Adaptive WaveNet Vocoder for Residual Compensation in GAN-Based Voice Conversion

Conference Paper · December 2018

DOI: 10.1109/SLT.2018.8639507

5 authors, including:

Berrak Sisman
Singapore University of Technology and Design
24 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Mingyang Zhang
National University of Singapore
12 PUBLICATIONS **96** CITATIONS
SEE PROFILE

Haizhou Li
National University of Singapore
760 PUBLICATIONS **10,165** CITATIONS
SEE PROFILE

Sakriani Sakti
Nara Institute of Science and Technology
221 PUBLICATIONS **1,261** CITATIONS
SEE PROFILE

ADAPTIVE WAVENET VOCODER FOR RESIDUAL COMPENSATION IN GAN-BASED VOICE CONVERSION

Berrak Sisman[1,2,3]*, Mingyang Zhang*[1]*, Sakriani Sakti*[2,3]*, Haizhou Li*[1]*, Satoshi Nakamura*[2,3]

[1]National University of Singapore, Singapore
[2]Nara Institute of Science and Technology, Japan
[3]RIKEN, Center for Advanced Intelligence Project AIP, Japan

berraksisman@u.nus.edu, mingyang.zhang@u.nus.edu, ssakti@is.naist.jp, haizhou.li@nus.edu.sg, s-nakamura@is.naist.jp

ABSTRACT

In this paper, we propose to use generative adversarial networks (GAN) together with a WaveNet vocoder to address the over-smoothing problem arising from the deep learning approaches to voice conversion, and to improve the vocoding quality over the traditional vocoders. As GAN aims to minimize the divergence between the natural and converted speech parameters, it effectively alleviates the over-smoothing problem in the converted speech. On the other hand, WaveNet vocoder allows us to leverage from the human speech of a large speaker population, thus improving the naturalness of the synthetic voice. Furthermore, for the first time, we study how to use WaveNet vocoder for residual compensation to improve the voice conversion performance. The experiments show that the proposed voice conversion framework consistently outperforms the baselines.

Index Terms— voice conversion, generative adversarial networks, adaptive Wavenet, residual compensation

1. INTRODUCTION

Voice conversion (VC) converts one speaker's voice to sound like that of another. With the advancement of the technology, voice conversion has enabled many applications such as personalized speech synthesis, spoofing attacks, and dubbing of movies. A voice conversion system typically consists of a module that converts speech parameters, also called features, and followed by a vocoder. In this paper, we would like to study deep learning approaches to improve both the speech parameter converter and the vocoder.

The early studies of speech parameter conversion were focused on spectrum mapping between source and target speakers [1,2]. The statistical parametric approaches, such as Gaussian mixture model [3], partial least square regression [4] and dynamic kernel partial least squares regression (DKPLS) [5] marked a success in spectrum conversion. As a solution to the limited training data problem, non-negative matrix factorization based voice conversion frameworks [6–11] have been proposed to address the over-smoothing problem in voice conversion.

Recently, deep learning approaches such as restricted Boltzmann machine (RBM) [12, 13], long short-term memory (LSTM) [14, 15], and autoencoder [16, 17] are shown to be effective in modeling the relationship between source and target features more accurately than conventional techniques. However, the resulting speech parameters from these models tend to be over-smoothed. One way to improve the speech quality is to reduce the difference between natural and the converted speech parameters. Global variance (GV) [3] and modulation spectrum (MS) [18], among others [19, 20], are typical examples to address the problem. However, how to improve the perceptual quality of converted voice remains a research problem.

To address the quality problem, GAN was proposed [17, 21, 22] as a solution. GAN has shown to be effective in many fields including speech enhancement [23, 24], speech synthesis [25,26]. It consists of two neural networks: a discriminator to distinguish natural and generated samples, and a generator to deceive the discriminator. In this paper, we use a training criterion for the spectrum mapping; which is the weighted sum of the conventional minimum generation error (MGE) training and an adversarial loss. The adversarial loss makes the discriminator recognize the generated speech parameters as natural. Since the objective of the GANs is to minimize the divergence (i.e., the distribution difference) between the natural and generated speech parameters, this method effectively alleviates the problem of over-smoothing the generated speech parameters.

The speech synthesized by traditional parametric vocoders lacks naturalness due to the over-simplified assumptions in signal processing. WaveNet vocoder [28], that directly estimates waveform samples from the input feature vectors, potentially addresses the problem. Speaker dependent and

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016, Non-parametric approach to voice morphing. Berrak Sisman is also funded by SINGA Scholarship under A*STAR Graduate Academy.

Speech Samples: https://sites.google.com/view/berraksisman/

independent WaveNet vocoders [27, 29] were studied and shown initial success. The WaveNet approach transforms the vocoder design into a data driven learning process. Through the learning, the network is expected to capture the dynamics of the complex mechanics of the human speech generation process. In practice, the voice conversion task is usually given a very limited amount of training data, that is not enough to train a speaker dependent WaveNet vocoder [29]. Therefore, a speaker independent WaveNet vocoder is usually used. Recently, a speaker adaptation for WaveNet vocoder [39] is shown to improve the performance under the sparse representation framework. In this paper, we take this adaptation idea one step further and propose to use WaveNet as a vocoder and also as a residual compensation module.

The main contributions of this paper are summarized as follows: 1) we propose a voice conversion framework that consists a GAN-based voice conversion and a WaveNet vocoder; and 2) we propose a novel WaveNet vocoder that also performs residual compensation. To our best knowledge, we are the first to study the use of adaptive WaveNet vocoder for residual compensation. For the first time, we also study the interaction between a WaveNet vocoder and a GAN-based voice conversion.

This paper is organized as follows: In Section 2, we explain the generative adversarial networks and their applications. In Section 3, we study the use of WaveNet vocoder in voice conversion. In section 4, we propose the novel idea of WaveNet-based residual compensation for GAN-based voice conversion, and formulate the training and run-time processes. We report the objective and subjective evaluation results in Section 5 and conclude in Section 6.

2. GENERATIVE ADVERSARIAL NETWORKS (GAN)

A generative adversarial network learns a deep network by simultaneously training two DNNs: a generator and discriminator $D(x;\theta_D)$, where θ_D is the model parameters for the discriminator. The posterior probability of an input x being a natural data, can be obtained by taking the sigmoid function from the discriminator's output, $1/(1+exp(-D(x)))$. The discriminator is trained to make the posterior probability 1 for natural data and 0 for generated data, while the generator is trained to deceive the discriminator.

GANs have recently been shown to be an effective training method and have been used for image generation [30], image synthesis [30], speech enhancement [23, 24], language identification [31], and text-to-speech synthesis [25, 26]. Moreover, Kaneko et al. [21] applied a GAN to sequence-to-sequence VC and demonstrated that the use of GAN-based training criteria outperforms the use of traditional mean squared error (MSE)-based training criteria. The subsequent GAN-based voice conversion techniques include CycleGAN [22] and starGAN [32], that achieve remarkable performance.

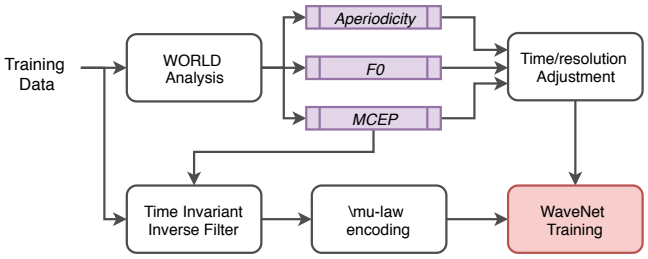


Fig. 1: The training phase of the Speaker Independent (SI) WaveNet vocoder [27]. The WaveNet vocoder learns the relationship between WORLD-analysized speech parameters and speech waveforms.

In this paper, we take the GAN-based voice conversion one step further by coupling with a WaveNet vocoder to improve the vocoding quality. The WaveNet vocoder also serves as a residual compensation module to improve the naturalness of converted speech.

3. WAVENET AS A VOCODER

Many of the state-of-the-art voice conversion frameworks [1–3,17,22,34–36] use a statistical parametric vocoder, which is generally designed to simulate the complex mechanics of the human speech generation process under certain simple assumptions, for example, the interaction between F0 and formant structure is ignored, the phase information is discarded [37], the assumption of stationary process in the short-time window, a time-invariant linear filter. As a result, the traditional vocoding voice lacks naturalness in general. Such a problem becomes more serious in voice conversion where the feature conversion changes both F0 and the formant structure of speech among others. We believe that a good vocoder can help to reconstruct the speech by harmonizing various changes. Motivated by this, recently in speech synthesis [38] and in GMM-based voice conversion [33], WaveNet vocoder is shown to be effective by improving the naturalness of the synthetic voice. Furthermore, in [39], an adaptation approach to the speaker independent WaveNet vocoder is shown to achieve a remarkable improvement over the speaker independent WaveNet vocoder for voice conversion.

3.1. WaveNet

WaveNet [28] is a well-known deep neural network that can generates raw audio waveforms. The joint probability of a waveform $x = x_1, x_2, ..., x_N$ is factorized as a product of conditional probabilities.

$$p(x) = \prod_{n=1}^{N} p(x_n | x_1, x_2, ..., x_{n-1}) \quad (1)$$

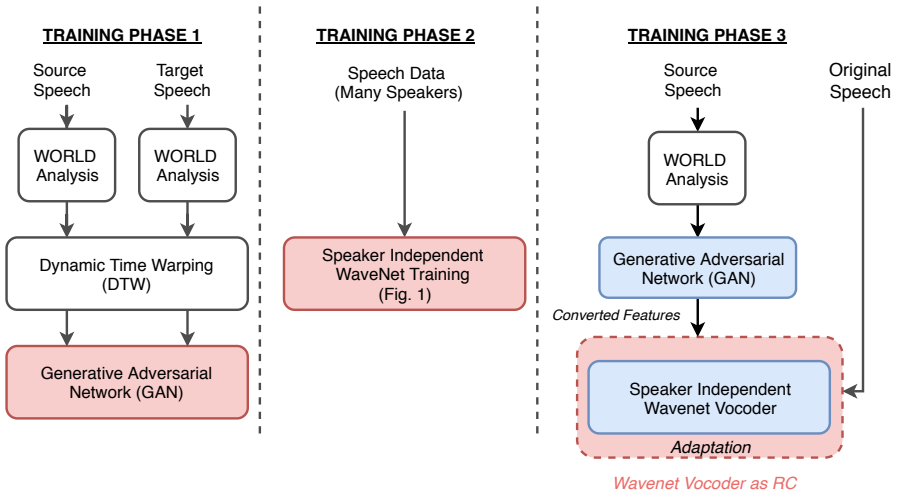


Fig. 2: The training phase of the proposed GAN-based voice conversion framework, where WaveNet acts as vocoder and a residual compensation module. Pink boxes represent the training stage of the networks, while blue boxes represent the networks that are already trained.

By canceling the effect of past samples, WaveNet approximates the conditional probability, that is given above. WaveNet consists of many residual blocks. Each of these residual blocks consist of 2×1 dilated causal convolutions, a gated activation function and 1×1 convolutions.

Using additional auxiliary features h, WaveNet can also model conditional distribution $p(x|h)$. Equation (1) can then be written as follows:

$$p(x|h) = \prod_{t=1}^{T} p(x_t|x_1, x_2, ..., x_{t-1}, h) \qquad (2)$$

By conditioning the network with auxiliary features, we can control the characteristics of generated samples. In original WaveNet [28], linguistic features and/or speaker codes are conditioned to generate speech samples based on given text information while keeping specific speaker characteristics.

3.2. WaveNet Vocoder

WaveNet vocoder [27, 33] has been proposed and shown to achieve remarkable sound quality improvement over the traditional vocoders. WaveNet vocoder is able to learn the relationship between input features and output waveforms, and also able to learn the interaction among the input features. Recently, a speaker independent WaveNet vocoder [27] is studied that takes the acoustic features such as F0, aperiodicity, and spectrum as the additional inputs to WaveNet. In doing so, WaveNet learns a sample-by-sample correspondence between the time-domain waveform and the corresponding acoustic features.

In this paper, we investigate the use of WaveNet vocoder in a GAN-based voice conversion. Different from all the previous studies, we propose to use WaveNet as a vocoder, as well as a residual compensation module for voice conversion.

4. WAVENET VOCODER FOR RESIDUAL COMPENSATION

The difference between the converted speech and the original speech represents the unwanted residual. We believe that the residual is a function that can be learnt and compensated. The study [7, 11] has shown that residual compensation always enhances the speech quality. We now study a novel technique for voice conversion, that uses WaveNet vocoder for residual compensation. We would like to mention that the proposed WaveNet vocoder can be easily generalized to work with any other voice conversion framework in the literature.

4.1. Training Phase

Figure 2 shows the training phase of the proposed voice conversion framework. In the proposed framework, we have 3 training phases as follows: 1) training a GAN-based voice conversion; 2) training the speaker independent WaveNet vocoder; and 3) adapting the WaveNet for residual compensation.

In training phase 1, we use parallel training data from source and target speakers. We first perform frame alignment by using dynamic time warping (DTW). Then, we train the GAN-based voice conversion framework to find a mapping between source and target spectral features. One can consider the source features as the input and the target features as the supervisor during training.

In training phase 2, we train the speaker independent

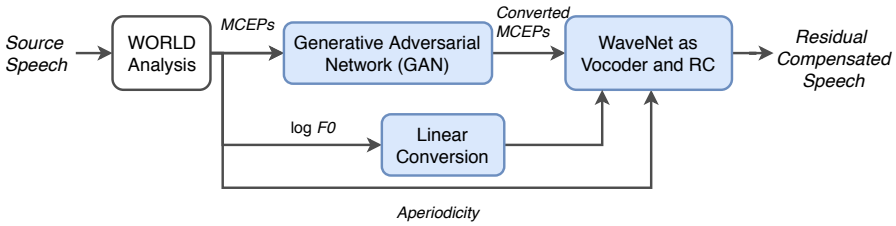


Fig. 3: The run-time conversion phase of the proposed GAN-based voice conversion with WaveNet vocoder.

WaveNet vocoder to generate the converted speech waveforms in a similar way that is described in [27] and shown in Fig. 1. WaveNet [28] is a fully convolutional neural network, where the convolutional layers have various dilation factors that allow its receptive field to grow exponentially with depth and cover thousands of timesteps. It is built with the acoustic features as the additional input. There are 30 hidden layers in our WaveNet Vocoder, and dilations are $[1, 2, 4, 8, 16, 32, 64, 128, 256, 512] \times 3$ respectively. The training data is about 5 hours of 5,408 utterances from CMU ARCTIC and Voice Conversion Challenge (VCC) 2016 datasets. It is important to mention that, when training the WaveNet vocoder, we exclude any source-target speakers, that would be used in voice conversion experiments.

In training phase 3, we first perform spectral parameter conversion over the training data by using the GAN, that is obtained in training phase 1. We use the speaker-independent WaveNet vocoder as the initialization for WaveNet adaptation. The adapted WaveNet vocoder has the same configuration as the speaker-independent WaveNet vocoder. During the adaptation, both the converted features and the original target speech are taken as the inputs. The difference between the converted speech and the original target speech represents the unwanted residual. At run-time, we only have the converted features, and we expect the adapted WaveNet vocoder to compensate the residuals for the target speaker. It is important to mention that for WaveNet adaptation, we use the same source-target utterances that have been used in the training of GAN.

4.2. Run-time Conversion Phase

The run-time conversion phase is depicted in Figure 3. We first use WORLD vocoder to obtain MCEPs, F0 and aperiodicity. Then, we use the trained GAN to perform spectral parameter conversion. Lastly, we use the adapted WaveNet vocoder to generate the converted speech waveforms. We believe that the adapted WaveNet vocoder is able to compensate the residual between the converted and original speech, hence achieve a natural sounding speech that is more similar to the target speaker. We note that F0 is converted linearly by normalizing the mean and variance of the source to that of the target, and we copy aperiodicity directly from source speech.

Framework	SNR	MCD
WaveNet Vocoder as RC	2.86	4.10
SA WaveNet Vocoder	2.46	4.34
SI WaveNet Vocoder	1.15	4.47
WORLD Vocoder	-3.21	4.07

Table 1: Comparison of spectral distortions between the proposed approach, denoted as WaveNet as RC, speaker adapted WaveNet, denoted as SA WaveNet, speaker independent WaveNet, denoted as SI WaveNet and the baseline WORLD. In all of the experiments, GAN is used as the spectrum conversion module.

Our framework differs from the previous studies of GAN-based voice conversion [22, 32] mainly in the WaveNet vocoder. For the first time, we study the interaction between GAN based voice conversion and WaveNet vocoder under different training scenarios, that includes speaker independent (SI) WaveNet vocoder, SI WaveNet vocoder adapted by the original target features that we call *Speaker Adapted (SA) WaveNet vocoder*, and SI WaveNet vocoder adapted by the converted features, that we call *WaveNet vocoder as RC*.

While the proposed WaveNet vocoder shares similar motivation with [39], it differs from [39] in many ways. For example, in [39], the adaptation of WaveNet vocoder is done with original target features. In this paper, WaveNet vocoder is adapted with the converted features and the original target speech, in a way that WaveNet vocoder learns the residual mapping between the converted and the original speech. For the first time in voice conversion, we propose to use WaveNet both as a vocoder and as a residual compensation module.

5. EXPERIMENTS

We conduct the experiments on CMU Arctic database to assess the performance of the proposed voice conversion framework. 50 source-target utterance pairs are used during the training of GAN. We train the speaker-independent WaveNet Vocoder with 5 hours of data from CMU Arctic and Voice Conversion Challenge (VCC) 2016 datasets. We set the batch size to 20,000 and the iteration number to 200,000. Then, we use this speaker-independent WaveNet Vocoder as the initial-

Framework	Best (%)	Worst (%)	Not Preferred (%)
WaveNet as RC	100	0	0
SA WaveNet	0	0	100
SI WaveNet	0	65	35
WORLD	0	35	65

Table 2: Voice quality and speaker similarity assesment using Best-Worst percentages on an aggregate level [41].

ized network for adaptation. During the adaptation, we use the same 50 utterances of speech from the source and target speakers, with a batch size is 20,000 and iteration number is 100,000.

5.1. Objective Evaluation

We implemented the following voice conversion schemes: 1) GAN-based voice conversion with WORLD vocoder; 2) GAN-based voice conversion with SI WaveNet vocoder; 3) GAN-based voice conversion with SA WaveNet vocoder; and 4) GAN-based voice conversion with WaveNet vocoder as residual compensation. As GAN-based voice conversion with WORLD vocoder is shown to outperform many of the state-of-the-art techniques, we use it as the baseline framework.

In order to conduct objective evaluation, we use the following two evaluation criteria Mel ceptral distortion (MCD) [40] and signal to noise ratio (SNR), which are calculated as follows:

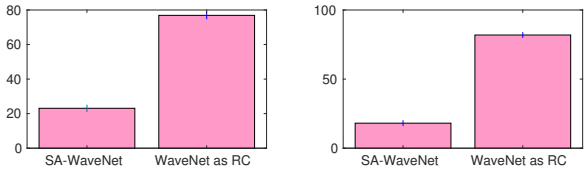
$$SNR[dB] = 10\log_{10}\left(\frac{\sum_{n=1}^{N} x(n)^2}{\sum_{n=1}^{N}(x(n)-y(n))^2}\right), \quad (3)$$

$$MCD[dB] = \frac{10}{\log 10}\sqrt{2\sum_{m=1}^{M}\left(c_t(m)-c_s(m)\right)^2} \quad (4)$$

where $x(n)$ is the converted waveform, $y(n)$ is the target waveform at time n, $c_s(m)$ represents the converted Mel-cepstrum, $c_t(m)$ represents the target Mel-cepstrum at frame m and M is the order of mel-cepstrum.

Objective evaluation results are shown in Table 1. Firstly, we focus on the SNR results. We can see that *WaveNet as RC* outperforms all other frameworks. We can also see that the speaker adapted WaveNet vocoder achieves better performance than the speaker independent WaveNet vocoder and WORLD vocoder.

Secondly, we would like to evaluate the on feature domain distortion and MCD. Traditionally, WORLD vocoder and WaveNet vocoder takes the same acoustic features as input and the MCD actually compares the acoustic features before the vocoder. Such MCD doesn't evaluate the effect of vocoders. However, in this paper, we would like to compare several different vocoding effects in conjunction with the GAN voice conversion module. To this end, we follow a procedure implemented in [27]. We first synthesize the converted speech, then calculate Mel-cepstral features one more time. We note that MCD is not a perfect measure to show the effect of WaveNet-based residual compensation. As previously shown [27], WaveNet vocoder doesn't offer a lower MCD than WORLD vocoder. However, we note that the proposed *WaveNet as RC* approach achieves as competitive as WORLD vocoder in terms of MCD, and outperforms other variants of WaveNet vocoders.



(a) Voice quality (b) Speaker similarity

Fig. 4: The preference test with 95 % confidence interval between the speaker adapted WaveNet vocoder and the proposed WaveNet vocoder as residual compensation, in terms of voice quality.

5.2. Subjective Evaluation

We further conduct four listening experiments to assess the performance of WaveNet vocoder in GAN-based voice conversion, in terms of voice quality and speaker similarity. 15 subjects participated in all the listening tests. Each listener listens to 20 converted utterances from 2 target speakers.

We conduct the first two listening experiments, as reported in Figure 4.a and 4.b, to examine the effect of residual compensation with WaveNet vocoder in terms of voice quality and speaker similarity. We note that the speaker adapted WaveNet vocoder is trained in a similar way as that in [39]. Each listener is asked to decide the better sample in terms of voice quality and speaker similarity. We observe that the proposed approach that uses the WaveNet vocoder for residual compensation (WaveNet as RC), outperforms the speaker adapted WaveNet (SA WaveNet) vocoder in both voice quality and speaker similarity.

Then, we evaluate the sound quality of the converted voices by using the mean opinion score (MOS). The listeners rate the quality of the converted voice using a 5-point scale: 5

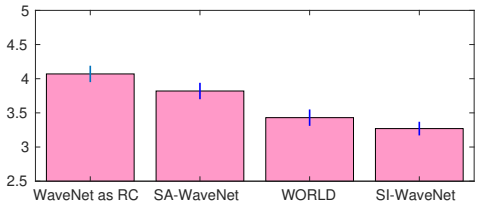


Fig. 5: Comparison of evaluated MOS for GAN with WaveNet as RC, GAN with SA WaveNet, GAN with SI WaveNet and GAN with WORLD vocoder.

for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Figure 5 suggest that the proposed WaveNet as RC framework, significantly outperforms other competing frameworks.

In VC literature, AB preference test and XAB test are the evaluation techniques which have been widely used [7]. We know that people are good at picking the extremes but their preferences for anything in between might be fuzzy and inaccurate. To alleviate such a problem, in this paper we also use an evaluation technique called Best-Worst Scaling (BWS) [41, 42], which can handle a long list of options and always generates discriminating results as the respondents are asked to choose the BEST and WORST option in a choice set. In Table 2, each listener was asked to listen to the converted samples and then decide the best and worst sample in terms of voice quality. The result shows that the proposed approach that uses WaveNet vocoder as residual compensation module, was chosen as the best-converted sample among all the speech samples. We also note that SI WaveNet was chosen as the worst-converted sample 65 % among all the speech samples, while WORLD vocoder chosen as the worst-converted sample 35 % among all the speech samples.

Overall, the results that reported in the listening experiments support the proposed idea of using WaveNet vocoder as a vocoder, as well as a residual compensation module.

6. CONCLUSION

We have studied the deep learning approaches to improve both the speech parameter converter and the vocoder. We propose to use WaveNet as a vocoder and also as a residual compensation module for voice conversion. We perform adaptation on WaveNet vocoder with a very limited amount of data from only the target speaker, that serves as a solution to the limited data problem. Experiment results show that the proposed idea of WaveNet vocoder and residual compensator in GAN-based voice conversion makes good use of the limited training data and outperforms the baselines in both objective and subjective evaluations.

The proposed WaveNet-based residual compensation module can also work together with a parallel-data-free voice conversion framework which will be an interesting future work.

7. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 655–658, 1988.

[2] Kiyohiro Shikano, Satoshi Nakamura, and Masanobu Abe, "Speaker Adaptation and Voice Conversion by Codebook Mapping," *IEEE International Sympoisum on Circuits and Systems*, pp. 594–597, 1991.

[3] Tomoki Toda, Alan W. Black, and Keiichi 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] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[5] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[6] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," *In IEEE SLT*, pp. 313–317, 2012.

[7] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou 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, 2014.

[8] Ryo Aihara, Kenta Masaka, Tetsuya Takiguchi, and Yasuo Ariki, "Parallel dictionary learning for multimodal voice conversion using matrix factorization," *In INTERSPEECH*, pp. 27–40, 2016.

[9] Zeyu Jin, Adam Finkelstein, Stephen Di Verdi, Jingwan Lu, and Gautham J Mysore, "Cute: a concatenative method for voice conversion using exemplar- based unit selection," *In ICASSP*, 2016.

[10] Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," *In ICASSP*, 2014.

[11] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," *IEEE ASRU*, 2017.

[12] Ling-hui Chen, Zhen-hua Ling, Li-juan Liu, and Li-rong Dai, "Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[13] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion," *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, , no. September, pp. 2278–2282, 2014.

[14] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional Long Short-Term Memory based Recurrent Neural Networks," *In ICASSP*, , no. 1, pp. 4869–4873, 2015.

[15] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *In IEEE ICME*, 2016.

[16] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC*, 2016.

[17] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," *arXiv*, 2017.

[18] S Takamichi, T. Toda, W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, 2016.

[19] S Takamichi, T. Toda, W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," *IEEE ICASSP*, 2015.

[20] T. Nose and A. Ito, "Analysis of spectral enhancement using global variance in HMM-based speech synthesis," *INTERSPEECH*, 2014.

[21] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *INTERSPEECH*, 2017.

[22] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv*, 2017.

[23] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *INTERSPEECH*, 2017.

[24] Daniel Michelsanti and Zheng-Hua Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *INTERSPEECH*, 2017.

[25] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," *ICASSP*, 2017.

[26] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

[27] Tomoki Hayashi, Akira Tamamori, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "An investigation of multi-speaker training for wavenet vocoder," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 712–718, 2017.

[28] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[29] Akira Tamamori, Tomoki Hayashi, and Kazuhiro Kobayashi, "Speaker-dependent wavenet vocoder," *INTERSPEECH*, 2017.

[30] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *ICCV*, 2017.

[31] Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai, "Conditional generative adversarial nets classifier for spoken language identification," *INTERSPEECH*, 2017.

[32] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv*, 2018.

[33] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, "Statistical voice conversion with wavenet-based waveform generation," *INTERSPEECH*, 2017.

[34] Heiga Zen, Yoshihiko Nankaku, and Keiichi Tokuda, "Probabilistic feature mapping based on trajectory HMMs," *In INTERSPEECH*, pp. 1068–1071, 2008.

[35] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data," *arXiv*, 2017.

[36] Wei-Ning Hsu, Yu Zhang, and James Glass, "Learning Latent Representations for Speech Generation and Transformation," *arXiv*, 2017.

[37] Sadaoki Furui, "Digital speech processing, synthesis, and recognition(revised and expanded)," *Digital Speech Processing, Synthesis, and Recognition*, 2000.

[38] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv:1712.05884*, 2018.

[39] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," *INTERSPEECH*, 2018.

[40] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Communications, Computers and Signal Processing*, pp. 125–128, 1993.

[41] Terry N. Flynn and Anthony A. J. Marley, "Best worst scaling: Theory and methods," *Handbook of choice modelling*, pp. 178–201, 2014.

[42] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Transformation of Prosody in voice conversion," *APSIPA ASC*, 2017.