INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Speaker-Dependent WaveNet for Voice Conversion with Non-Parallel Data

Xiaohai Tian[1,2], *Eng Siong Chng*[2] *and Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore
[2]School of Computer Engineering, Nanyang Technological University, Singapore

eletia@nus.edu.sg, aseschng@ntu.edu.sg, haizhou.li@nus.edu.sg

Abstract

In a typical voice conversion system, vocoder is commonly used for speech-to-features analysis and features-to-speech synthesis. However, vocoder can be a source of speech quality degradation. This paper presents a novel approach to voice conversion using WaveNet for non-parallel training data. Instead of reconstructing speech with intermediate features, the proposed approach utilizes the WaveNet to map the Phonetic PosteriorGrams (PPGs) to the waveform samples directly. In this way, we avoid the estimation errors arising from vocoding and feature conversion. Additionally, as PPG is assumed to be speaker independent, the proposed approach also reduces the feature mismatch problem in WaveNet vocoder based solutions. Experimental results conducted on the CMU-ARCTIC database show that the proposed approach significantly outperforms the traditional vocoder and WaveNet Vocoder baselines in terms of speech quality.

Index Terms: Voice conversion, WaveNet, Non-parallel data

1. Introduction

Voice conversion (VC) aims to modify the source speaker's voice to sound like that of the target speaker without changing the linguistic content. The challenge is to transform the speaker identity while maintaining the speech quality.

Various techniques have been proposed to convert spectral feature for speaker identity conversion. Among them, Gaussian mixture model (GMM) [1, 2, 3, 4] is one of the most popular methods, where the spectral feature is transformed by a statistical parametric model. However, it is known the GMM method doesn't capture the spectral details thus suffers from over-smoothing problem [3, 5]. To address these problems, frequency warping [5, 6, 7, 8] and exemplar based methods [9, 10] are also studied. More recently, with good regression performance, neural network methods are widely used in VC task, e.g. deep neural network (DNN) [11, 12, 13], long short-term memory (LSTM) [14] and generative adversarial networks (GAN) [15, 16].

Despite the research progress, the quality of converted speech varies at run-time. One reason is that most of the existing techniques perform the speaker identity conversion and speech reconstruction on the features analyzed by parametric vocoders. Conventional parametric vocoders (STRAIGHT [17] and WORLD [18]) are designed based on certain assumptions, e.g. source filter model, time invariant linear filter. Additionally, to simplify mathematical formulation of the parametric model, some information, e.g. the phase information, are usually discarded. As a result, the artifacts are introduced in both speech-to-features analysis stage and features-to-speech synthesis stage. To address the features-to-speech synthesis issue, a direct waveform modification technique based on spectrum differential is proposed in [19]. This method is able to generate high quality speech for intra-gender conversion. However, the performance is moderate for inter-gender speaker pairs due to the $f0$ transformation [20]. Recently, WaveNet [21] is proposed to directly estimate the time domain waveform samples conditioned on input features. A WaveNet vocoder [22, 23, 24], reconstructing waveforms from acoustic features extracted by parametric vocoders, is showed to improve the generated speech quality significantly. Its effectiveness has been demonstrated in several voice conversion studies [20, 25, 26, 27] to replace the traditional vocoders for high quality speech generation. However, in these approaches, the WaveNet vocoder is usually trained with acoustic features extracted from natural speech, while the converted acoustic features are used for run-time generation. This feature mismatch between the training and generation may result in undesired noise-like signals, which are observed in the WaveNet generated speech as reported in [20, 25, 28].

In this paper, we introduce a voice conversion approach using WaveNet for non-parallel training data, where the traditional parametric vocoder is not required for both intermediate spectral feature extraction and speech reconstruction. Inspired by [29, 30], the proposed method first encodeds a speech signal into speaker independent (SI) feature representations, e.g. Phonetic PosteriorGrams (PPG) [29, 30]. Then, a WaveNet is trained to predict the corresponding time-domain speech signals with SI features as the local conditioning. At run-time, the same SI features extracted by given speech are used to drive the WaveNet to generate the converted speech. Note that, the conversion model is trained between SI features and the corresponding time-domain speech signals of the same speaker. Hence, the parallel data is not required for the proposed method.

We make three contributions in this paper.

1) We propose a voice conversion framework that doesn't require a traditional parametric vocoder.
2) By doing so, we avoid the feature extraction and speech reconstruction errors arising from the parametric vocoder, and achieve better voice quality than state-of-the-art systems.
3) Bypassing the the intermediate vocoder features for conversion, the proposed approach reduces the feature mismatch problem of WaveNet vocoder based VC approaches.

2. Voice Conversion with WaveNet Vocoder

In this section, we discuss the advantages and limitations of the WaveNet vocoder based voice conversion techniques.

2.1. WaveNet Vocoder

WaveNet vocoder [22] is a conditional WaveNet [21]. It can reconstruct the time-domain audio signals conditioned on the

http://dx.doi.org/10.21437/Interspeech.2019-1514

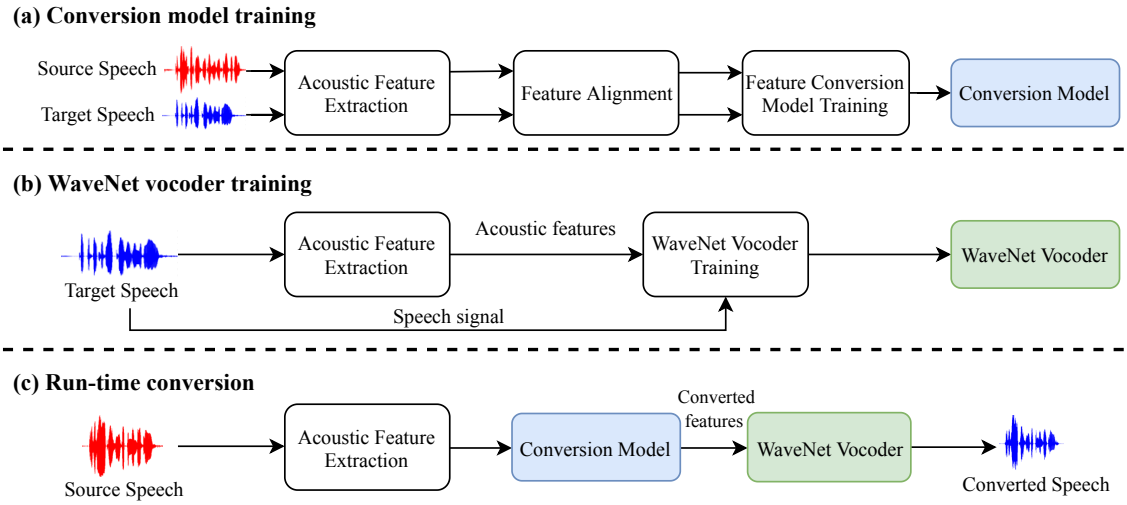

Figure 1: *Block diagram of voice conversion system with WaveNet vocoder.*

acoustic features extracted from traditional vocoders, e.g. aperiodicity, f0 and spectral features. Given a waveform sequence $x = [x_0, x_1, ..., x_T]$ and the additional local conditioning input $\mathbf{h}$, WaveNet vocoder can model the conditional distribution $p(x|\mathbf{h})$ as follows:

$$p(x|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, x_2, \cdots, x_{t-1}; \mathbf{h}). \quad (1)$$

In order to model the long-range temporal dependencies of audio samples, an architecture based on dilated causal convolutions and a gated activation unit is proposed. Deep residual learning framework is also utilized to speed up the convergence and train a deep model (e.g. 30 layers). For the i-th residual block, the gated activation function is expressed as:

$$\mathbf{z}_i = \tanh(\mathbf{W}_{f,i} * \mathbf{x} + \mathbf{V}_{f,i} * \mathbf{h}) \circ \sigma(\mathbf{W}_{g,i} * \mathbf{x} + \mathbf{V}_{g,i} * \mathbf{h}), \quad (2)$$

where $*$ and $\circ$ denote the convolution and element-wise product operator respectively. $\mathbf{W}$ and $\mathbf{V}$ are the trainable convolution filters, f and g denote the filter and gate, respectively.

2.2. The Limitations

The WaveNet vocoder has been adopted in voice conversion tasks [20, 25, 26, 27] to replace the traditional vocoders for high quality speech generation. One of the successful example is proposed in [20], where the WaveNet vocoder is integrated in a GMM based VC framework. Fig. 1 (a) and (b) show its conversion model and WaveNet vocoder training processes, while Fig. 1 (c) shows its conversion process. During training, two models are built. The GMM model is trained between the aligned source and target feature pairs for feature conversion. While, a WaveNet vocoder is trained with the acoustic features, e.g. spectral feature $\mathbf{F}_{\text{mgc}}$, pitch $\mathbf{F}0$, voiced/unvoiced flag (vuv) $\mathbf{F}_{\text{vuv}}$ and aperiodicity $\mathbf{F}_{\text{AP}}$, extracted from original target speech as the local conditioning input for speech generation. At runtime, the acoustic features extracted by the traditional vocoders, e.g. $\mathbf{F}_{\text{mgc}}$, $\mathbf{F}0$, are first converted by GMM VC model. Converted $\mathbf{F}'_{\text{mgc}}$ and $\mathbf{F}0'$ features with original vuv and aperiodicity are then used as the additional input of the WaveNet vocoder to generate the converted speech.

While the WaveNet vocoder based VC is able to generate high quality speech, the converted features, especially spectral features $\mathbf{F}'_{\text{mgc}}$, used in run-time generation are very different from the original target features $\mathbf{F}_{\text{mgc}}$ used for training, which results in the noise-like signals or irregular impulses in some speech segments [28]. Similar findings are also reported in recent studies [20, 25, 28].

3. Converting PPG to Target Speech with WaveNet

Phonetic PosteriorGrams (PPG) based voice conversion [29, 30] has been proposed to model the relationship between PPG features to acoustic features. PPG is a sequence of probability vectors estimated by an automatic speech recognition (ASR) system. As the ASR system is designed to generate the outputs invariant to the input speaker, the PPG feature is considered to be speaker independent, that can be used to represent the linguistic content in voice conversion. In this paper, we investigate the effectiveness of using PPG as a local conditioning input of a WaveNet for voice conversion. Rather than separating the conversion and generation model in WaveNet vocoder based VC, the proposed approach takes speaker independent PPG features as WaveNet input to generate the target speaker's voice directly. In this way, we achieve the speaker identity conversion and speech generation with a single WaveNet, and avoid the estimation errors arising from feature conversion model mentioned in Section 2.2.

The proposed framework is presented in Fig. 2, which consists of two steps: (a) WaveNet conversion model training and (b) run-time conversion. The details will be described as follows.

Fig. 2(a) shows the WaveNet conversion model training process. Given speech data of the target speaker, we first extract PPGs $\mathbf{L} \in \mathbb{R}^{D \times N}$, where, D and N are the feature dimension and frame number respectively. In order to control the prosody of generated speech, $f0$ and voiced/unvoiced flag (vuv) features are also extracted, denoted as $\mathbf{F}0 \in \mathbb{R}^{1 \times N}$ and $\mathbf{F}_{\text{vuv}} \in \mathbb{R}^{1 \times N}$, respectively. To facility the WaveNet training, the PPGs, $f0$ and vuv are extended to match the temporal resolution of the time domain signals, denoted as $\widehat{\mathbf{L}} \in \mathbb{R}^{D \times T}$, $\widehat{\mathbf{F}0} \in \mathbb{R}^{1 \times T}$ and

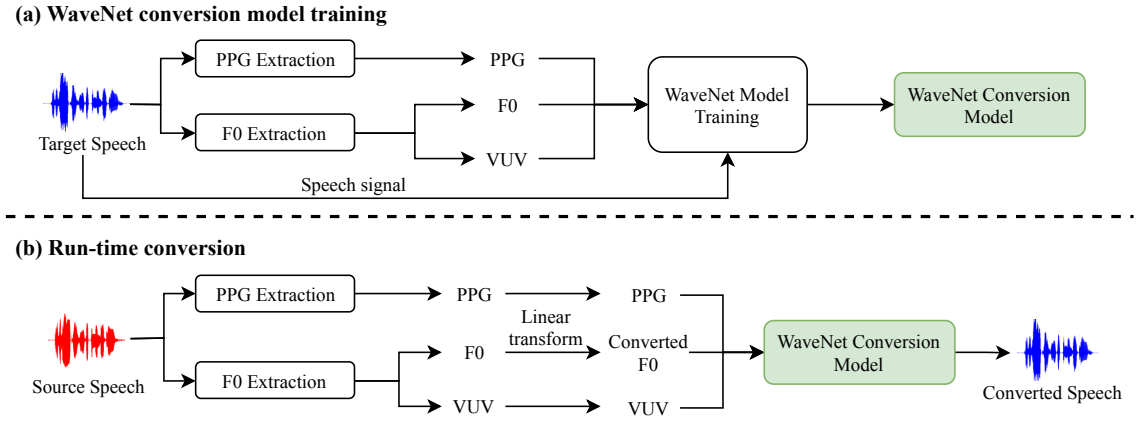

Figure 2: *Block diagram of the proposed WaveNet voice conversion approach, where speaker independent PPG is used as conditional input of WaveNet to generate target speaker's voice.*

$\widehat{\mathbf{F}_{\text{vuv}}} \in \mathbb{R}^{1\times T}$. Then, the local conditioning input $\mathbf{h}$ in Eq.(2) can be expressed as $\mathbf{h} = [\widehat{\mathbf{L}}^\top, \widehat{\mathbf{F0}}^\top, \widehat{\mathbf{F}_{\text{vuv}}}^\top]^\top$.

At run-time (see Fig. 2(b)), given a source speech, we extract the PPG, $f0$ and vuv features. A linear transformation is applied on the extracted $f0$, expressed as:

$$f0'_y = \exp((\log\widehat{f0_x} - \mu_x)\frac{\sigma_y}{\sigma_x} + \mu_y), \tag{3}$$

where μ_x and σ_x are the mean and variance of the input source speech sample's $f0$ in logarithmic domain, respectively. μ_y and σ_y are the mean and variance of the target speaker's $f0$ in logarithmic domain over all training samples. $f0'_y$ is the converted $f0$ of the target speaker. Then we adjust the temporal resolution of the PPG and feed them into the trained WaveNet conversion model to generate converted speech. As PPG is considered to be speaker independent, feature mismatch only appears between target and converted $f0$. This reduces the generation problem of WaveNet vocoder based approaches as mentioned in Section 2.2.

4. Experimental Setup

4.1. Database and feature extraction

The voice conversion experiments were conducted on the CMU-ARCTIC database [31]. Four speakers were selected consisting of two male speakers, *bdl* and *rms*, and two female speakers, *slt* and *clb*. Intra-gender and inter-gender conversions were conducted between following pairs: *rms* to *bdl* (M2M), *clb* to *slt* (F2F), *clb* to *bdl* (F2M) and *rms* to *slt* (M2F). 500 utterances were used for training, another 20 non-overlap utterances of each speaker were used for evaluation.

WORLD vocoder [18] was used to extract the 513-dimensional spectrum, 1-dimensional aperiodicity coefficients and F_0 with 5 ms frame step. Then 40-dimensional MCCs were calculated from the spectrum using Speech Signal Processing Toolkit (SPTK) [1]. The 42-dimensional phonetic posteriorgram (PPG) features were extracted by the PPG extractor trained on the Wall Street Journal corpus (WSJ) [32]. The detailed information can be found in [30]. All the audio files were resampled at 16 kHz.

[1] https://sourceforge.net/projects/sp-tk/

4.2. Baselines and setup

The details of reference and the proposed WaveNet voice conversion approaches were introduced as follows.

4.2.1. Reference Systems

- **GMM-WORLD**: We implemented the joint-density Gaussian mixture model with maximum likelihood parameter conversion [2] for feature conversion. The WORLD vocoder was used for speech generation. The source and target MCC features were aligned using dynamic time warping (DTW) [33]. Both static and its dynamic features were used in this implementation. The mixtures number of GMM is set to 128.
- **GMM(GV)-WORLD**: We use the same setting as GMM-WORLD, and the converted MCC features were enhaced by GV processing as proposed in [34].
- **GMM-WaveNet**: We use the same setting as GMM-WORLD with WaveNet vocoder for speech generation.
- **GMM(GV)-WaveNet**: We use the same setting as GMM(GV)-WORLD with WaveNet vocoder for speech generation.

4.2.2. The Proposed WaveNet VC

- **WaveNet-PPG**: The proposed WaveNet based voice conversion system with non-parallel data. The 42-dimensional PPG was used as the local condition of the WaveNet.
- **WaveNet-VC**: The proposed WaveNet based voice conversion system with non-parallel data. The 42-dimensional PPG, voiced/unvoiced flag and converted $f0$ were used as the local condition of the WaveNet. In total, the feature dimension was 44.

We trained the WaveNet vocoder and WaveNet conversion models for each target speaker. Both WaveNet vocoder and WaveNet conversion models shared the same network architecture. The WaveNet consisted of 3 dilated residual blocks. Each residual block contained of 10 dilated causal convolution layers. In each block, the dilation started from 1 and exponentially increased by a factor of 2. The hidden units of residual connection and gating layers was set to 512, while the skip connection channels was set to 256. The networks were trained using the Adam optimization method with a constant learning rate of

0.0001. The mini batch size was 15,000 samples and the training steps was set to 200,000. The waveform sample values were encoded by 16 bits μ-law.

5. Evaluations

5.1. Objective evaluation

We conducted objective evaluation to assess the effectiveness of WaveNet-VC approach. Root Mean Square Error (RMSE) was employed as the objective measure the distortion between target and converted speech. Magnitude features were extracted every 5ms with a window of 25ms. FFT length is set to 512. RMSE of j^{th} frame was calculated as: RMSE[dB] $= \sqrt{\frac{1}{F}\sum_{f=1}^{F}(20 * \log_{10}(\frac{|Y(f)_i|}{|Y(f)_i^{\text{conv}}|}))^2}$, where $Y(f)_i$ and $Y(f)_i^{\text{conv}}$ are the i^{th} magnitude features of target and converted speech respectively. F is the total number of the frequency bins. A lower RMSE indicates the smaller distortion.

Table 1: *Comparison of the Root Mean Square Errors (RMSEs) between the proposed WaveNet VC and the reference systems.*

Conversion Method	Intra	Inter	Average
GMM-WORLD	11.36	11.39	11.38
GMM(GV)-WORLD	11.90	11.99	11.94
GMM-WaveNet	13.52	13.46	13.49
GMM(GV)-WaveNet	13.85	14.06	13.96
WaveNet-PPG	14.32	14.89	14.61
WaveNet-VC	13.62	13.84	13.73

Table 1 shows the RMSE results for all the baseline methods. Firstly, we examine the effect of the $f0$ and voiced/unvoiced flag as a additional condition for WaveNet voice conversion. It is observed that WaveNet-VC consistently outperforms WaveNet-PPG in both intra- and inter-gender conversions.

Then, we further compare the performance of WaveNet-VC with other baseline methods. We observe that WaveNet-VC performs close to two WaveNet vocoder baselines, GMM-WaveNet and GMM(GV)-WaveNet, with averaged RMSEs over all testing pairs of 13.73 dB, 13.49 dB and 13.96 dB respectively. The systems using WORLD vocoder outperform those with WaveNet vocoder. GMM-WORLD achieves the lowest RMSE of 11.38 dB.

Objective metric evaluates the spectral distortion that reflects the how close the generated voice is to the target speech. However, it is an indirect measurement. Typically, speech generated by traditional vocoders give a lower objective measure than that of WaveNet [22, 24].

5.2. Subjective evaluation

AB preference tests and XAB tests were conducted to assess the speech quality and speaker similarity respectively. In AB preference tests, each paired samples A and B were randomly selected from the proposed method and one of the baseline methods, respectively. Each listener was asked to choose the sample with better quality. While, in XAB preference tests, X indicated the reference target sample, A and B were the converted samples randomly selected from the comparison methods. Noted that X, A and B have the same language content. The listeners were asked to listen to the samples, then decided A and B which is closer to the reference sample or no preference. For each test, 20 sample pairs were randomly selected from the 80 paired samples. 10 subjects participated in each tests. Only the WaveNet vocoder based VC baselines, GMM-WaveNet and GMM(GV)-WaveNet were included in the listening tests.

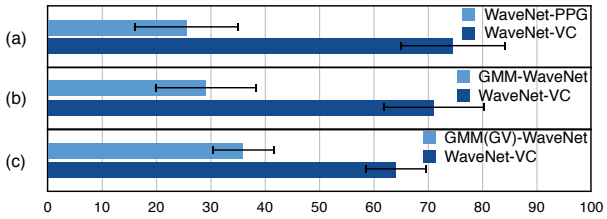

Figure 3: *Results of quality preference tests with 95% confidence intervals for different methods.*

Subjective results of quality preference tests are presented in Fig. 3. Results showed in Fig. 3 (a) suggests that speech quality of WaveNet-VC significantly outperforms that of WaveNet-PPG. Similar results were also observed in Fig. 3 (b) and Fig. 3 (b), which suggest that the proposed WaveNet-VC significantly outperforms WaveNet Vocoder baselines in terms of speech quality.

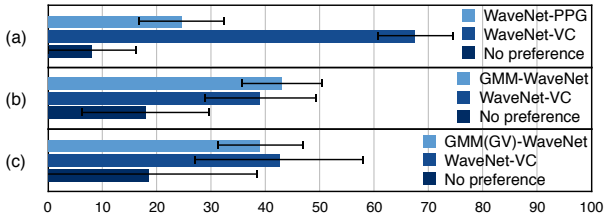

Figure 4: *Results of similarity preference tests with 95% confidence intervals for different methods.*

Subjective results of speaker identity are presented in Fig. 4. As shown in Fig. 4 (a), we observed that WaveNet-VC significantly outperforms WaveNet-PPG in terms of similarity. While, in the experiments of WaveNet-VC vs. GMM-WaveNet and WaveNet-VC vs. GMM(GV)-WaveNet (see Fig. 4 (b) and Fig. 4 (c)), the identification rates fall into each other's confidence intervals. This indicates that they are not significantly different in terms of speaker identity. (Converted samples are available via: https://xhtian.github.io/WaveNet-VC-Demo/)

6. Conclusions

This paper presents a voice conversion approach using WaveNet for non-parallel data. The proposed approach does not rely on the vocoder features for conversion, which 1) avoids the feature analysis and speech synthesis problems arise from vocoding; 2) reduces the feature mismatch problem in WaveNet vocoder based approaches. Experiment results show that the WaveNet-VC significantly outperforms the baseline methods in terms of quality, while maintain the speaker identity.

7. Acknowledgment

This research is supported by the NUS Start-up Grant, Non-parametric approach to voice morphing.

8. References

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

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 1998, pp. 285–288.

[3] 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, 2007.

[4] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *INTERSPEECH*, 2011, pp. 669–672.

[5] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[6] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[7] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 211–215.

[8] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. S. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Spoken Language Technology workshop (SLT)*, 2012, pp. 313–317.

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

[11] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 3893–3896.

[12] 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 Transactions on Speech and Audio Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[13] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (se) minimization training of neural network for voice conversion," in *INTERSPEECH*, 2014.

[14] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[15] 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," *Interspeech*, 2017.

[16] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[17] 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.

[18] 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*, pp. 1877–1884, 2016.

[19] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Interspeech*, 2014.

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

[21] 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." in *SSW*, 2016, p. 125.

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

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

[24] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *ICASSP*, 2018.

[25] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The nu non-parallel voice conversion system for the voice conversion challenge 2018," *Odyssey*, 2018.

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

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

[28] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda, "Collapsed speech segment detection and suppression for WaveNet vocoder," *Interspeech*, 2018.

[29] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016.

[30] X. Tian, J. Wang, H. Xu, E.-S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 227–232.

[31] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[32] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992.

[33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[34] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion." in *INTERSPEECH*, 2012.