INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Statistical voice conversion with WaveNet-based waveform generation

Kazuhiro Kobayashi[1], *Tomoki Hayashi*[2], *Akira Tamamori*[3], *Tomoki Toda*[1]

[1] Information Technology Center, Nagoya University, Japan
[2] Graduate School of Information Science, Nagoya University, Japan
[3] Institute of Innovation for Future Society, Nagoya University, Japan

{kobayashi.kazuhiro,hayashi.tomoki,tamamori}@g.sp.m.is.nagoya-u.ac.jp,
tomoki@icts.nagoya-u.ac.jp

Abstract

This paper presents a statistical voice conversion (VC) technique with the WaveNet-based waveform generation. VC based on a Gaussian mixture model (GMM) makes it possible to convert the speaker identity of a source speaker into that of a target speaker. However, in the conventional vocoding process, various factors such as F_0 extraction errors, parameterization errors and over-smoothing effects of converted feature trajectory cause the modeling errors of the speech waveform, which usually bring about sound quality degradation of the converted voice. To address this issue, we apply a direct waveform generation technique based on a WaveNet vocoder to VC. In the proposed method, first, the acoustic features of the source speaker are converted into those of the target speaker based on the GMM. Then, the waveform samples of the converted voice are generated based on the WaveNet vocoder conditioned on the converted acoustic features. In this paper, to investigate the modeling accuracies of the converted speech waveform, we compare several types of the acoustic features for training and synthesizing based on the WaveNet vocoder. The experimental results confirmed that the proposed VC technique achieves higher conversion accuracy on speaker individuality with comparable sound quality compared to the conventional VC technique.
Index Terms: voice conversion, WaveNet, vocoder, Gaussian mixture model, deep neural networks.

1. Introduction

The variation of voice characteristics, such as voice timbre and fundamental frequency (F_0) patterns, produced by individual speakers are always restricted by their own physical constraint due to the speech production mechanism. This constraint is helpful for making it possible to produce a speech signal capable of simultaneously conveying not only linguistic information but also non-linguistic information such as speaker individuality. However, it also causes various barriers to produce a desire voice characteristics of the individual speaker. If the individual speakers freely produced various voice characteristics over their own physical constraints, it would break down these barriers and open up an entirely new speech communication style.

Voice conversion (VC) is a potential technique to enable us for producing speech sounds beyond our own physical constraints [1]. VC research was originally started to implement the speaker conversion which enables a source speaker to transform his/her speaker individuality into another target speaker while preserving the linguistic content [2]. A mainstream of VC is a statistical approach to developing a conversion function using a parallel data set consisting of utterances of the source and target speakers. Several techniques based on the statistical conversion models such as Gaussian mixture model (GMM) [3, 4, 5], Gaussian process regression [6, 7] deep neural networks [8, 9], and non-negative matrix factorization [10, 11] have been proposed to convert acoustic features of the source speaker into those of the target speaker. Also, to alleviate the sound quality degradation of the converted voice due to an over-smoothing effect of the converted feature trajectory, several techniques to model additional features such as global variance (GV) [4] and modulation spectrum (MS) [12] have been proposed. However, the sound quality of the converted voices is still obviously degraded compared to that of the natural voices. One of the major factors causing this degradation is the waveform generation process using a vocoder [13]. It is difficult to address this issue as long as using the vocoder even when using high-quality vocoder systems [14, 15, 16, 17].

There are two approaches to avoid the sound quality degradation caused by waveform generation based on the vocoding. First, in VC, it is possible to directly use an original waveform of the source voice for the converted voice. To implement this approach, a direct waveform modification technique using a spectral differential between the source and target speakers, which is estimated based on a differential GMM, has been proposed [18, 19]. Although this technique achieves significant improvements in terms of the sound quality in intra-gender speaker pairs, the sound quality in inter-gender speaker pairs usually deteriorates compared to the intra-gender speaker pair because F_0 transformation and aperiodicity modification cause the sound quality degradation. Second, in statistical parametric speech synthesis [20], WaveNet [21] has been proposed as a technique to directly estimate waveform samples from given input feature vectors such as context labels and logarithmic F_0. This technique achieves remarkable sound quality improvements of the generated speech sounds. Additionally, it is reported that the architecture of WaveNet is capable of utilizing as a waveform generator like the vocoder substituting acoustic features extracted from an original waveform for the input feature vectors [22]. However, in VC, the effectiveness of the WaveNet waveform generation technique has not been confirmed yet. Therefore, it is worthwhile to investigate the effectiveness of the WaveNet-based waveform generation technique.

In this paper, to achieve higher sound quality and conversion accuracy on speaker identity in statistical VC, we propose VC based on the GMM with WaveNet-based waveform generation. In the proposed method, the acoustic features of the source speaker are converted into those of the target speaker based on the GMM as the same manner as the conventional VC. Then, the waveform samples of the converted voice are synthesized based on the WaveNet-based waveform generation technique conditioned on the converted acoustic features. In this paper, we conduct both objective and subjective evaluations. Although the objective evaluation demonstrates that the proposed

http://dx.doi.org/10.21437/Interspeech.2017-986

VC technique is worth than the conventional VC technique, the proposed VC technique achieves higher conversion accuracy on speaker individuality with comparable sound quality in the subjective evaluation.

2. VC based on GMM

VC based on GMM consists of a training process and a conversion process.

In the training process, a joint probability density function of acoustic features of the source and target speakers' voices are modeled with a GMM using a parallel data set. As the acoustic features of the source and target speakers, we employ $2D$-dimensional joint static and dynamic feature vectors $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ of the source and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ of the target consisting of D-dimensional static feature vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ and their dynamic feature vectors $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ at frame t, respectively, where $\top$ denotes the transposition of the vector. Their joint probability density modeled by the GMM is given by

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \quad (1)$$

where $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component. The GMM is trained using joint vectors of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ in the parallel data set, which are automatically aligned to each other by dynamic time warping.

In the conversion process, the acoustic features of the source voice is converted into those of the target voice using maximum likelihood estimation (MLE) of speech parameter trajectory [4]. Time sequence vectors of the source features and the target features are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$, where T is the number of frames included in the time sequence of the given source feature vectors. A time sequence vector of the converted static features $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (2)$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [23]. To alleviate the over-smoothing effects that usually make the converted voice sound muffled, GV is also considered to compensate the variation of converted feature vector sequence.

3. VC with WaveNet-based waveform generation

We propose a technique to generate waveform samples of the converted voice based on WaveNet-based waveform generation in statistical VC. Figure 1 illustrates the conversion process of VC based on the GMM with WaveNet-based waveform generation.

3.1. WaveNet-based waveform generation for VC

In the conventional vocoding process of VC, various assumptions (e.g., a stationary process in the analysis window,

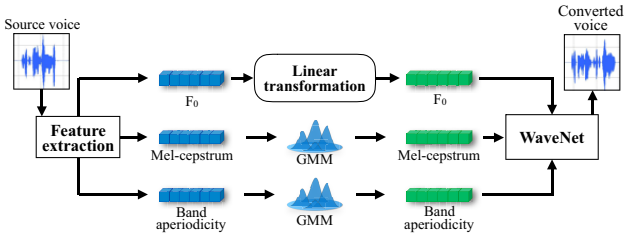

Figure 1: *Conversion process of VC based on the GMM with WaveNet-based waveform generation.*

Gaussianity, a time-invariant linear filter) usually cause the sound quality degradation of the converted voice. To overcome this problem, we apply the speaker-dependent WaveNet vocoder [22], which utilizes the acoustic features of the existing vocoder for a WaveNet auxiliary feature, into VC based on the GMM. The WaveNet vocoder does not require not only involving a filter of articulatory movements with the excitation signals, but also any mathematical assumptions to the data. Therefore, it is expected that detailed time information of the waveform sequence, which is usually lost in the conventional vocoder, can be recovered in the synthesizing process.

The network architecture of WaveNet mainly consists of a stack of one-dimensional convolution layers called dilated causal convolution layer. The input vector passes through these layers and gated activation functions. Finally, WaveNet generates values of the converted waveform sample encoded by μ-law compressor [24], which maximize the posterior probability of the output layer based on the softmax functions. The form of the gated activation function of the WaveNet vocoder is defined as follows:

$$\boldsymbol{z} = \tanh\left(W_f * \boldsymbol{i} + V_f * \boldsymbol{k}\right) \odot \sigma\left(W_g * \boldsymbol{i} + V_g * \boldsymbol{k}\right), \quad (3)$$

where $*$ and $\odot$ denote a convolution operator and an element-wise product operator, respectively. $\sigma(\cdot)$ denotes a sigmoid function. $\boldsymbol{i}$ and $\boldsymbol{z}$ are the input and output vectors of the activation, respectively. f and g represent filter and gate, respectively. W and V indicate convolution weights for input and auxiliary vectors, respectively. The auxiliary feature vector $\boldsymbol{h}$ consisting of the acoustic features such as transformed F_0, converted aperiodicity, and converted spectral feature $\hat{\boldsymbol{y}}$ is transformed into $\boldsymbol{k}$ to make sure to adjust the resolution of the time series data to $\boldsymbol{i}$.

3.2. Alleviation of the mismatch between training and conversion data of WaveNet using intra-speaker conversion

In GMM-based VC, the converted feature trajectory becomes smoother than the target feature trajectory due to the over-smoothing effect. Therefore, in the conversion process, it is assumed that a mismatch between the training and synthesizing data for the WaveNet vocoder may cause the serious errors of the waveform generation. To alleviate this mismatch, we utilize the over-smoothed acoustic feature of the target speaker for the training data of the WaveNet vocoder. However, it is not straightforward to apply the converted feature converted from the source speaker into the target speaker, because the duration between the source and target features is usually different. To address this issue, we use an intra-speaker conversion technique and utilize its converted feature for the training data of the WaveNet vocoder, where the intra-speaker conversion is a technique to convert acoustic features within an identical speaker based on an intra-speaker GMM [25]. The probability density

function of the intra-speaker GMM is defined by

$$P\left(\boldsymbol{Y}_t, \boldsymbol{Y}'_t|\boldsymbol{\lambda}^{(YXY)}\right) = \sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t\\ \boldsymbol{Y}'_t\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}\\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)}\\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \quad (4)$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\boldsymbol{\Sigma}_m^{(XY)}, \quad (5)$$

where $\boldsymbol{\lambda}^{(YXY)}$ indicates the parameter set of the intra-speaker GMM. Based on this intra-speaker GMM, the acoustic feature of the target speaker into converted into the smoothed acoustic feature of the target speaker $\overline{\boldsymbol{y}}$ in the same manner as MLE described in Sect. 2. Note that this converted feature vector $\overline{\boldsymbol{y}}$ has the same length of the natural acoustic feature of the target voice.

3.3. Post-filtering using spectral differential

In this paper, we also propose a technique to modify spectral envelopes based on the spectral differential. In the proposed method, the converted voice sometimes causes serious buzzy sounds. To suppress this unstable sounds, we filter the converted voice with the spectral differential between converted spectral feature of the VC based on GMM and spectral feature extracted from the converted voice. The spectral differential is calculated as

$$\boldsymbol{d} = \hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}^{(\boldsymbol{WN})}, \quad (6)$$

where $\hat{\boldsymbol{y}}^{(\boldsymbol{WN})}$ indicates the spectral feature extracted from the converted voice based on VC with the WaveNet vocoder.

4. Experimental evaluation

To examine the effectiveness of the proposed methods, we compared following methods.

- The NU-NAIST VC system: the conventional VC method submitted to the Voice Conversion Challenge 2016 [19].
- WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$): the proposed VC method with the WaveNet vocoder modeled by natural mel-cepstrum and synthesizes using converted mel-cepstrum based on the GMM.
- WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}^{(GV)}$): the proposed VC method with the WaveNet vocoder modeled by target mel-cepstrum and synthesizes using converted mel-cepstrum based on the GMM considering GV.
- WaveNet VC ($\overline{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}$): the proposed VC method with the WaveNet vocoder modeled by intra-speaker converted mel-cepstrum and synthesizes using converted mel-cepstrum based on the GMM.
- WaveNet VC ($\overline{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}^{(GV)}$): the proposed VC method with the WaveNet vocoder modeled by intra-speaker converted mel-cepstrum and synthesizes using converted mel-cepstrum based on the GMM considering GV.
- WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$: the proposed VC method with the WaveNet vocoder modeled by natural mel-cepstrum and synthesizes using converted mel-cepstrum based on the GMM, and post-filtering the spectral differential between converted mel-cepstrum based on GMM considering GV and mel-cepstrum extracted from converted voice.

Table 1: *Mel-cepstral distortions of several conversion methods.*

Method	Mel-CD [dB]
The NU-NAIST VC system	5.55
WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$)	6.75
WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}^{(GV)}$)	6.84
WaveNet VC ($\overline{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}$)	7.12
WaveNet VC ($\overline{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}^{(GV)}$)	7.17
WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$	5.46

4.1. Experimental conditions

We evaluated sound quality and speaker identity to compare the performance of the conventional and proposed methods. We used the ARCTIC speech database [26]. We used "bdl" and "slt" for the source speaker and "clb" and "rms" for the target speaker. The number of combinations of the source and target speaker was 4. The number of sentences uttered by each speaker was 1132. The sampling frequency was set to 16 kHz.

STRAIGHT [14] was used to extract spectral envelope, which was parameterized into the 1-24th mel-cepstral coefficients as the spectral feature. The frame shift was 5 ms. As the source excitation features, we used F_0 and aperiodic components extracted with STRAIGHT [27]. The aperiodic components were averaged over five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, to be modeled with the GMM.

We used 1028 sentences for training and the remaining 104 sentences were used for evaluation. The speaker-dependent GMMs were separately trained for the individual source and target speaker pairs. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64.

The WaveNet models were trained for the individual target speakers. Considering one layer of dilated causal convolution, gate activated function, and residual as one block, we connected the 30 residual blocks in total. Specifically, dilations in 10 layers were set to 2^0, 2^1, 2^2, $\cdots$, 2^9, and this was repeated three times to form a total of 30 dilated causal convolution layers. The number of channels of (dilated) causal convolution and 1×1 convolution in the residual block was set to 256. The number of 1×1 convolution channel between skip-connection and softmax layer was set to 2,048. Adam algorithm [28] was used for network learning, and its learning rate was manually adjusted to 0.001 as an initial value, and attenuation schedule was adjusted. The mini batch size was 20,000 samples. In addition to the converted mel-cepstrum, we used transformed F_0 and converted aperiodic components for the auxiliary features in both modeling and synthesizing process for the WaveNet vocoder.

4.2. Objective evaluation

In the objective evaluation, we compared the mel-cepstral distortions (Mel-CD) between the target and converted voice. The Mel-CD was calculated as

$$\text{Mel-CD}\ [dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(X)} - mc_d^{(Y)}\right)^2}, \quad (7)$$

where $mc_d^{(X)}$ and $mc_d^{(Y)}$ represent the d-th dimensional component of the converted mel-cepstrum extracted from the converted voice and the original mel-cepstrum of the target speaker, respectively.

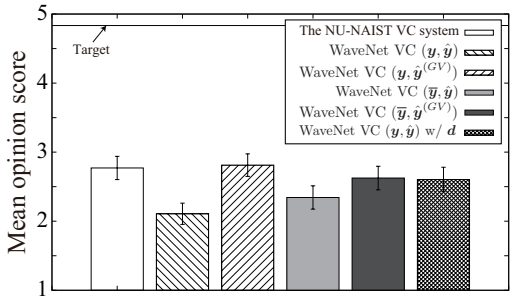

Figure 2: *Sound quality of converted voice.*

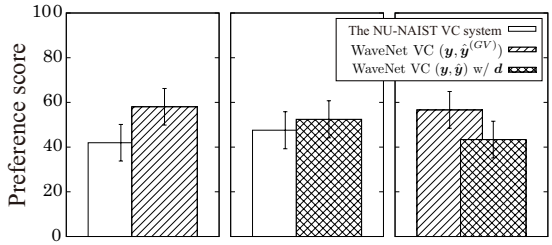

Figure 3: *Comparison of conversion accuracy on speaker identity.*

Table 1 indicates the experimental results of the Mel-CD between the target mel-cepstrum and the converted mel-cepstram extracted from the converted voice. We can see that the Mel-CDs of the proposed methods are significantly larger than the Mel-CD of the NU-NAIST VC system. In particular, there is a tendency to increase the Mel-CD when considering GV in the proposed methods. In terms of the WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$ method, the Mel-CD is similar to that of the NU-NAIST VC system because its spectral feature is modified using the spectral differential between VC considering GV and WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$. Therefore, it can be said that the converted voice of the WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$ method consists of the excitation signal synthesized by the WaveNet vocoder and spectral envelopes of the GMM-based VC.

4.3. Subjective evaluation

Two subjective evaluations were conducted. In the first test, we evaluated the sound quality of the converted voices using a mean opinion score (MOS). The natural and converted voice samples generated by the conventional and proposed VC methods were presented to subjects in random order. The subjects rated the quality of the converted voice using a 5–point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. The number of evaluation sentences in each subject was 84. The number of subjects was 8.

In the second test, conversion accuracy on speaker individuality was evaluated. In this test, we chose three different methods of The NU-NAIST VC system, WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}^{(GV)}$), and WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$ in order to reduce the cost for evaluation in each subject. A natural voice sample of the target speaker was presented to the subjects first as a reference. Then, the converted voice samples generated by these techniques for the same sentences were presented in random order. The subjects selected which sample was more similar to the reference natural voice in terms of speaker identity. Each subject evaluated 39 sample pairs. The number of subjects was 11.

Subjects were not native English speakers and they were allowed to replay each sample pair as many times as necessary in both evaluations.

4.4. Experimental results

Figure 2 indicates the results of the MOS test for the sound quality. We can see that there are almost equivalent performances of the sound quality between the conventional and proposed methods. The conventional method can generate the converted voice with steady sound quality over all frames. On the other hand, although the proposed methods can generate the converted voice with the quite better sound quality compared to the conventional method, these methods sometime cause serious buzzy sounds. It is considered that these buzzy sounds are derived from the less training data for the WaveNet vocoder. Therefore, it is expected that the increase of the training utterances can improve the sound quality of the converted voice in the proposed methods. As for a comparison between the WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) and WaveNet VC ($\overline{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}$) methods, we can see that the WaveNet modeling using intra-speaker converted features slightly improves the sound quality by avoiding of the inconsistency of the training and synthesizing data. In terms of a comparison between w/ and w/o GV, the techniques considering GV achieve higher sound quality compared to the techniques without considering GV. These results demonstrate that the WaveNet vocoder modeled using natural acoustic features and synthesizing using converted feature considering GV contributes the sound quality improvements.

Figure 3 indicates the results of the preference test for speaker identity. We can see the WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}^{(GV)}$) method achieves higher conversion accuracy on speaker identity compared to the other methods. Also, there is not a large difference between the NU-NAIST VC system and WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}$) w/ $\boldsymbol{d}$ because the spectral feature of these techniques is almost equaled. From these results, it can be assumed that the WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}^{(GV)}$) method can implement further restoration of the speaker individuality, is usually lost in the conventional VC based on GMM.

These results suggest that the WaveNet VC ($\boldsymbol{y}$, $\hat{\boldsymbol{y}}^{(GV)}$) method is the best conversion technique in terms of sound quality and conversion accuracy on speaker identity.

5. Conclusions

This paper describes a technique to convert speaker individuality of a source speaker into that of a target speaker with Gaussian mixture model (GMM)-based voice conversion (VC) and WaveNet-based waveform generation. In order to improve the sound quality and conversion accuracy of the speaker individuality in VC, we propose a waveform generation technique based on the WaveNet vocoder conditioned on the converted acoustic features such as transformed F_0, converted aperiodicity, and converted mel-cepstrum. The experimental results demonstrated that the WaveNet-based waveform generation technique using natural acoustic features for modeling and converted acoustic features considering GV for synthesizing achieves higher conversion accuracy on speaker identity with comparable sound quality compared to the NU-NAIST VC system submitted Voice Conversion Challenge 2016. In future work, we plan to implement a technique for enabling stability for the WaveNet-based waveform generation technique.

6. Acknowledgements

This work was supported in part by JSPS KAKENHI Grant-in-Aid for JSPS Research Fellow Number 16J10726, by JSPS KAKENHI Grant Number 15H02726, and by JST, PRESTO Grant Number JPMJPR1657.

7. References

[1] T. Toda, "Augmented speech production based on real-time statistical voice conversion," *Proc. GlobalSIP*, pp. 755–759, Dec. 2014.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

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

[6] N. Pilkington, H. Zen, and M. Gales, "Gaussian process experts for voice conversion," *Proc. INTERSPEECH*, pp. 2761–2764, Aug. 2011.

[7] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, and Z. Yang, "Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data," *Speech Communication*, vol. 58, pp. 124–138, Mar. 2014.

[8] 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 Trans. ASLP*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

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

[10] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," *IEICE Trans. Inf. and Syst.*, vol. E96-A, no. 10, pp. 1946–1953, Oct. 2013.

[11] Z. Wu, T. Virtanen, E. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. ASLP*, vol. 22, no. 10, pp. 1506–1521, June 2014.

[12] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Trans. ASLP*, vol. 24, no. 4, pp. 755–767, Jan. 2016.

[13] H. Dudley, "Remaking speech," *JASA*, vol. 11, no. 2, pp. 169–177, 1939.

[14] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[15] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. SAP*, vol. 9, no. 1, pp. 21–29, 2001.

[16] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J-STSP*, vol. 8, no. 2, pp. 184–194, 2014.

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. and Syst.*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[18] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," *Proc. IEEE SLT*, pp. 693–700, Dec. 2016.

[19] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," *Proc. INTERSPEECH*, Sept. 2016.

[20] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[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," *CoRR*, vol. abs/1609.03499, 2016.

[22] A. Tamamori, T. Hayashi, T. Toda, and K. Takeda, "Speech waveform synthesis based on wavenet considering speech generation process," *IEICE Tech. Rep. SP2016-77 (Japanese edition)*, no. 477, pp. 1–6, Mar. 2017.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, pp. 1315–1318, June 2000.

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

[25] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G.Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," *IEICE Trans. Inf. and Syst.*, vol. E97-D, no. 6, pp. 1419–1428, 2014.

[26] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," *Tech. Rep. CMU-LTI-03-177*, 2003.

[27] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight," *Proc. MAVEBA*, pp. 13–15, Sept. 2001.

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