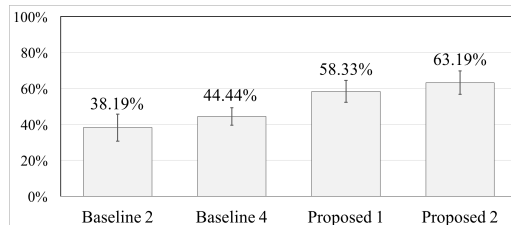
(c) Result of speaker similarity scores, which were aggregated from "same sure" and "same not-sure" decisions. Error bars indicate the 95% confidence intervals.

Fig. 7: Results of objective and and subjective evaluations.

Fig. 7(a) shows the GV over the MCCs extracted from the converted voices of the compared methods. In the VC literature, the GV values reflect perceptual quality and intelligibility, and we generally seek for a GV curve close to that of the target. In light of this, the following are some observations from the graph:

- Baselines 2 and 4 had GV very close to that of the natural target, compared to Baselines 1 and 3, which is consistent with the results of combining the GV post-filtering method with the WaveNet vocoder as in [23].
- The GV of Baseline 3 was far from that of the natural target compared to that of Baseline 4, possibly because the post-filtered converted features are closer to the natural features used to train/fine-tune the WaveNet vocoder.
- Proposed 1 had higher GV values than Baseline 3. This indicates that the WaveNet vocoder acts like a post-filter that compensates the over-smoothed converted features in our proposed framework.
- Nonetheless, the GV values of Proposed 1 and Proposed 2 were nearly identical and still lower than that of the natural target. We speculate that since we penalize WaveNet by training it using distorted features, degradation in speech quality might still be introduced.

Subjective tests were conducted to evaluate the naturalness

and similarity. A 5-scaled mean opinion score (MOS) test was performed to assess the naturalness. For the speaker similarity, each listener was given a pair of audio stimuli, consisting of a natural speech of a target speaker and a converted speech, and asked to judge whether they were produced by the same speaker, with a confidence decision, i.e., sure and not sure. A subset of speakers (SF1, SM1, SF3, SM3, TF1, TM1, TF2, TM2) were chosen. Ten subjects were recruited.

Fig. 7(b) shows that our proposed methods outperformed all baseline systems, even though the proposed methods had smaller GV values compared with Baselines 2 and 4. It could be inferred that the proposed methods not only alleviated the over-smoothing problem of the converted features but also improved other aspects of the generated speech, which might correspond to the naturalness of speech. Furthermore, Fig. 7(c) indicates that our proposed methods also improved the speaker similarity of the generated speech.

V. Conclusion

In this work, we have proposed a refinement framework of VAE-VC with the WaveNet vocoder. We utilize the self-reconstruction procedure in the VAE-VC framework to fine-tune the WaveNet vocoder to alleviate the performance degradation issue caused by the mismatch between the training phase and conversion phase of the WaveNet vocoder. Evaluation results show the effectiveness of the proposed method in terms of naturalness and speaker similarity. In the future, we plan to investigate the use of VAE-reconstructed features in the multi-speaker training stage of the WaveNet vocoder to further improve the robustness. Speech samples are available at https://unilight.github.io/VAE-WNV-VC-Demo/

VI. Acknowledgement

References

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

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

[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, July 2010.

[4] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, Dec 2014.

[5] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. SLT*, 2012, pp. 313–317.

[6] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, Oct 2014.

[7] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. Interspeech*, 2016, pp. 1652–1656.

[8] D. P. Kingma and Max W., "Auto-encoding variational bayes.," *CoRR*, vol. abs/1312.6114, 2013.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APISPA ASC*, 2016, pp. 1–6.

[10] B. S. Atal and Suzanne L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[11] 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 Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[12] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. 99, pp. 1877–1884, 2016.

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

[14] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 712–718.

[15] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[16] P. L. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "NU voice conversion system for the voice conversion challenge 2018," in *Proc. Odyssey*, 2018, pp. 219–226.

[17] Y. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the voice conversion challenge 2018," in *Proc. Odyssey*, 2018, pp. 211–218.

[18] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted WaveNet vocoder," in *Proc. Interspeech*, 2018, pp. 1978–1982.

[19] W.-C. Huang, C.-C. Lo, H.-T. Hwang, Y. Tsao, and H.-M. Wang, "WaveNet vocoder and its applications in voice conversion," in *Proc. The 30th ROCLING Conference on Computational Linguistics and Speech Processing (ROCLING)*, Oct 2018.

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

[21] P. L. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6815–6819.

[22] J. Shen, R. Pang, R. J. Weiss, M Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, Rj Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

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

[24] 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*, 2018, pp. 195–202.

[25] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Voice conversion based on cross-domain features using variational auto encoders," in *Proc. ISCSLP*, 2018.

[26] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5664–5668.

[27] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implemen global variance in statistical speech synthesis," in *Proc. Interspeech*, 2012, pp. 1436–1439.