STATISTICAL VOICE CONVERSION BASED ON WAVENET

Jumpei Niwa, Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Depatment of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

This paper proposes a voice conversion technique based on WaveNet to directly generate target audio waveforms from acoustic features of a source speaker. In voice conversion based on statistical models, the relation between acoustic features, such as spectral parameters, extracted from source and target audio waveforms is generally modeled using statistical models, such as Gaussian mixture models and neural networks. Although modeling the relation between acoustic features is reasonable and efficient, these models are not optimized for predicting target audio waveforms because the vocoder parameters are used as intermediate representations. To overcome this problem, we developed a voice conversion method to model the relation between target audio waveforms and acoustic features extracted from source audio waveforms using WaveNet, which is a generative model for audio waveforms. The proposed model can directly generate converted audio waveforms without vocoders. Experimental results indicate that the proposed method can generate a more naturally sounding converted speech than that using a conventional DNN method.

Index Terms— Voice conversion, WaveNet, Deep Neural Network, statistical model

1. INTRODUCTION

Voice conversion is a technique for converting a certain speaker's voice into another voice while maintaining linguistic information. The technique has been applied to many tasks, such as speech enhancement, emotion conversion, speaking assistance, post-processing of text-to-speech (TTS), and other applications [1, 2].

In voice conversion studies, statistical approaches have been widely used for mapping acoustic features of a source speaker to those of a target one. The framework of a statistical voice conversion typically uses a parallel data set, which consists of pairs of speech data from source and target speakers uttering the same sentences, to estimate the voice conversion model. Conventional statistical voice conversion framework is often based on a Gaussian mixture model (GMM) [3, 4]. This method achieves continuous mapping on the basis of soft clustering and converts spectral parameters frame-by-frame on the basis of the minimum mean square error. A more recent framework that has been widely investigated is based on deep neural networks (DNNs) [5–8]. DNN-based voice conversion can represent complex mapping functions from acoustic features for source speech to ones for target speech. It has been often reported that such a new approach performs better than a conventional ones, such as a GMM, in voice conversion. The constructed model can be used to convert the identity of the source speaker's arbitrary utterances to that of the target speaker. However, the naturalness and similarity of the converted voices are still degraded compared to the natural voices. One of the major factors causing this degradation is the use of a vocoder.

Recently, a deep neural network called WaveNet [9] has been proposed as a generative model that operates directly on audio waveforms. WaveNet can model audio waveforms accurately, therefore, it can directly generate natural-sounding speech without vocoders. In [9], WaveNet was applied to text-to-speech (TTS) by using linguistic features as additional inputs and achieves improvements from the state-of-the-art DNN-based method. Additionally, a speaker-dependent WaveNet vocoder has been proposed. In this method, WaveNet is used as a waveform generator like the vocoder by utilizing acoustic features for existing vocoders as additional inputs for WaveNet. It has been demonstrated that the sound quality of the WaveNet vocoder was significantly improved compared to the mel-cepstrum vocoder, and could capture source excitation information more accurately. Also, a technique based on a WaveNet vocoder for voice conversion was proposed [10]. In this study, the acoustic features of the source speaker are converted into those of the target speaker on the basis of GMMs. The converted acoustic features are then passed through the WaveNet vocoder, and the converted speech is generated from the WaveNet vocoder, as shown in Fig. 1(b). However, in this study, the GMM-based conversion model and the WaveNet vocoder are modeled independently. Therefore, the models are not optimized for generating converted audio waveforms. In this paper, we apply WaveNet to voice conversion as shown in Fig. 1(c). The proposed method can model audio waveforms of the target speaker by using the acoustic features of the source speaker as additional inputs for WaveNet. Consequently, the proposed model is optimized for predicting the target audio waveforms from acoustic features of the source speaker, i.e., it can directly generate audio waveforms with the target speaker's voice characteristics from acoustic features of the source speaker.

The rest of this paper is organized as follows. Sections 2 and 3 describe WaveNet and voice conversion based on WaveNet, respectively. The experimental conditions and experimental results are given in Section 4. Concluding remarks and future work are presented in Section 5.

2. WAVENET

WaveNet is a generative model for audio waveforms. The input to the network is a sequence of waveform samples. The joint probability of an audio sample sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$ is factorized as a product of conditional probabilities as follows:

$$P(\boldsymbol{x}) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1}). \tag{1}$$

Each audio sample x_t is conditioned on the samples from all the previous time steps. This conditional probability distribution is modeled by dilated causal convolutions with gated activation units. The form of the gated activation units of WaveNet is defined as follows:

$$\boldsymbol{z} = \tanh(\boldsymbol{W_f} * \boldsymbol{x}) \odot \sigma(\boldsymbol{W_g} * \boldsymbol{x}), \tag{2}$$

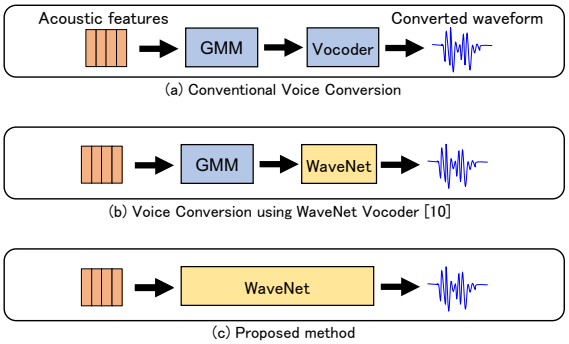

Fig. 1. Difference between conventional voice conversion, voice conversion using WaveNet vocoder, and proposed method.

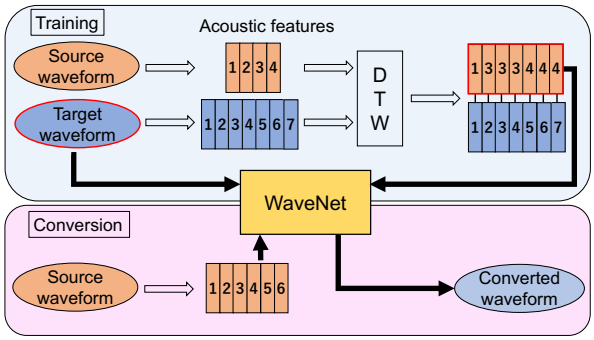

Fig. 2. Overview of the proposed method.

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are the input and output to the activation units, respectively. $*$ is a convolution operator and $\odot$ is an element-wise product operator. $\sigma(\cdot)$ and $\tanh(\cdot)$ represent a sigmoid function and a hyperbolic tangent, respectively. $\boldsymbol{W}$ represents a convolution weight for input. The subscripts f and g represent a filter and gate, respectively. In addition, residual [11] and parameterized skip connections are used throughout the network to speed up convergence and enable training of much deeper models. The network has no pooling layers, and the output of the model has the same time dimensionality as the input. The network outputs a categorical distribution over the next value x_t with a softmax layer.

WaveNet can model the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{h})$ by giving additional inputs $\boldsymbol{h}$:

$$P(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1}, \boldsymbol{h}). \tag{3}$$

The form of the gated activation units of the WaveNet is defined as follows:

$$\begin{aligned} \boldsymbol{z} = \tanh(\boldsymbol{W_f} * \boldsymbol{x} + \boldsymbol{V_f} * \boldsymbol{y}(\boldsymbol{h})) \odot \\ \sigma(\boldsymbol{W_g} * \boldsymbol{x} + \boldsymbol{V_g} * \boldsymbol{y}(\boldsymbol{h})). \end{aligned} \tag{4}$$

where $\boldsymbol{V_f}$ is the convolution weight for the auxiliary features. $\boldsymbol{V_f} * \boldsymbol{y}(\boldsymbol{h})$ and $\boldsymbol{V_g} * \boldsymbol{y}(\boldsymbol{h})$ represent a 1×1 convolution calculation. The variable $\boldsymbol{y}(\boldsymbol{h})$ is an extended time series of the original auxiliary features $\boldsymbol{h}$ to be adjusted to $\boldsymbol{x}$. For TTS, linguistic features, which represent utterance content, are used as auxiliary features. By using vocoder parameters, e.g., mel-cepstral coefficients, fundamental frequency (F_0), and voiced/unvoiced values, as auxiliary features, WaveNet can be used as a vocoder [12].

3. VOICE CONVERSION BASED ON WAVENET

In voice conversion based on statistical models, the relation between acoustic features extracted from source and target audio waveforms is modeled by statistical models. In the conversion step, acoustic features extracted from a source speaker's audio waveform are converted into acoustic features with a target speaker's characteristics by the trained statistical model, and then audio waveforms are generated by inputting the converted acoustic features into a vocoder. Although to model the relation between acoustic features is reasonable and efficient, the model is not optimized for predicting target audio waveforms because the vocoder parameters are used as intermediate representations. To overcome this problem, we propose a voice conversion based on WaveNet to directly generate target audio waveforms from acoustic features of a source speaker.

An overview of the proposed method is shown in Fig. 2. First, acoustic features are extracted from audio waveforms of source and target speakers. Then, time alignments between the extracted source and target feature sequences are obtained by dynamic time warping (DTW) [13]. In the time alignment step, a constraint for target feature sequences that are not warped is applied to keep the relation between the acoustic feature sequence and the sequence of audio waveform samples of the target speaker. Finally, a WaveNet-based voice conversion model is trained from the time-aligned acoustic feature sequences of the source speaker and the audio waveforms of the target speaker. The model is optimized to predict the target audio waveforms from the acoustic features of the source speaker, i.e., it can directly generate audio waveforms with the target speaker's voice characteristics from acoustic features of the source speaker.

In the training of the proposed model, mel-cepstral coefficients extracted from the source speaker's waveforms and log F_0 and voiced/unvoiced values extracted from the target speaker's waveforms are used as additional inputs $\boldsymbol{h}$. Only sequences of mel-cepstral coefficients are warped so that the length of the sequences becomes the same as that of the mel-cepstral coefficients of the target speaker. When converting the source speaker's waveforms, the log F_0 extracted from the waveforms is converted by a simple linear transformation to equalize the mean and variance of the converted and target log F_0:

$$p_t^{(Y)} = \frac{p_t^{(X)} - \mu^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + \mu^{(Y)}, \tag{5}$$

where $p_t^{(Y)}$ and $p_t^{(X)}$ are the converted log F_0 and the original log F_0, respectively. $\mu^{(X)}$ and $\mu^{(Y)}$ are the means, and $\sigma^{(X)}$ and $\sigma^{(Y)}$ are the standard deviations of the training data for the source and the target speakers, respectively. The mel-cepstral coefficients, voiced/unvoiced values, and transformed log F_0 are then input to the trained WaveNet. The extracted acoustic features generally have a lower sampling frequency than that of the audio samples. Therefore, the acoustic features are transformed to sequences with the same time resolution as the audio samples by upsampling and linear interpolation. An overview of the time resolution adjustment of auxiliary features is shown in Fig. 3.

4. EXPERIMENTS

Evaluation results of the proposed WaveNet-based voice conversion are presented in this section. Two subjective evaluations were con-

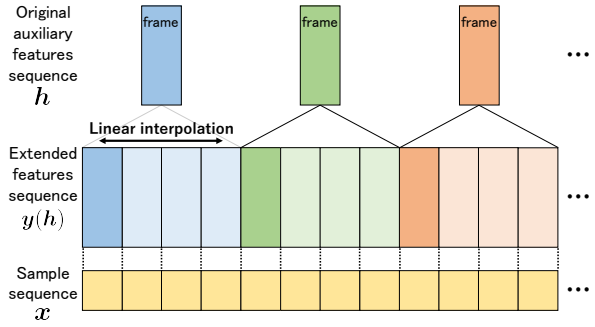

Fig. 3. Overview of time resolution adjustment of auxiliary features.

ducted to evaluate the naturalness and similarity of converted speech.

4.1. Experimental conditions

A Japanese speech database, which was constructed by our research group, was used in the experiments. The database contains a set of 503 phonetically balanced sentences uttered by three male speakers and one female speaker. The set is the same as the B-set of the ATR phonetically balanced Japanese speech database [14]. From the set, 450 sentences were used as training data, with the remaining 53 sentences used as test data. We selected a set consisting of source and target utterances from two male speakers and a set of consisting of those from the other male and female speakers. Speech signals were sampled at 16 kHz, and acoustic features were extracted with a 5-ms shift. Acoustic feature vectors, consisting of 0^{th} through 34^{th} mel-cepstral coefficients, a log F_0 value, and a voiced/unvoiced value, were extracted from the smoothed spectrum analyzed by STRAIGHT [15].

In these experiments, the proposed WaveNet-based voice conversion system (**WaveNet-VC**) was evaluated by comparing it with three systems: **DNN-VC**, **WaveNet-vocoder**, and **DNN-VC+WN-vocoder**.

- **WaveNet-VC**: A WaveNet-based model with 3 blocks (30 layers in total) was used. Specifically, dilations in 10 layers were set to 2^0, 2^1, 2^2, ..., 2^9, and this was repeated three times to form a total of 30 dilated causal convolution layers. The number of channels for dilated causal convolutions and residual connections were 256 and 512, respectively. The Adam algorithm [16] was used for network learning, and its learning rate was manually adjusted to 0.0001 as an initial value. Audio waveforms were 8-bit μ-law [17] encoded.
- **DNN-VC**: A conventional DNN-based voice conversion system. The DNN used in this system was trained from mel-cepstral coefficients and their dynamic features. The architecture of the DNN was a 3-hidden-layer feed-forward neural network with 1024 units per hidden layer. The sigmoid activation function was used in the hidden layers, and the linear activation function was used in the output layer. To obtain a smooth trajectory of spectral features considering the relation between static and dynamic features, maximum likelihood parameter generation (MLPG) [18] was applied to the converted mel-cepstral coefficients. From the smoothed mel-cepstral coefficients, audio waveforms were generated using a mel-log spectrum approximation (MLSA) filter [19].
- **WaveNet-vocoder**: A vocoder rather than a voice conversion system. The architecture of WaveNet for **WaveNet-vocoder** was same as the one for **WaveNet-VC**. For the input features of WaveNet, mel-cepstral coefficients, log F_0 and voiced/unvoiced values extracted from target speaker's speech were used.
- **DNN-VC+WN-vocoder**: The DNN and WaveNet were used in **DNN-VC** and **WaveNet-vocoder**, respectively. The output of the DNN was applied as the input for the WaveNet.

We conducted mean opinion score (MOS) tests [20] to evaluate the naturalness of the converted speech and degradation MOS (DMOS) tests to evaluate the similarity between the target and converted speech samples in terms of speaker characteristics. The opinion score was set on a five-point scale (5: for excellent, 4: for good, 3: for fair, 2: for poor, 1: for bad) in the MOS tests. In the DMOS tests, a difference five-point scale is defined (5: for very similar, 4: for quite similar, 3: for similar, 2: for different, 1: for very different). Fifteen sentences were selected randomly from test data for each subject. There were ten subjects, who were all Japanese.

4.2. Experimental results

Figures 4 and 5 show the results of the MOS and DMOS tests for male-to-male conversion. It can be seen that **DNN-VC+WN-vocoder** outperformed **DNN-VC**. These results indicate that the WaveNet vocoder is able to synthesize much higher quality speech compared to a MLSA filter. Comparing **WaveNet-VC** with**DNN-VC+WN-vocoder**, **WaveNet-VC** obtained a higher score in the MOS test than **DNN-VC+WN-vocoder**. This result clearly shows that the proposed method, **WaveNet-VC**, can generate more naturally sounding converted speech than the system using the WaveNet vocoder, **DNN-VC+WN-vocoder**. In addition, **WaveNet-VC** showed a large improvement from **DNN-VC+WN-vocoder** on the DMOS test. The proposed method can convert voice characteristics more accurately than **DNN-VC+WN-vocoder**. These results indicate that direct modeling from the source speaker's acoustic features to the target speaker's waveforms is effective and the proposed method can improve naturalness and speaker similarity. However, the performance of the proposed method did not reach the performance of **WaveNet-vocoder**. It seems that this degradation was due to the mismatch between the target waveforms and input time-warped mel-cepstral features.

Figures 6 and 7 show the results of the MOS and DMOS tests for male-to-female conversion. It can be seen from Figure 7 that the scores for speaker similarity show a similar trend as observed in male-to-male conversion. However, in Figure 6, **WaveNet-VC** showed the worst score in naturalness. This is because the intelligibility of speech output from **WaveNet-VC** was degraded. The speech converted by the proposed method often include pronunciation errors. It seems that these errors may be due to the mis-alignments between the target waveforms and source acoustic features by DTW. Impacts of the alignment on the performance of **WaveNet-VC** will be investigated in the future.

5. CONCLUSIONS

In this paper, we proposed a WaveNet-based voice conversion model that can directly generate audio waveforms from a source speaker's acoustic features. Experimental results of subjective evaluations showed that the proposed method outperforms a conventional DNN-based method in terms of speaker similarity. Future work includes

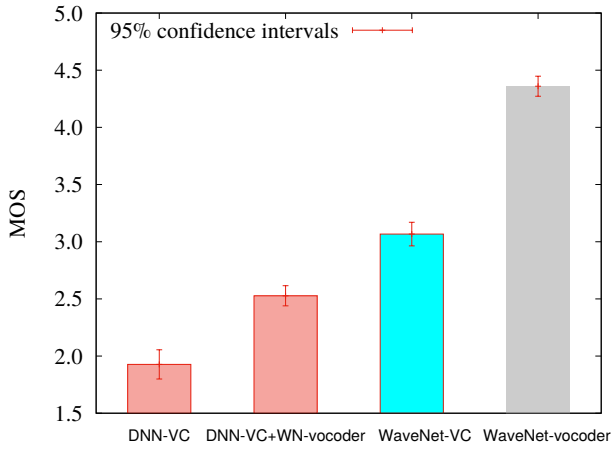

Fig. 4. Mean opinion scores for naturalness (male-to-male voice conversion).

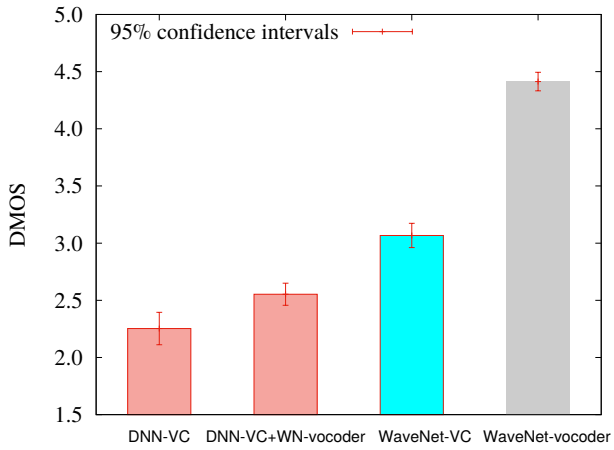

Fig. 5. Degradation mean opinion scores for similarity (male-to-male voice conversion).

the evaluation of the proposed model with larger/smaller databases, the investigation of the auxiliary features for voice conversion based on WaveNet, and applying the proposed approach to many-to-one and one-to-many voice conversion.

6. ACKNOWLEDGEMENTS

This research and development work was partly supported by the MIC/SCOPE #162106106

7. REFERENCES

[1] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," *Proc. of ICASSP*, pp. 301–304, 2001.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. of ICASSP*, pp. 285–288, 1998.

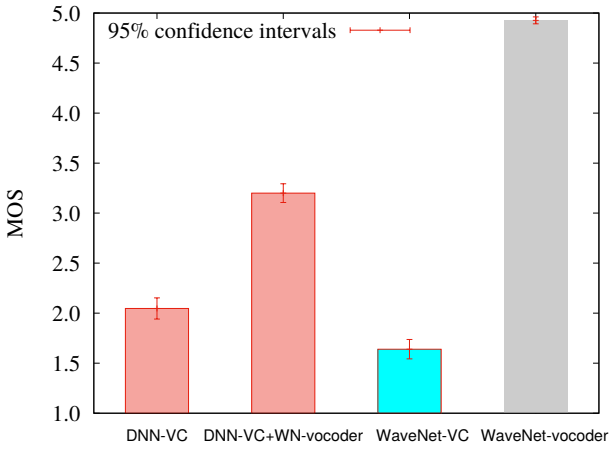

Fig. 6. Mean opinion scores for naturalness (male-to-female voice conversion).

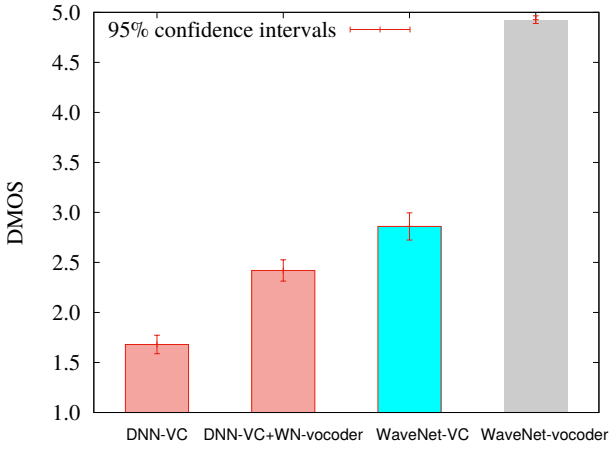

Fig. 7. Degradation mean opinion scores for similarity (male-to-female voice conversion).

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. Speech Audio Process*, pp. 131–142, 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," *Proc. of ICASSP*, pp. 9–12, 2005.

[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using articial neural networks," *Proc. of ICASSP*, pp. 3893–3896, 2009.

[6] Z. Chen and L. H. Zhang, "A ANN Based High Quality Method for Voice Conversion," *Proc. of WiCOM*, pp. 1–4, 2010.

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

[8] N. Hosaka, K. Hashimoto, K. Oura, Y. Nankaku, and K.Tokuda, "Voice conversion based on trajectory model train-

ing of neural networks considering global variance," *Proc. of INTERSPEECH*, pp. 307–311, 2016.

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

[10] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical Voice Conversion with WaveNet-Based Waveform Generation," *Proc. of INTERSPEECH*, pp. 1138–1142, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. of IEEE*, pp. 770–778, 2016.

[12] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-Dependent WaveNet Vocoder," *Proc. of INTERSPEECH*, pp. 1118–1122, 2017.

[13] H. Sakoe and S. Chiba, "Dynamic programing algorithm optimization for spoken word recognition," *Proc. of IEEE*, pp. 43–49, 1978.

[14] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikan, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, pp. 357–363, 1990.

[15] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, pp. 187–207, 1999.

[16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. of ICLR*, 2015.

[17] ITU-T. Recommendation G. 711., "Pulse Code Modulation (PCM) of voice frequencies," 1988.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM based speech synthesis," *Proc. of ICASSP*, pp. 936–939, 2000.

[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptative algorithm for mel-cepstral analysis of speech," *Proc. of ICASSP*, pp. 137–140, 1992.

[20] "Mean opinion score (MOS) terminology," *Proc. of ITU-T*, 2003.