Interspeech 2018
2-6 September 2018, Hyderabad

High-quality Voice Conversion Using Spectrogram-Based WaveNet Vocoder

Kuan Chen, Bo Chen, Jiahao Lai, Kai Yu

Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

{azraelkuan, bobmilk, ljhao1993, kai.yu}@sjtu.edu.cn

Abstract

Waveform generator is a key component in voice conversion. Recently, WaveNet waveform generator conditioned on the Mel-cepstrum (Mcep) has shown better quality over standard vocoder. In this paper, an enhanced WaveNet model based on spectrogram is proposed to further improve voice conversion performance. Here, Mel-frequency spectrogram is converted from source speaker to target speaker using an LSTM-RNN based frame-to-frame feature mapping. To evaluate the performance, the proposed approach is compared to an Mcep based LSTM-RNN voice conversion system. Both STRAIGHT vocoder and Mcep-based WaveNet vocoder are elected to produce the converted speech for Mcep conversion system. The fundamental frequency (F_0) of the converted speech in different systems is analyzed. The naturalness, similarity and intelligibility are evaluated in subjective measures. Results show that the spectrogram based WaveNet waveform generator can achieve better voice conversion quality compared to traditional WaveNet approaches. The Mel-spectrogram based voice conversion can achieve significant improvement in speaker similarity and inherent F_0 conversion.

Index Terms: voice conversion, WaveNet vocoder, mel-frequency spectrogram, LSTM-RNN

1. Introduction

Voice Conversion(VC) is a technique to modify the speech of the source speaker to sound like the target speaker while preserving the linguistic content[1]. Conventional voice conversion techniques focus on developing conversion functions using some parallel data which the source speaker and target speaker speak the same sentences. Some conversion models like Gaussian mixture model(GMM)[2], deep neural networks[3, 4] have been applied to convert the acoustic features from the source speaker to the corresponding target speaker.

The sound quality of the converted speech is always attractive to researchers. There are always distortions in the converted speech, e.g. over-smoothing, lack of similarity and etc. In parametric voice conversion, several techniques have been proposed to enhance the sound quality, e.g. modeling additional features(Global Variance[5], Spectrum envelope[6]) and post-filtering[7]. However, the quality of the converted speech is still not as natural as the target speaker. One important factor is that the acoustic features used for parametric voice conversion are usually vocoder parameters (e.g. Mel-cepstrum, F_0) whose conversion can lead to quality distortion when generating waveform with the converted vocoder parameters.

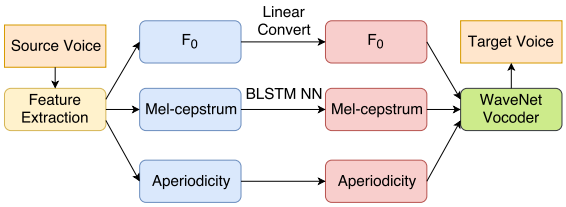
Recently, a high-quality vocoder [8] has been proposed with WaveNet speech generation model. WaveNet [9] is the state-of-art natural waveform generation technique that can produce high quality speech waveform. One of its advantages is that the WaveNet speech generation model is able to generate waveform on specific conditions like linguistic information or acoustic features. It has been applied to many applications like text-to-speech [9, 10, 11], voice conversion[6] and speech vocoder[8]. The WaveNet waveform generation in Voice Conversion has been proposed in [6]. Similar to the WaveNet vocoder[8], the acoustic features in [6] are mainly the Mel-cepstrum(Mcep) and fundamental frequency(F_0) which are widely used for speech synthesis. The sound quality of the WaveNet-vocoded converted voice is comparable to the STRAIGHT-vocoded[12] voice. Very recently, Tacotron 2[10] has been proposed as a sequence-to-sequence model with attention in end-to-end speech synthesis. Comparing to Tacotron 1[13], one of its advantages is that the speech signals are generated with WaveNet architecture conditioning on Mel-frequency spectrogram. It draws our interest that: does Mel-frequency spectrogram work better in other speech generation tasks? We will give an investigation on introducing the Mel-frequency spectrogram into voice conversion tasks.

In this paper, we propose a high quality voice conversion architecture with mel-frequency spectrogram as acoustic features. The converted features are then vocoded into waveform using a Mel-spectrogram based WaveNet vocoder. A Mcep-based voice conversion system we proposed before [14] (Group 'G' in VCC2016[15]) is used for comparison. The Mel-spectrogram and Mcep in different systems are trained using similar LSTM-RNN neural networks for frame-to-frame feature mapping. The converted Mcep and F_0s are vocoded to waveform using STRAIGHT vocoder and an Mcep-based WaveNet vocoder. The F_0 contours of converted waveform, which is an important factor of the speech quality, are analyzed in detail for different systems. The naturalness, similarity and intelligibility are subjectively evaluated by human listeners. The result shows that, voice conversion with Mel-frequency spectrogram can produce high quality converted voice especially in similarity.

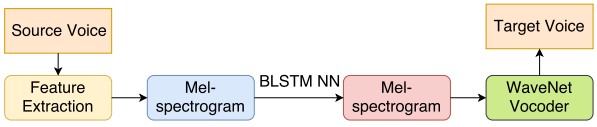
The rest of this paper is organized as follows: Section 2 gives an introduction of the parallel data voice conversion and introduces the LSTM-RNN acoustic feature conversion architecture. Section 3 proposes the Mel-spectrogram based voice conversion technique with Mel-spectrogram WaveNet vocoder. Section 4 describes the experiments with measurements. Section 5 gives the conclusion and the future work.

Bo Chen is the co-first author. The corresponding author is Kai Yu. This work has been supported by the National Key Research and Development Program of China under Grant No.2017YFB1002102, and the Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) (No.17JC1404104). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

10.21437/Interspeech.2018-1528



(a) *Mel-cepstral based Voice Conversion*



(b) *Mel-spectrogram based Voice Conversion*

Figure 1: *Architectures of voice conversion systems with WaveNet vocoder*

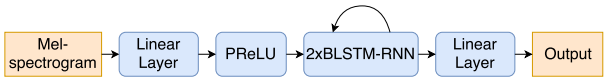


Figure 2: *BLSTM frame to frame voice conversion.*

2. Parallel Data Voice Conversion

This section gives an introduction of the parallel data voice conversion framework. Fig.1-(a) shows the architecture of a Mcep-based parallel data voice conversion system. The acoustic features of the source speaker are converted to the target speaker in different feature streams. The converted features are then vocoded into audio signals. This architecture is also a general parametric voice conversion framework[16] in which the general treatments are replaced by specific methods (e.g. BLSTM-NN, WaveNet Vocoder).

For a speech pair with the same text, the acoustic features $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_1, .., \hat{\boldsymbol{x}}_m$ from the source speaker and the corresponding acoustic features $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}_1, .., \hat{\boldsymbol{y}}_n$ from the target speaker are first aligned into the same length T. The alignment is usually applied directly by Dynamic Time Wrapping (DTW)[17]. Also, there are techniques to get a more accurate feature alignment with the help of automatic speech recognition (ASR) techniques [18, 14, 19]. The aligned feature sequences $\boldsymbol{x} = x_1, .., x_T$ and $\boldsymbol{y} = y_1, .., y_T$ are then converted frame by frame in different methods (e.g. GMM, LSTM). In this paper, the Mcep is converted using a BLSTM-NN architecture shown in Fig.2. The training cost is simply measured by the mean square error as shown in Eq.1 where M_{xy} is the Mcep converting model from source speaker to target speaker. The F_0 is converted linearly and the aperiodicity is not converted in this work.

$$\mathcal{L} = \sum_{i=1}^{T} |M_{xy}(\boldsymbol{x}_i) - \boldsymbol{y}_i|_2 \quad (1)$$

We observed that the intelligibility of the converted speech may degrade with WaveNet Vocoder. We tried to improve the intelligibility using un-parallel voice conversion techniques. A simple dual training strategy is applied to train M_{xy} and M_{yx} together as in [20]. Unfortunately, we only observed minor improvements in preliminary test. We plan to fully import CycleGAN [20] to improve the intelligibility in future work.

Table 1: *Fundamental frequency(RMSE)*

System	bdl-rms	clb-rms	bdl-slt	clb-slt
MSP-WaveNet	10.18	10.28	9.15	9.1
Mcep-WaveNet	11.22	10.85	11.76	11.06

3. Voice conversion with Mel-spectrogram

3.1. Mel-spectrogram conversion

Mel-spectrogram is a very low level acoustic presentation of the speech waveform. It has not yet been imported as acoustic features in voice conversion tasks, since there is not a good Vocoder for Mel-spectrogram before.

We propose a very simple architecture[1] to convert the speech waveform with Mel-spectrogram as shown in Fig.1(b). The speech waveform is only analyzed into Mel-spectrogram. Then the Mel-sepctogram is converted frame-by-frame following the architecture in Fig.2. Compared to the conventional Mcep-based voice conversion, F_0 is not necessary to be converted explicitly as a separate feature stream. It has been addressed in [15] that F_0 and duration patterns may be parameterized to properly handle their supra-segmental characteristics, which are not well converted within the frame-wise conversion process. However, in the proposed system, F_0 is converted inherently while converting the Mel-spectrograms. The performance of the F_0 conversion will be analyzed in detail in the experiments.

3.2. WaveNet vocoder

The conventional vocoder of voice conversion makes various assumptions which usually cause the sound quality degradation of the converted voice. Therefore, Wavenet Vocoder mainly based on Mel-cepstrum and F_0 was proposed[6] to overcome this problem. The result shows that the Speaker-Dependent Wavenet Vocoder[8] can generate better waveform than MLSA[21].

The Mel-spectrogram based WaveNet follows the architecture in Tacotron 2[13], which can produce high quality speech waveform in end-to-end text-to-speech task. The architecture of conditional WaveNet is shown in Fig.3[8]. It consists of a stack of dilated causal convolution layers, each can process the input vector in parallel. Two transposed convolution layers are added for upsampling. Also, the gated activation functions are used in WaveNet with the mechanism to condition extra information such as acoustic or linguistic features:

$$z = \tanh(W_f * i + V_f * c) \odot \sigma(W_g * i + V_g * c) \quad (2)$$

where $*$ denotes a convolution operator, and $\odot$ denotes an element-wise multiplication operator. $\sigma(\cdot)$ denotes a sigmoid function. i is the input vector and c is the extra condition feature like Mel-spectrogram and one hot of speaker identity. f and g represent filter and gate,respectively. W and V are learnable weights. Instead of using 8-bit(μ-law)[22], the signal samples

[1]The method to convert Mel-spectrogram can be investigated in future works. In this paper, we want to address that the simplest way can also achieve good performance.

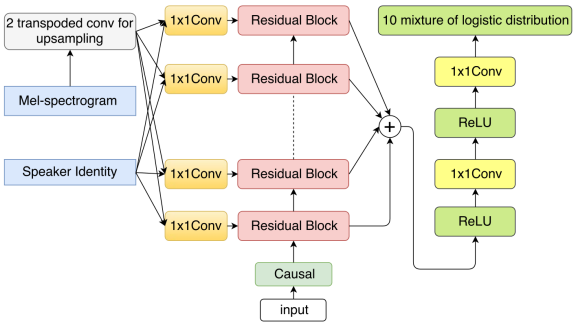


Figure 3: *Architecture of Conditional WaveNet Vocoder*

Table 2: *Comparison of voiced/unvoiced decision error(%)*

System	bdl-rms	clb-rms	bdl-slt	clb-slt
Msp-WaveNet	3.38	3.1	2.63	4.01
Mcep-WaveNet	3.46	3.21	2.71	3.63

are modelled with the discretized mixture of logistics distribution introduced in [23, 24].

4. Experiments and Results

4.1. Experiment setup

The experiments were conducted on CMU ARCTIC dataset[25] using PyTorch[26]. The sentences in the dataset are randomly divided into train, develop and test set, each with 957, 107, 55 sentences. The waveform is sampled at 16kHz sampling rate. The Mel-spectrograms are extracted through a short-time Fourier transform (STFT) using a 50ms frame size, 12.5 ms frame hop and a Hann window function as in [10]. The baseline system uses the same LSTM-RNN voice conversion system in [14]. The converted acoustic features are vocoded into speech waveform using both MLSA and Mcep-based WaveNet vocoder[8]. The Mcep-based WaveNet Vocoder proposed in [6] follows the best vocoder trained on natural acoustic features. The Mceps are extracted at 5ms frame shift. But different from [6], we use the conversion model in [14] and trained a speaker dependent WaveNet Vocoder using 8 bits μ-law.

In the system proposed in this paper, we first trained a speaker independent WaveNet vocoder on all waveforms in the CMU ARCTIC dataset except the utterance in the test set. The WaveNet network was trained for 1000k steps with Adam optimizer with a mini batch of 16 on 4 GTX1080TI, and it has 24 layers , which is divided into 4 groups. The hidden units of residual connection and gating layers are 512, the skip connection of the output layer is 256. we also use 10 mixture components for the mixture of logistics output distribution[24]. Then we trained a converting model based on LSTM network, which has two layers and the hidden units is 256. Before the lstm layer, we use two dense layer with PReLU[27] activation. And we apply a global mean-variance transformation for source and target speaker. To ensure that both WaveNet vocoders were well-trained. The training procedure is stopped after the WaveNet vocoder can generate convincing speech on the training set.

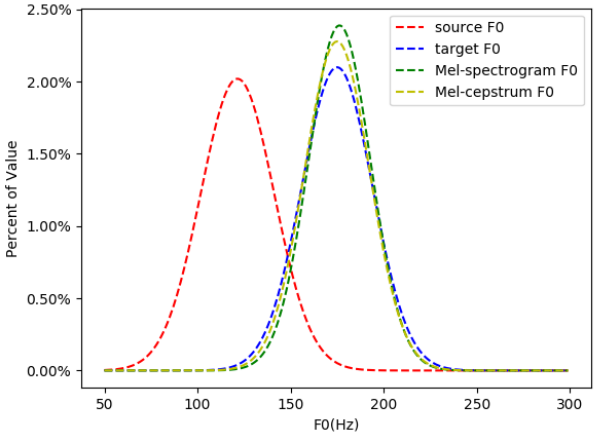


Figure 4: *The Distribution of F_0 in converted speech.*

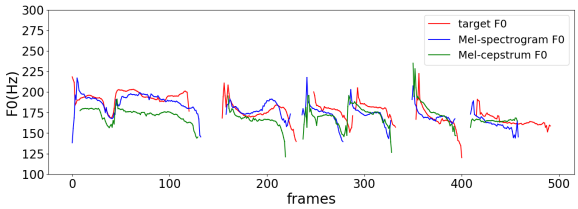


Figure 5: *The F_0 contour extracted from the converted speech*

4.2. Objective Measure

F_0 is an important acoustic features that affect the speech quality a lot. In the Mel-spectrogram based voice conversion, all the acoustic information is maintained in the lower level spectrogram representation. Therefore, F_0 is converted inherently during the Mel-spectrogram conversion. We first present an evaluation of the F_0 contour of the converted speech.

The F_0 contours are extracted from both natural and converted speech using WORLD[28]. Fig.5 shows an example of F_0 contours[2]. We can see that the F_0 contour from the Mel-spectrogram converted voice is closer to the target speech, even F_0 is not explicitly converted. We draw a distribution of F_0 in Fig.4, the system we proposed and the system based on Mel-cepstrum all have a close mean and standard deviation with the target speech. Exactly, the F_0 in the system based on Mel-cepstrum is converted by a global mean-variance transformation between source utterances and target utterances. So it is confirmed that the system proposed in this paper can obtain better F_0 without any prior condition.

Table 1 indicates the objective measure of F_0 error. Before we evaluate, DTW is applied to align the natural target utterance and the converted utterance. The system we proposed has higher accuracy than the system based on Mel-cepstrum. And table 2 lists unvoiced/voiced (U/V) decision errors.It is believed that the proposed system could capture the U/V information with comparable accuracy with the system based on Mel-cepstrum.

[2]The sentence is b0185. The audio is converted from `bdl` to `slt`. Since `bdl` and `slt` have similar speaking rates, we can directly look into their F_0 contours in parallel.

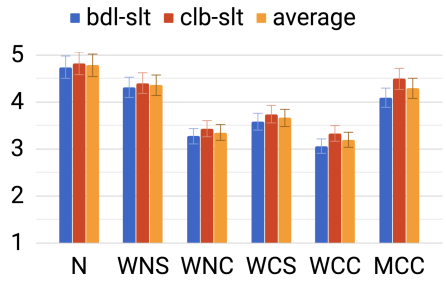


Figure 6: *MOS on intelligibility of the converted speech.*

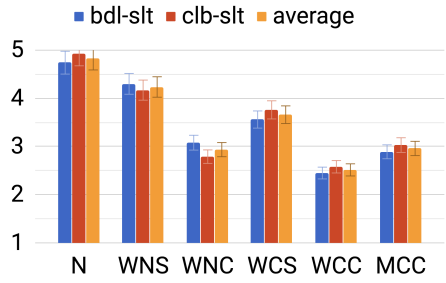


Figure 7: *MOS on naturalness of the converted speech.*

4.3. Subjective Measure

All the subjective tests are conducted in both intra-gender and cross-gender cases. In the listening test, we use (clb→slt) as the intra-gender pair and (bdl→slt) as the cross-gender pair. All 55 sentences in the test set are used for listening tests[3]. Every sentence is presented to at least 6 listeners in each test. The listeners are all non-native speakers.

4.3.1. Naturalness

We run a Mean Opinion Score (MOS) evaluation on the speech naturalness. The Mel-spectrogram is shorten as `Msp`. The evaluated experiment sets are listed below:

- Natural speech (N)
- WaveNet-vocoded speech on natural `Msp` (WNS[4])
- WaveNet-vocoded speech on natural `Mcep` (WNC)
- WaveNet-vocoded speech on converted `Msp` (WCS)
- WaveNet-vocoded speech on converted `Mcep` (WCC)
- MLSA-vocoded speech on converted `Mcep` (MCC)

4.3.2. Intelligibility

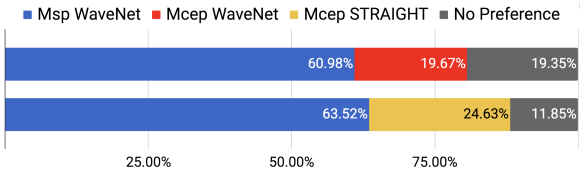
We observed that the contextual information may be distorted with WaveNet vocoder (both Msp and Mcep). So we also run a MOS evaluation on the intelligibility of the converted speech.

4.3.3. Similarity

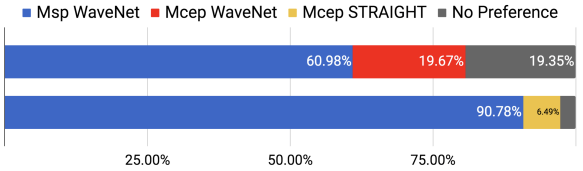
We run an preference test to evaluate the similarity. The converted speech from the two systems are provided to the listeners in random orders along with the natural speech of the same sentence from the target speaker. The listeners were asked to select which sentence sounds more like the target speaker.

[3]samples:https://azraelkuan.github.io/High-quality-Voice-Conversion-Using-Spectrogram-Based-WaveNet-Vocoder/

[4]The first char refers the vocoder type (WaveNet/MLSA); the second char refers to the acoustic features (Natural/Converted); the third char refers to the acoustic feature type (Mel-Spectrogram/Mel-Cepstrum).



(a) *bdl to slt*



(b) *clb to slt*

Figure 8: *Preference test on similarity.*

4.4. Experiment Results

Fig.7 shows the result of naturalness of the converted speech. We can see that WNS performs better than WNC, which means the Mel-spectrogram conversion has higher upper bound in speech naturalness, which can be further investigated. In addition to this, WCS achieves much better performance compared to WCC and MCC, which indicates that the Mel-spectrogram based voice conversion can achieve good naturalness.

Fig.6 shows the result of intelligibility of the converted speech. MCC achieves better performance than WCS and WCC. One of the reasons is that MCC can generate converted voice with steady quality in all the frames, the other one is that WaveNet Vocoder will generate buzzy voice sometimes, which can be considered as the lack of training data for WaveNet Vocoder. This might also indicate the reason why Mcep-based WaveNet vocoder has a similar speech quality MOS compared to MLSA in [15] even with a much higher naturalness.

Apart from this, we can also see that the WNS performs much better than WNC, which means the Mel-spectrogram contains more information than Mel-cepstrum.

Fig.8 shows the results of similarity of different systems compared to the target speaker. It shows that Msp Wavenet performs significantly better than Mcep WaveNet and Mcep STRAIGHT on intra-gender and cross-gender case.

5. Conclusion and Future Work

This paper presents a voice conversion technique to generate high quality voice from source speaker to target speaker with LSTM network and Mel-spectrogram based WaveNet Vocoder. Instead of using a conventional feature of STRAIGHT, we apply Mel-spectrogram in the pipelines of the proposed system. The experiment shows that Mel-spectrogram based WaveNet Vocoder performs much better than Mel-cepstrum based WaveNet Vocoder in voice conversion task in naturalness, similarity and intelligibility. In future work, we plan to build a transform learning technique to enable WaveNet Vocoder to generate better steady voice in small dataset, and further investigate the modelling algorithm on Mel-spectrogram.

6. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] 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.

[3] 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 (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.

[4] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4869–4873.

[5] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

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

[7] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 290–294.

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

[9] 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.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[11] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," *arXiv preprint arXiv:1803.00860*, 2018.

[12] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[14] J. Lai, B. Chen, T. Tan, S. Tong, and K. Yu, "Phone-aware lstm-rnn for voice conversion," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 2016, pp. 177–182.

[15] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *INTERSPEECH*, 2016, pp. 1632–1636.

[16] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The nu-naist voice conversion system for the voice conversion challenge 2016." in *INTERSPEECH*, 2016, pp. 1667–1671.

[17] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[18] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, "High quality voice conversion using prosodic and high-resolution spectral features," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5265–5285, 2016.

[19] B. Çişman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 677–684.

[20] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[21] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[22] G. Recommendation, "Pulse code modulation (pcm) of voice frequencies," *ITU*, 1988.

[23] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[24] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

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

[26] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration," 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[28] 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, no. 7, pp. 1877–1884, 2016.