F0 TRANSFORMATION TECHNIQUES FOR STATISTICAL VOICE CONVERSION WITH DIRECT WAVEFORM MODIFICATION WITH SPECTRAL DIFFERENTIAL

Kazuhiro Kobayashi[1], Tomoki Toda[2], Satoshi Nakamura[1]

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

ABSTRACT

This paper presents several F_0 transformation techniques for statistical voice conversion (VC) with direct waveform modification with spectral differential (DIFFVC). Statistical VC is a technique to convert speaker identity of a source speaker's voice into that of a target speaker by converting several acoustic features, such as spectral and excitation features. This technique usually uses vocoder to generate converted speech waveforms from the converted acoustic features. However, the use of vocoder often causes speech quality degradation of the converted voice owing to insufficient parameterization accuracy. To avoid this issue, we have proposed a direct waveform modification technique based on spectral differential filtering and have successfully applied it to intra-gender singing VC (DIFFSVC) where excitation features are not necessary converted. Moreover, we have also applied it to cross-gender singing VC by implementing F_0 transformation with a constant rate such as one octave increase or decrease. On the other hand, it is not straightforward to apply the DIFFSVC framework to normal speech conversion because the F_0 transformation ratio widely varies depending on a combination of the source and target speakers. In this paper, we propose several F_0 transformation techniques for DIFFVC and compare their performance in terms of speech quality of the converted voice and conversion accuracy of speaker individuality. The experimental results demonstrate that the F_0 transformation technique based on waveform modification achieves the best performance among the proposed techniques.

Index Terms— voice conversion, speaker identity, F_0 transformation, Gaussian mixture model, direct waveform modification.

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 constraint, 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 constraint [1]. VC research was originally started to develop a speaker conversion technique to transform speaker identity of a source speaker's voice into that of a target speaker's voice 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 utterance pairs 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 develop a better conversion function, 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 sophisticated vocoding techniques, such as STRAIGHT [12] with mixed excitation [13]. Furthermore, a real-time conversion process has also been successfully implemented for the state-of-the-art GMM-based VC method [14]. However, 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 convert singing voices rather than normal voices, 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, 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 VC method. Moreover, to apply intra-gender DIFFSVC framework to cross-gender SVC as well, we have proposed an F_0 transformation technique with direct residual signal modification [20] based on time-scaling and resampling. In the proposed technique, waveform similarity-based overlap-add (WSOLA) [21] is applied to time-scale modification to avoid quality degradation caused by error of automatic pitch mark detection, which is often needed in other OLA methods, such as time domain pitch-synchronous overlap-add (TD-PSOLA) or linear prediction PSOLA (LP-PSOLA) [22]. We have found that the DIFFSVC framework can significantly improve speech quality of the singing converted voices compared to the conventional framework using the vocoding process.

Motivated by this success of the DIFFSVC framework in SVC, we have started to apply it to normal speech conversion (i.e., normal VC). However, we have found that it is not straightforward to apply the DIFFSVC framework to normal VC because more complicated F_0 transformation is necessary in VC compared to SVC; e.g., even if using a simple F_0 transformation method with a constant F_0 transformation ratio [23], such a ratio widely varies depending on a combination of the source and target speakers in normal VC although it can be fixed to double or half in cross-gender SVC. In the Voice Conversion Challenge 2016 (VCC 2016) [24], only a part of the DIFFSVC framework has been successfully applied to normal VC and our developed VC system still needs to use the vocoding processing for performing the F_0 transformation because the direct residual signal modification tends to cause quality degradation depending on a setting of the F_0 transformation ratio. Although our developed VC system (the NU-NAIST VC system) has been evaluated as one of the best systems achieving the best conversion accuracy on speaker identity and high speech quality in VCC 2016 [25], we have confirmed that its performance is still comparable to that of our conventional VC system [26].

In this paper, we propose several F_0 transformation techniques for normal VC with direct waveform modification with spectral differential (DIFFVC) to make it possible to widely accept various F_0 transformation ratios. The following F_0 transformation techniques using with or without the vocoder process are investigated: 1) DIFFVC with F_0 transformation using STRAIGHT vocoder (the NU-NAIST VC system for VCC 2016 [26]), 2) DIFFVC with F_0 transformation based on the direct residual signal modification using time-scaling and resampling [20], and 3) DIFFVC with F_0 transformation based on waveform modification using time-scaling and resampling. The experimental results demonstrate that the DIFFVC with F_0 transformation based on waveform modification using time-scaling and resampling achieves the highest speech quality and conversion accuracy when the F_0 transformation ratio nearly equals to 1.0 and its performance is still comparable to the others even if changing the F_0 transformation ratio from 0.5 to 2.0.

2. STATISTICAL VOICE CONVERSION WITH DIRECT WAVEFORM MODIFICATION WITH SPECTRAL DIFFERENTIAL (DIFFVC)

DIFFVC consists of a training process and a conversion process. In the training process, a joint probability density function of spectral features of a source speaker and the differential between the source and target speakers is modeled with a differential GMM, which is directly derived from a traditional GMM. As the spectral features of the source and target speaker, 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. As shown in [27], 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 a Gaussian 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. Then, the differential GMM is analytically derived from the trained GMM by transforming the parameters. 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 joint probability density function of the source and differential spectral features is shown as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{D}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(D)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XD)} \\ \boldsymbol{\Sigma}_m^{(DX)} & \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right) \quad (2)$$

$$\boldsymbol{\mu}_m^{(D)} = \boldsymbol{\mu}_m^{(Y)} - \boldsymbol{\mu}_m^{(X)} \quad (3)$$

$$\boldsymbol{\Sigma}_m^{(XD)} = {\boldsymbol{\Sigma}_m^{(DX)}}^\top = \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(XX)} \quad (4)$$

$$\boldsymbol{\Sigma}_m^{(DD)} = \boldsymbol{\Sigma}_m^{(XX)} + \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(YX)}. \quad (5)$$

In the conversion process, the converted spectral feature differential is estimated from the source speaker's spectral features based on the differential GMM in the same manner as maximum likelihood estimation of speech parameter trajectory with the GMM [10]. The voice timbre of the source speaker is converted into that of the target speaker by directly filtering the speech waveform of the input natural voice with the converted spectral feature differential. Time sequence vectors of the source features and the spectrum feature differential are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{D} = [\boldsymbol{D}_1^\top, \cdots, \boldsymbol{D}_T^\top]^\top$ 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 $\hat{\boldsymbol{d}} = [\hat{\boldsymbol{d}}_1^\top, \cdots, \hat{\boldsymbol{d}}_T^\top]^\top$ is determined as follows:

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d} \quad (6)$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [28].

3. INVESTIGATION OF THE F_0 TRANSFORMATION TECHNIQUES FOR DIFFVC FRAMEWORK

In this paper, we apply the following three F_0 transformation techniques to DIFFVC: 1) F_0 transformation based on STRAIGHT vocoder (which is also used in the NU-NAIST VC system for VCC 2016), 2) F_0 transformation based on residual signal modification using time-scaling and resampling, and 3) F_0 transformation based on waveform modification. Figure 1 describes the conversion processes of the DIFFVC methods using these techniques.

3.1. DIFFVC with F_0 transformation using STRAIGHT vocoder

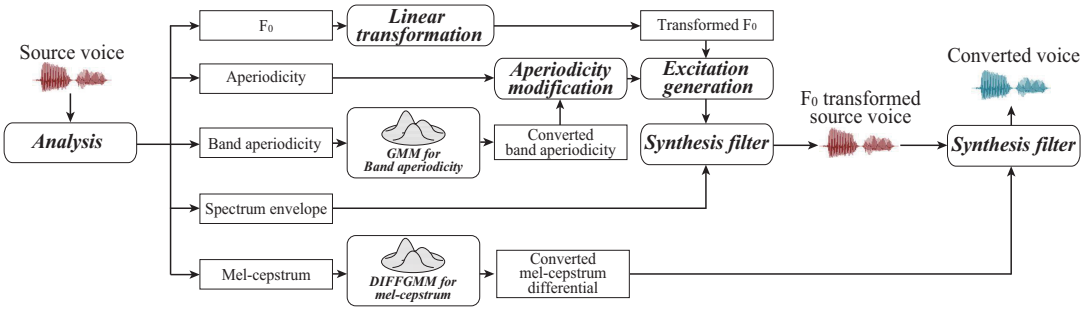
Figure 1 (a) describes the conversion process of the DIFFVC method with the F_0 transformation based on STRAIGHT vocoder. In this method, several acoustic features such as F_0, aperiodicity, and spectral envelope are extracted from the source voice using STRAIGHT analysis framework [29]. For the excitation conversion, F_0 is transformed based on global linear transformation in the same manner as the traditional VC method [10]. The aperiodic components at all frequency bins are shifted using band-averaged aperiodic differentials between the extracted and converted ones as a global bias term. Then, an F_0 transformed source voice is synthesized using full representation of STRAIGHT spectral envelope, the transformed F_0, and the transformed aperiodic components. Finally, spectral envelope of the F_0 transformed source voice is converted using the converted mel-cepstrum differentials with DIFFGMM in the same manner as the DIFFSVC.

This method is capable of converting the excitation parameters including not only F_0 but also aperiodic components as accurately as in the conventional VC. Therefore, it is expected that the conversion accuracy of speaker identity is almost equivalent to that of the conventional VC. On the other hand, this method ruins the advantage of the DIFFVC method, i.e., achievement of a high-quality converted voice by avoiding the vocoding process. Consequently, this method significantly suffers from quality degradation of the converted voice caused by F_0 extraction errors, unvoiced/voiced decision errors, lack of natural phase components, and so on.

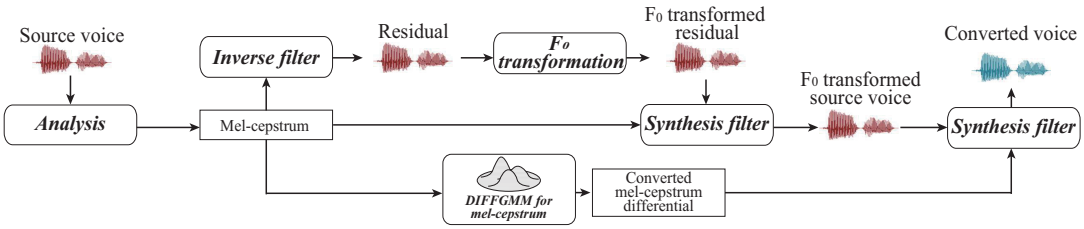
3.2. DIFFVC with F_0 transformation using residual signal modification

Figure 1 (b) describes the conversion process of the DIFFVC method with F_0 transformation based on residual signal modification. In this method, the F_0 transformation is carried out by directly modifying the residual signal. For the excitation conversion, the residual signal composed of harmonic and aperiodic components is extracted from the source voice with inverse filtering based on the extracted mel-cepstrum. Then, the time-scaling with WSOLA and resampling is performed on the residual signal in order to transform F_0. For instance, if F_0 is transformed to higher, the residual signal is expanded to make its duration longer, followed by using down-sampling to restore the length of the residual signal. If F_0 is transformed to lower, the residual signal is shrunk to make its duration shorter, followed by using up-sampling to restore its length. We further need to perform an additional process when decreasing F_0, making high frequency components of the transformed residual signal vanish. To reconstruct these vanished frequency components, they are generated using a noise excitation signal because the high frequency components of a speech signal tend to be less periodic and be well modeled with noise components. The F_0 transformed source voice is generated by filtering the resulting residual signal again using the extracted mel-cepstrum. Finally, spectral envelope of the F_0 transformed source voice is converted using the converted mel-cepstrum differentials with DIFFGMM in the same manner as the DIFFSVC. Note that we set the F_0 transformation ratio to a constant value for each speaker pair.

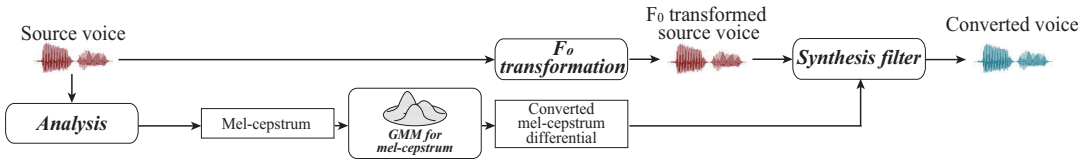
In this technique, a part of natural phase components of the source voice is well preserved because the F_0 transformation is performed by directly modifying the residual signal without the vocoding process. Moreover, this technique makes it possible to freely control the F_0 transformation ratio without changing DIFFGMM for the spectral differential conversion because the original spectral envelope is also preserved through the F_0 transformation. On the other hand, it is



(a) DIFFVC with F_0 transformation using STRAIGHT vocoder



(b) DIFFVC with F_0 transformation using residual signal modification



(c) DIFFVC with F_0 transformation using waveform modification

Fig. 1. Conversion process of several DIFFVC techniques.

possible to cause speech quality degradation of the converted voice due to some essentially difficult processes, e.g., the difficulty of extracting the residual signal by perfectly removing the effect of spectral envelope.

3.3. DIFFVC with F_0 transformation using waveform modification

Figure 1 (c) describes the conversion process of the DIFFVC method with the F_0 transformation using waveform modification. In this technique, the F_0 transformation using WSOLA and resampling based on linear interpolation is directly applied to an original waveform of the source voice. Because this direct waveform modification causes frequency warping, spectral envelope also changes according to the F_0 transformation ratio. Therefore, we need to use DIFFGMM capable of converting such a frequency warped source voice. We train the joint GMM using the F_0 transformed source voices and the natural target voices. For spectral conversion, the converted voice is generated by filtering the F_0 transformed source voice with converted mel-cepstrum differential determined with DIFFGMM derived from the corresponding joint GMM. The F_0 transformation ratio is set to a constant value for each speaker pair. Note that this F_0 transformation doesn't cause any problems even when decreasing F_0 because the high frequency components are generated with aliasing caused by the linear interpolation and the resulting spectral envelope is modeled with the joint GMM and also DIFFGMM.

In this technique, there is no approximation error caused by the vocoding process and the other processes, such as inverse filtering. Therefore, it is expected that this method achieves high-quality of the converted voice. Moreover, this method is based on quite simple processes, and therefore, it is easy to implement it to the real-time VC system [14]. On the other hand, we need to separately train the joint GMM for each different setting of the F_0 transformation ratio because spectral envelope of the F_0 transformed source voice depends on the F_0 transformation ratio.

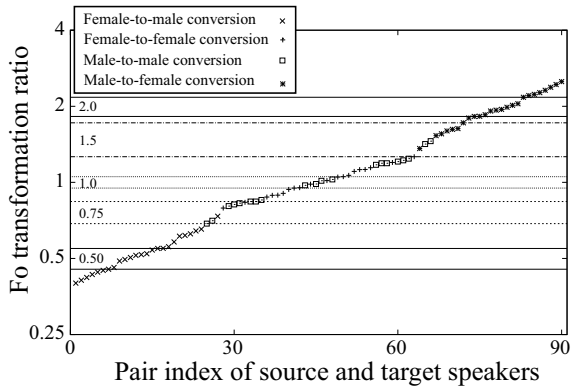


Fig. 2. F_0 transformation ratios between source and target speakers.

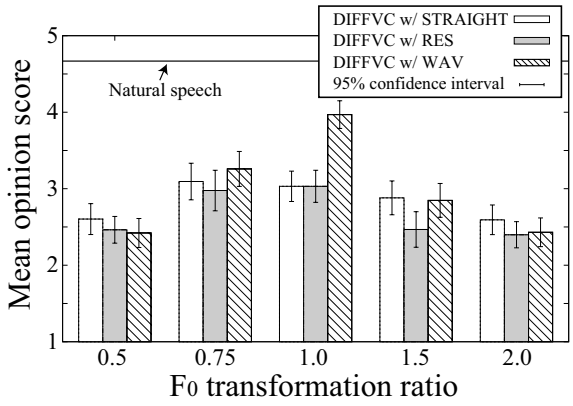


Fig. 3. Sound quality of converted voice.

4. EXPERIMENTAL EVALUATION

In this section, we evaluate performance of the following DIFFVC methods using different F_0 transformation techniques:

- DIFFVC w/ STRAIGHT: The NU-NAIST VC system submitted to the VCC 2016 [26] described in Sect. 3.1,
- DIFFVC w/ RES: The DIFFVC method with F_0 transformation based on the residual signal modification [20] described in Sect. 3.2,
- DIFFVC w/ WAV: The DIFFVC method with F_0 transformation based on the waveform modification described in Sect. 3.3.

4.1. Experimental conditions

We evaluated speech quality and speaker identity of the converted voices to compare performance of the different F_0 transformation techniques in both intra-gender and cross-gender conversions tasks. We used the English speech database used in the VCC 2016. The number of evaluation speakers was 10 including 5 females and 5 males, and the number of combinations of source and target speakers was 90. 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 envelope, 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 [30] was used as the synthesis filter. As the source excitation features, we used F_0 and aperiodic components extracted with STRAIGHT [29]. 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 investigated F_0 transformation ratios for all speaker possible pairs from 10 evaluation speakers (i.e., 45 speaker pairs in total) as shown in Figure 2, and selected 10 speaker pairs in each F_0 transformation ratio (0.5, 0.75, 1.0, 1.5, and 2.0) as the source and target speaker pairs. We used 162 sentences for training and the remaining 54 sentences were used for evaluation. The speaker-dependent GMMs were separately trained for the individual source and target speaker pairs. We performed MS-based postfilter for the converted mel-cepstrum differential. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64. The number of subjects was 8 and they were not native English speakers.

Two subjective evaluations were conducted. In the first test, we evaluated the speech quality of the converted voices using a mean opinion score (MOS). The natural and converted voice samples generated by three different DIFFVC 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 128.

In the second test, conversion accuracy in speaker identity was evaluated. In this test, F_0 transformation ratios were set to 0.5, 1.0, and 2.0. A natural voice sample of the target speaker was presented to the subjects first as a reference. Then, the converted voice samples generated by three different DIFFVC methods 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 90 sample pairs. They were allowed to replay each sample pair as many times as necessary.

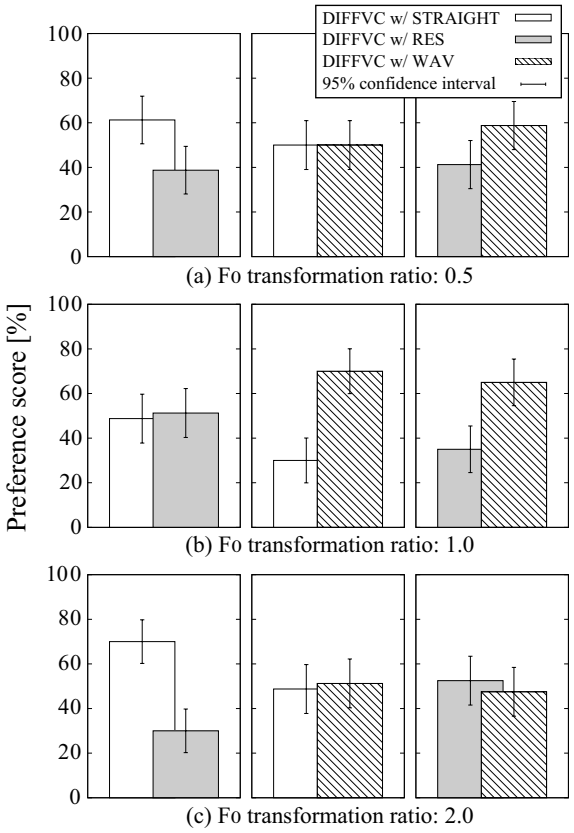


Fig. 4. Comparison of conversion accuracy on speaker identity.

4.2. Experimental results

Figure 3 indicates the results of the MOS test for speech quality. We can see a general tendency that speech quality degradation is caused by setting the F_0 transformation ratio to higher/lower values in all methods. When the F_0 transformation ratio is set to around 1.0, DIFFVC w/ WAV can achieve the highest speech quality. The speech quality achieved by DIFFVC w/ WAV rapidly degrades when setting the F_0 transformation ratio to higher or lower values than 1.0. On the other hand, DIFFVC w/ STRAIGHT and DIFFVC w/ RES tend to make such a quality degradation more gradually compared to DIFFVC w/ WAV. Nevertheless, the speech quality achieved by DIFFVC w/ WAV is still comparable to the other methods even if setting the F_0 transformation ratio to around 0.5 or 2.0. As for a comparison between DIFFVC w/ STRAIGHT and DIFFVC w/ RES, we can see that DIFFVC w/ STRAIGHT is slightly better than DIFFVC w/ RES when setting the F_0 transformation ratio to higher values (i.e., around 1.5 and 2.0). These results demonstrate that DIFFVC w/ WAV outperforms DIFFVC w/ STRAIGHT and DIFFVC w/ RES in terms of speech quality of the converted voices.

Figures 4 (a), (b) and (c) indicate the results of the preference test for speaker identity. We can see a tendency similar to that observed in the previous test on the converted speech quality; i.e., 1) DIFFVC w/ WAV yields better conversion accuracy for speaker identity than the other methods when setting the F_0 transformation ratio to around 1.0; 2) DIFFVC w/ WAV is still comparable to the other methods even when setting the F_0 transformation ratio to around 0.5 and 2.0; and 3) as for a comparison between DIFFVC w/ STRAIGHT and DIFFVC w/ RES, DIFFVC w/ STRAIGHT yields better conversion accuracy for speaker identity when setting the F_0 transformation ratio to around 0.5 and 2.0. Therefore, DIFFVC w/ WAV outperforms the other methods in terms of conversion accuracy for speaker identity as well.

These results suggest that DIFFVC w/ WAV is the best approach to implementing F_0 transformation to the DIFFVC framework in terms of both converted speech quality and conversion accuracy for speaker identity. Note that DIFFVC w/ WAV can also significantly reduce a computational cost in conversion.

5. CONCLUSIONS

In this paper, we have investigated the effectiveness of several F_0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential (DIFFVC), such as 1) F_0 transformation based on STRAIGHT vocoder (the NU-NAIST VC system for VCC 2016), 2) F_0 transformation based on residual signal modification, and 3) F_0 transformation based on waveform modification. We have compared their performance in terms of speech quality of the converted voices and conversion accuracy for speaker identity. The experimental results have demonstrated that 1) the F_0 transformation method based on waveform modification achieves significantly higher speech quality and conversion accuracy compared to the other methods when setting the F_0 transformation ratio close to 1.0, and 2) this method still achieves comparable performance to the other methods in terms of both speech quality and conversion accuracy for speaker identity even when setting the F_0 transformation ratio to higher or lower (e.g., around 2.0 or 0.5).

Thanks to the DIFFVC method using F_0 transformation based on waveform modification, voice conversion performance has been significantly improved for speaker pairs whose F_0 ranges are similar to each other but the performance is still comparable to the traditional conversion method using vocoder for other speaker pairs whose F_0 ranges are quite different from each other (e.g., in cross-gender conversion). In future work, we plan to improve performance of the F_0 transformation techniques in cross-gender conversion.

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] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "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] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467, Dec. 1990.

[23] D. T. Chappell and J. H. L. Hansen, "Speaker-specific pitch contour modeling and modification," *Proc. ICASSP*, vol. 2, pp. 885–888, May 1998.

[24] 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.

[25] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 evaluation results," *Proc. INTERSPEECH*, Sept. 2016.

[26] 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.

[27] Y. Kawakami, H. Banno, and F. Itakura, "GMM voice conversion of singing voice using vocal tract area function," *IEICE technical report. Speech (Japanese edition)*, vol. 110, no. 297, pp. 71–76, Nov. 2010.

[28] 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.

[29] 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.

[30] 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.