INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016

Kazuhiro Kobayashi[1], *Shinnosuke Takamichi*[1], *Satoshi Nakamura*[1], *Tomoki Toda*[2]

[1]Nara Institute of Science and Technology (NAIST), Japan
[2]Information Technology Center, Nagoya University, Japan
[1]{kazuhiro-k, shinnosuke-t, s-nakamura}@is.naist.jp, [2]tomoki@icts.nagoya-u.ac.jp

Abstract

This paper presents the NU-NAIST voice conversion (VC) system for the Voice Conversion Challenge 2016 (VCC 2016) developed by a joint team of Nagoya University and Nara Institute of Science and Technology. Statistical VC based on a Gaussian mixture model makes it possible to convert speaker identity of a source speaker' voice into that of a target speaker by converting several speech parameters. However, various factors such as parameterization errors and over-smoothing effects usually cause speech quality degradation of the converted voice. To address this issue, we have proposed a direct waveform modification technique based on spectral differential filtering and have successfully applied it to singing voice conversion where excitation features are not necessary converted. In this paper, we propose a method to apply this technique to a standard voice conversion task where excitation feature conversion is needed. The result of VCC 2016 demonstrates that the NU-NAIST VC system developed by the proposed method yields the best conversion accuracy for speaker identity (more than 70% of the correct rate) and quite high naturalness score (more than 3 of the mean opinion score). This paper presents detail descriptions of the NU-NAIST VC system and additional results of its performance evaluation.

Index Terms: voice conversion challenge 2016, speaker identity, segmental feature, Gaussian mixture model, STRAIGHT analysis.

1. Introduction

Varieties 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 identity. However, it also causes various barriers in speech communication; e.g., severe vocal disorders are easily caused even if speech organs are partially damaged; and we hesitate to talk about something private using a cell phone if we are surrounded by others. 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 make it possible for us to produce speech sounds beyond our own physical constraints [1]. VC research was originally started to achieve speaker conversion to make it possible to transform the voice identity of a source speaker into that of a 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. As one of the most popular statistical VC methods, a regression method using a Gaussian mixture model (GMM) was proposed [3]. To improve performance of the GMM-based VC method, various VC methods have been proposed by implementing more sophisticated techniques, such as Gaussian process regression [4, 5] deep neural networks [6, 7], non-negative matrix factorization [8, 9], and so on. We have also significantly improved performance of the standard GMM-based VC method by incorporating a trajectory-based conversion algorithm to make it possible to consider temporal correlation in conversion [10], modeling additional features to alleviate an over-smoothing effect of the converted speech parameters, such as global variance (GV) [10] and modulation spectrum (MS) [11], and implementing STRAIGHT [12] with mixed excitation [13]. Furthermore, a real-time conversion process has also been successfully implemented for state-of-the-art GMM-based VC [14]. However, the speech quality of the converted voices is still obviously degraded compared to that of the natural voices. One of the biggest factors causing this quality degradation is the waveform generation process using a vocoder [15], which is still observed even when using high-quality vocoder systems [12, 16, 17].

In singing VC (SVC), to avoid the quality degradation caused by the vocoding process [15], we have proposed an intra-gender SVC method with direct waveform modification based on spectrum differential (DIFFSVC) [18] considering global variance (GV) [19], focusing on F_0 transformation is not necessary in the intra-gender SVC. The DIFFSVC framework can avoid using the vocoder by directly filtering an input singing voice waveform with a time sequence of spectral parameter differentials estimated by a differential GMM (DIFFGMM) analytically derived from the conventional GMM used in the standard method. Moreover, to apply this DIFFSVC framework to cross-gender DIFFSVC as well, we have proposed an F_0 transformation technique with direct residual signal modification [20] based on time-scaling with waveform similarity-based overlap-add [21] and resampling.

In this paper, we develop a new VC system for speaker conversion based on the direct waveform modification technique, which was submitted to the Voice Conversion Challenge 2016 (VCC 2016) [22] from our joint team of Nagoya University and Nara Institute of Science and Technology (NAIST) as the NU-NAIST VC system (called "new NAIST VC system"). The following techniques are newly implemented for our GMM-based VC system: 1) voice conversion with direct waveform modification with spectral differential (DIFFVC), 2) speech parameter trajectory smoothing in the GMM training, 3) post-filtering process based on MS for DIFFVC, and 4) excitation conver-

http://dx.doi.org/10.21437/Interspeech.2016-970

sion (EC) using STRAIGHT as preprocessing of spectral conversion. The results of the VCC 2016 have demonstrated that the NU-NAIST VC system (system "J") achieved the best conversion accuracy on speaker identity and high naturalness (more than 3 on the mean opinion score scale). In this paper, we also conduct subjective evaluations, demonstrating that the NU-NAIST VC system achieves high speech quality and conversion accuracy comparable to our conventional GMM-based VC system.

2. VC based on GMM

In the conventional VC, acoustic features such as spectral features and aperiodic components of a source speaker are converted into those of a target speaker based on previously trained GMMs. F_0 is transformed to compensate for the difference in pitch between the source and target speakers based on frame-by-frame linear conversion. Finally, the converted voice is generated by synthesizing these converted acoustic features using a vocoder.

2.1. Acoustic feature mapping based on GMM

Acoustic feature mapping 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 speaker' 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

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

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 speaker are converted into that of the target speaker using maximum likelihood estimation (MLE) of speech parameter trajectories using the GMM and GV [10].

2.2. F_0 transformation

In both intra- and cross-gender conversions, F_0 is transformed frame-by-frame in order to line up pitch differences between source and target speakers.

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)}, \quad (2)$$

where x_t and $\hat{y}_t$ are a log-scaled F_0 of the source speaker and the converted one at frame t. $\mu^{(x)}$ and $\sigma^{(x)}$ are the mean and standard deviation of log-scaled F_0 of the source speaker and $\mu^{(y)}$ and $\sigma^{(y)}$ are those of the target speaker.

3. The NU-NAIST VC system for VCC 2016

In this paper, we proposed the following techniques: 1) DIFFVC, 2) GMM training with smoothed speech parameter trajectory, 3) post-filtering process based on modulation spectrum (MS) for DIFFVC, and 4) excitation conversion with F_0 and aperiodic components transformations using a vocoder. Figure 1 indicates the conversion flow of the NU-NAIST VC system for the VCC 2016. The NU-NAIST VC system performs excitation and spectral conversion. During excitation conversion, F_0 values and aperiodic components extracted from a source voice are transformed within an analysis/synthesis framework using a vocoder. During spectral conversion, spectral features of the source voice are converted into spectral feature differentials based on the DIFFGMM. Next, MS-based post-filtering is applied to the spectral feature differential. Finally, the converted speech waveform is generated by directly filtering the analysis-synthesized speech waveform generated during the excitation conversion step using the post-filtered spectral feature differentials.

3.1. DIFFVC based on DIFFGMM

As part of the modelling step, the DIFFGMM is analytically derived from the traditional GMM (in Eq. (3)). Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the static and dynamic differential feature vector, where $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$, the DIFFGMM is derived by transforming model parameters in the same manner as DIFFSVC [18] as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{D}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(D)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XD)}\\ \boldsymbol{\Sigma}_m^{(DX)} & \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right). \quad (3)$$

During the conversion step, a time sequence of the D-dimensional converted spectral feature differentials, $\hat{\boldsymbol{d}}$, is determined using MLE of the speech parameter trajectory using the DIFFGMM [18]. Then, the converted speech waveform is generated by directly filtering an input speech waveform with a time-variant synthesis filter designed from the spectral feature differential sequence. This filtering process modifies a spectral envelope sequence while basically preserving the natural excitation signals of the input speech waveform.

3.2. Speech parameter trajectory smoothing

Modulation Spectrum (MS) [11] is defined as the log-scaled power spectrum of the parameter sequence; i.e., temporal fluctuation of the parameter sequence is decomposed into individual modulation frequency components and their power values are represented as the MS. The MS, $\boldsymbol{s}(\boldsymbol{y})$, of the parameter sequence $\boldsymbol{y}$ is defined as:

$$\boldsymbol{s}(\boldsymbol{y}) = \left[\boldsymbol{s}_1(\boldsymbol{y})^\top, \cdots, \boldsymbol{s}_d(\boldsymbol{y})^\top, \cdots, \boldsymbol{s}_D(\boldsymbol{y})^\top\right]^\top, \quad (4)$$

$$\boldsymbol{s}_d(\boldsymbol{y}) = \left[s_{d,0}(\boldsymbol{y}), \cdots, s_{d,f}(\boldsymbol{y}), \cdots, s_{d,D_\mathrm{s}-1}(\boldsymbol{y})\right]^\top, \quad (5)$$

where $2D_\mathrm{s}$ is the length of the discrete Fourier transform, and $s_{d,f}(\boldsymbol{y})$ is the f-th MS of the d-th dimension of the parameter sequence $[\boldsymbol{y}_1(d), \cdots, \boldsymbol{y}_T(d)]^\top$. f is the modulation frequency index. As reported in [23, 24], the higher modulation frequency components (more fluctuating components of a temporal sequence) of spectral parameter sequences are negligible for speech quality. By applying a low-pass filter (LPF) that removes the higher modulation frequency components (e.g., more than 50 Hz ($f > D_\mathrm{s}/2$)), we can improve training accuracy

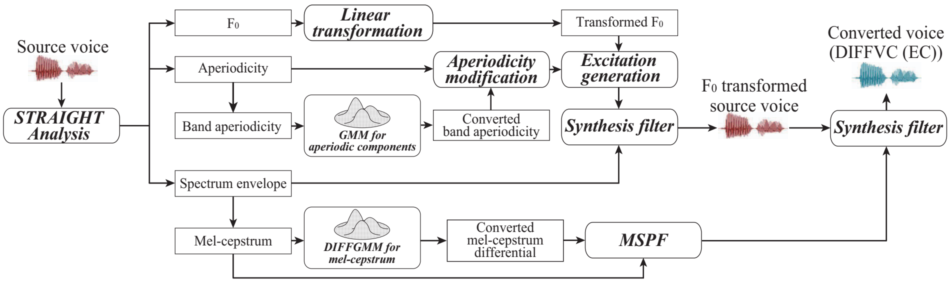

Figure 1: Conversion process of the NU-NAIST VC system for VCC 2016.

of acoustic models as done for hidden Markov model-based speech synthesis [25]. Here, source and target speakers' speech parameter sequences, $\boldsymbol{x}$ and $\boldsymbol{y}$, are LPFed, then the LPFed sequences, $\boldsymbol{x}^{(\mathrm{LPF})}$ and $\boldsymbol{y}^{(\mathrm{LPF})}$, are used to train the GMM. In conversion, $\boldsymbol{x}^{(\mathrm{LPF})}$ is used to generate the spectral differentials.

3.3. MS-based post-filter for VC with spectral differentials

Statistical modeling tends to deteriorate MSs of the converted speech parameters, and keeping natural MSs is strongly effective for improving the quality of the converted speech. An MS-based post-filter (MSPF) [11], which is applied after speech parameter conversion in conventional GMM-based VC, modifies a converted speech parameter sequence so that the sequence has the target speaker's natural MS. Here, we propose an MS-based post-filtering process that modifies spectral differentials, $\hat{\boldsymbol{d}}$, such that the finally synthesized speech has the target speaker's natural MS.

In training, we calculate MS statistics for target speaker's natural and converted speech parameters from the training data, $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}} = [\hat{\boldsymbol{d}} + \boldsymbol{x}^{(\mathrm{LPF})}]$. Here, let $\mu_{d,f}^{(y)}$ and $\mu_{d,f}^{(\tilde{y})}$ be the mean of $s_{d,f}(\boldsymbol{y})$ and $s_{d,f}(\tilde{\boldsymbol{y}})$, and let $\sigma_{d,f}^{(y)}$ and $\sigma_{d,f}^{(\tilde{y})}$ be their variance. The $\hat{\boldsymbol{d}}$ is generated by converting $\boldsymbol{x}^{(\mathrm{LPF})}$.

In conversion, $\boldsymbol{x}^{(\mathrm{LPF})}$ is first added to the generated $\hat{\boldsymbol{d}}$. Then, the MS, $s_{d,f}(\tilde{y})$ is converted as follows:

$$s'_{d,f}(\tilde{y}) = \frac{\sigma_{d,f}^{(y)}}{\sigma_{d,f}^{(\tilde{y})}} \left(s_{d,f}(\tilde{y}) - \mu_{d,f}^{(\tilde{y})} \right) + \mu_{d,f}^{(y)}. \tag{6}$$

The converted $\tilde{\boldsymbol{y}}$ is determined using the converted MS and the original phase components. The MSPFed spectral differentials, $\hat{\boldsymbol{d}}^{(\mathrm{MSPF})}$ can be determined by subtracting $\boldsymbol{x}^{(\mathrm{LPF})}$ from the converted $\tilde{\boldsymbol{y}}$ [1]. Note that, in this paper, we use mean-normalized MSs and adopt a segment-level post-filtering process [11].

3.4. Excitation conversion based on F_0 and aperiodicity transformations using a vocoder

Although we initially tried implementing the F_0 transformation technique with direct residual signal modification [20] for singer conversion, we found that this technique was not effective for speaker conversion. In speaker conversion, we need to handle larger acoustic differences in excitation signals between the source and target speakers compared to singing voice conversion. To address this issue, we implemented excitation conversion using STRAIGHT [26] as high-quality vocoder. For the F_0 transformation, we perform the global linear transformation as described in Sect 2.2. For the aperiodic components, band-averaged aperiodic components are extracted and converted with the GMM as in the conventional method [13]. Then, original aperiodic components at all frequency bins are shifted using aperiodic differentials between the extracted and converted band-averaged aperiodic components. Finally, analysis-synthesized speech is generated from these transformed excitation parameters using STRAIGHT. Note that full STRAIGHT spectral representation is directly used in synthesis.

This excitation conversion method actually causes significant quality degradation because original phase information is discarded. Nevertheless, we have found that this method yields better speech quality as well as better conversion accuracy than the direct residual signal modification [20].

4. Experimental evaluation

In this section, we show results of the VCC 2016 to demonstrate performance of the NU-NAIST VC system. Moreover, we compare the following three systems:

- DIFFVC (EC): The NU-NAIST VC system submitted to the VCC 2016,
- VC: Our conventional VC system [13],
- DIFFVC: The NU-NAIST VC system w/o excitation conversion.

4.1. Experimental conditions

We evaluated speech quality and speaker identity of the converted voices to compare performance of the different VC systems in both intra-gender and cross-gender conversion tasks. We used the English speech database used in the VCC 2016. The number of source speakers was 5 including 3 females and 2 males, and that of the target speakers was 5 including 2 females and 3 males who were different from the source female and male speakers. The number of sentences uttered by each speaker was 216. The sampling frequency was set to 16 kHz.

STRAIGHT [12] was used to extract spectral envelopes, which was parameterized into the 1-24th mel-cepstral coefficients as the spectral feature. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [27] was used as the synthesis filter. As the source excitation features, we used F_0 and aperiodic components extracted with STRAIGHT [26]. 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 162 sentences for training and the remaining 54 sentences were used for evaluation. The speaker-dependent GMMs were separately trained for all combinations of source and target speaker pairs. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64.

Two preference tests were conducted. In the first test, speech quality of the converted voices was evaluated. The converted voice samples generated by two different VC systems for the same sentences were presented to subjects in random order. The subjects selected which sample had better speech quality.

[1] Note that, because the MSPF process is non-linear to the speech parameter sequence, the sequence that $\boldsymbol{x}^{(\mathrm{LPF})}$ is subtracted from the converted $\tilde{\boldsymbol{y}}$ is not equal to $\hat{\boldsymbol{d}}$.

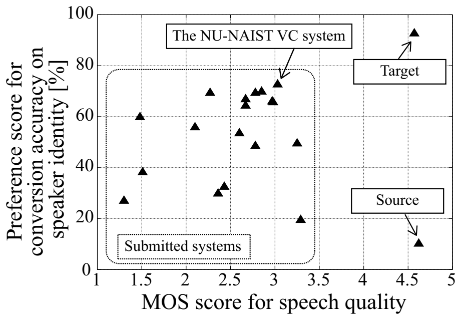

Figure 2: Sound quality and conversion accuracy on speaker identity in VCC 2016.

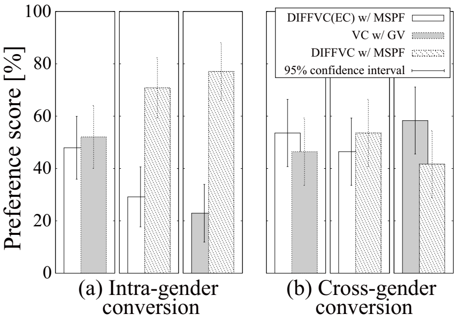

Figure 3: AB preference test for speech quality.

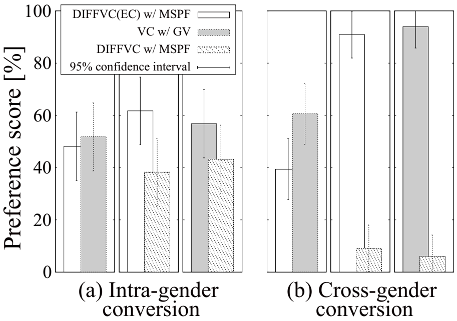

Figure 4: XAB test for conversion accuracy on speaker identity.

In the second test, conversion accuracy in speaker identity was evaluated. A natural voice sample of the target speaker was presented to the subjects first as a reference. Then, the converted voice samples generated by two different VC systems 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. The number of subjects was 10 and each listener evaluated 54 sample pairs in each evaluation. They were allowed to replay each sample pair as many times as necessary.

4.2. Results of the VCC 2016

Figure 2 indicates an overall result of the VCC 2016. The NU-NAIST VC system achieved quite high speech quality over 3.0 of MOS and the best conversion accuracy (about 74%) among all submitted VC systems. In terms of the conversion accuracy, our system achieved successful performance even though very simple prosodic conversion was performed. However, it is observed that there is still a large gap between the converted voices and the natural target voices. It is expected that further improvements will be yielded by implementing a conversion method of prosodic patterns or asking the source speakers to mimic target prosodic patterns, which would be possible in several practical applications. In terms of speech quality, the NU-NAIST VC system causes serious quality degradation compared to natural voices, i.e., from 4.6 to 3.0 in MOS. This quality degradation is mainly caused by using a vocoder to perform the excitation conversion as shown in the next section. Therefore, it is expected that the converted speech quality will be significantly improved by developing a better analysis/synthesis technique than STRAIGHT.

4.3. Results of subjective evaluation

Figures 3 (a) and (b) indicate the results of the preference test for speech quality. DIFFVC (EC) achieves equivalent speech quality compared to VC in both intra/cross-gender conversions. On the other hand, DIFFVC achieves significantly higher speech quality compared to the other two methods in the intra-gender conversion. This is because DIFFVC can avoid using vocoding to generate converted speech waveforms, making the conversion process free from various errors, such as F_0 extraction errors and unvoiced/voiced decision errors. Note that DIFFVC in cross-gender conversion condition does not result in any significant quality improvements as it suffers from mismatches between spectral envelope and F_0 in the cross-gender conversion.

Figures 4 (a) and (b) indicate the results of the preference test for speaker identity. Although DIFFVC (EC) has equivalent conversion accuracy compared to VC in the intra-gender conversion, it tends to be degraded in the cross-gender conversion. It is expected that the residual spectral envelope preserved in the direct waveform modification process still includes speaker-dependent or gender-dependent features, and that this causes adverse effects on conversion accuracy.

These results suggest that 1) the NU-NAIST VC system demonstrating the best conversion accuracy and high speech quality in the VCC 2016 has an almost equivalent performance compared to the conventional VC system in both intra-gender and cross-gender conversions, and 2) the direct waveform modification technique achieves significantly higher converted speech quality compared to the conventional VC system if the excitation conversion is not necessary as in the intra-gender conversion, and therefore, there is still large room to improve the converted speech quality of the NU-NAIST VC system.

5. Conclusions

This paper describes the details of the NU-NAIST voice conversion (VC) system for the Voice Conversion Challenge 2016 (VCC 2016) developed by a joint team of Nagoya University and Nara Institute of Science and Technology. In order to improve the quality of statistical VC based on Gaussian Mixture Model (GMM), we applied the following techniques: 1) voice conversion with direct waveform modification with spectral differential (DIFFVC), 2) speech parameter trajectory smoothing, 3) post-filtering based on modulation spectrum for DIFFVC, and 4) preprocessing for excitation conversion with F_0 and aperiodic component transformations using high-quality vocoding. The experimental results demonstrated that the NU-NAIST VC system was highly ranked in the VCC 2016, its performance was comparable to our conventional VC system, and the DIFFVC technique showed large potential to significantly improve the converted speech quality of the NU-NAIST VC system. In future work, we plan to implement high quality F_0 and aperiodicity transformation for the DIFFVC technique.

Acknowledgements This work was supported in part by JSPS KAKENHI Grant Number 26280060 and Grant-in-Aid for JSPS Research Fellow Number 16J10726.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] S. Takamichi, T. Toda, A. W. 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.

[12] 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.

[13] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Proc. INTERSPEECH*, pp. 2266–2269, Sept. 2006.

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

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

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

[17] 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.

[18] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," *Proc. INTERSPEECH*, pp. 2514–2418, Sept. 2014.

[19] ——, "Statistical singing voice conversion based on direct waveform modification with global variance," *Proc. INTERSPEECH*, pp. 2754–2758, Sept. 2015.

[20] K. Kobayashi, T. Toda, and S. Nakamura, "Implementation of f0 transformation for statistical singing voice conversion based on direct waveform modification," *Proc. ICASSP*, pp. 5670–5674, Mar. 2016.

[21] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *Proc. ICASSP*, pp. 554–557 vol.2, Apr. 1993.

[22] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," *Proc. INTERSPEECH*, Sept. 2016.

[23] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," *Proc. ICASSP*, Apr. 2015.

[24] ——, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," *Proc. ICASSP*, Apr. 2015.

[25] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The naist text-to-speech system for the blizzard challenge 2015," *Proc. Blizzard Challenge workshop*, Sept. 2015.

[26] 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*, Sept. 2001.

[27] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," *Proc. ICSLP*, pp. 1043–1045, 1994.