Generalization of Spectrum Differential based Direct Waveform Modification for Voice Conversion

Wen-Chin Huang¹, Yi-Chiao Wu¹, Kazuhiro Kobayashi¹, Yu-Huai Peng², Hsin-Te Hwang², Patrick Lumban Tobing¹, Yu Tsao², Hsin-Min Wang² Tomoki Toda¹,

¹ Nagoya University, Japan
² Academia Sinica, Taiwan
wen.chinhuang@sp.m.is.nagoya-u.ac.jp

Abstract
We present a modification to the spectrum differential based direct waveform modification for voice conversion (DIFFVC) so that it can be directly applied as a waveform generation module to voice conversion models. The recently proposed DIFFVC avoids the use of a vocoder, meanwhile preserves rich spectral details hence capable of generating high quality converted voice. To apply the DIFFVC framework, a model that can estimate the spectral differential from the F0 transformed input speech needs to be trained beforehand. This requirement imposes several constraints, including a limitation on the estimation model to parallel training and the need of extra training on each conversion pair, which make DIFFVC inflexible. Based on the above motivations, we propose a new DIFFVC framework based on an F0 transformation in the residual domain. By performing inverse filtering on the input signal followed by synthesis filtering on the F0 transformed residual signal using the converted spectral features directly, the spectral conversion model does not need to be retrained or capable of predicting the spectral differential. We describe several details that need to be taken care of under this modification, and by applying our proposed method to a non-parallel, variational autoencoder (VAE)-based spectral conversion model, we demonstrate that this framework can be generalized to any spectral conversion model, and experimental evaluations show that it can outperform a baseline framework whose waveform generation process is carried out by a vocoder.

Index Terms: voice conversion, direct waveform modification, F0 transformation, collapsed waveform detection

1. Introduction
Voice conversion (VC) aims to convert the speech from a source to that of a target without changing the linguistic content. Numerous approaches have been proposed, such as Gaussian mixture model (GMM)-based methods [1]-[2], deep neural network (DNN)-based methods [3]-[4], and exemplar-based methods [5]-[7]. While most VC researchers have focused on the conversion of spectral features, the waveform generation process, in fact, plays an important role in a VC system. Conventional VC frameworks employ parametric vocoders [8],[9],[10] as their synthesis module, which impose many overly simplified assumptions that discard the phase information and result in unnatural excitation signals, and thus cause a significant degradation in the quality of the converted speech.

In recent years, there are two mainstreams that try to improve the waveform generation module. One direction is to develop neural vocoders [11],[12],[13],[14],[15],[16],[17],[18],[19],[20], which are capable of reconstructing the phase and excitation information, and thus generate extremely natural sounding speech. The converted speech is then obtained by directly filtering the transformed signal with a trained differential GMM (DIFFGMM). Such a GMM model is termed DIFFGMM.

State-of-the-art VC systems have shown remarkable results by combining such neural waveform generation process with conversion models such as GMMs [20], or other methods based on recent DNN methods such as bidirectional long short term memory [21], autoencoders [22], sequence-to-sequence learning [23] or generative adversarial networks [24]. Nonetheless, neural vocoders usually require a large amount of training data and are computationally expensive.

Another approach which is simple, efficient yet free from any extra training data is the direct waveform modification framework based on spectrum differential, which was first applied to VC in [25] (DIFFVC). The DIFFVC framework first estimates a sequence of spectral differential from the F0 transformed signal with a trained differential GMM (DIFFGMM). The converted speech is then obtained by directly filtering the F0 transformed speech with the spectral differential. In the Voice Conversion Challenge (VCC) 2016 [26], a predecessor version [25] was evaluated as one of the best systems achieving the best conversion accuracy on speaker identity and high speech quality. It was then adopted as the baseline system of the VCC2018, and achieved the second highest sound quality for the same-gender speaker conversion pairs [27].

Despite the success of DIFFVC, one flaw is its inflexibility. First, in order to train a model that can estimate the spectral differential, parallel data is needed. This makes application of DIFFVC to non-parallel VC methods infeasible. Moreover, the differential estimation model in the state-of-the-art DIFFVC framework needs to be trained using the features extracted from the F0 transformed signal. In other words, for an arbitrary VC model that we wish to apply DIFFVC to, it requires retraining the model for each source-target conversion pair using the corresponding F0 transformed features. This limitation significantly increases the inflexibility of the framework.

To this end, our goal is to develop a flexible method which is free from parallel data or any extra training process so that one can directly apply the vocoder-free DIFFVC framework to an arbitrary VC model. Our proposed method is based on
the residual signal modification based F0 transformation implementation, which is similar to the method proposed in [28]. Our contributions are:

- We generalize the DIFFVC framework such that it is applicable to any VC model as long as it is able to perform spectral conversion from the normal source speech, regardless of how it was trained. In other words, the VC model need not to estimate the spectrum differential nor use the F0 transformed features as input, thus free from parallel data or extra retraining procedure.

- We propose several techniques to address some details that need to be taken care of when adopting the residual signal modification based F0 transformation, such as power compensation, collapsed speech detection and feature substitution.

As a proof of concept, we show the effectiveness of our proposed method by applying it to a variational autoencoder based VC model, which we will refer to as VAE-VC [29]. It was trained with non-parallel data, and takes normal, "un-F0-transformed" as input during conversion. Through subjective evaluations, we show that our proposed framework can bring a significant performance boost in some conversion gender pairs, compared to a baseline system which employs a conventional parametric vocoder, WORLD [10], in terms of speech naturalness and conversion accuracy. Note that our goal is to demonstrate the ability of our method to generalize to any VC model, so it is not restricted to VAE-VC.

2. Spectrum Differential based Direct Waveform Modification for Voice Conversion (DIFFVC)

2.1. DIFFVC based on DIFFGMM

DIFFVC is a conversion framework (not restricted to VC but also other applications like singing VC) that does not employ a parametric vocoder as the waveform generation module [25, 27, 28, 30]. In this section we describe the DIFFVC framework in [30]. In the offline stage, instead of training a VC model capable of mapping source features to target features, DIFFVC requires to model the spectral differential between the source and target. The differential is obtained by subtracting the target feature with the source feature. To realize this, a joint density GMM [2] was first trained using the joint vectors of source and target features in the parallel dataset. Then, the DIFFGMM can be analytically derived by transforming the trained model parameters. Note that to train such a model, a parallel dataset is required, which is a fundamental limitation of this method.

The conversion process is depicted in Figure 1. First, an F0 transformation process is performed to obtain the F0 transformed waveform. Then, the mcp differential is estimated via maximum likelihood speech parameter estimation [2] based on the DIFFGMM, optionally enhanced by a global variance (GV) postfilter [21]. Finally, the converted speech is obtained by directly filtering the F0 transformed signal with the differential.

2.2. F0 transformation techniques

In this subsection, we describe two F0 transformation techniques proposed in [30]. Note that in both techniques, we assume the F0 transformation ratio is time-invariant and set it to a constant value, which is calculated using the training data of the source and target speakers.

2.2.1. F0 transformation by residual signal modification

In this method, the F0 transformation is performed by modifying the residual signal. According to the source-filter model, given a series of waveform signals $S(z)$ and the spectral envelope $H(z)$, the ideal excitation signal (or residual) $E(z)$ can first be computed by filtering $S(z)$ with the inverse filter $H(z)$:

$$E(z) = \frac{1}{H(z)} S(z),$$

where the spectral features extracted from $S(z)$ are employed as coefficients of the time-invariant filter $H(z)$. Then, a WSOLA [32] and resampling process can be performed on the residual signal in order to transform F0. Specifically, the residual signal is shrunk then up-sampled if the F0 transformation ratio is smaller than 1 and, conversely, expanded then down-sampled if the ratio larger than 1. Finally, the F0 transformed speech is restored by filtering the modified residual signal using the same $H(z)$.

As discussed in [28], when the ratio is smaller than 1, the high frequency components of the transformed residual signal vanish, thus need to be reconstructed. [28] proposed to reconstruct them by adding a high-pass filtered noise excitation signal to the F0 transformed residual signal, as they claimed that the high frequency components of a speech signal tend to be less periodic and well modeled with noise components.

This technique has several flaws. First, the WSOLA process needs to be applied to the residual signal, which presents a challenge for WSOLA since it was originally designed for normal speech. In addition, the generation of the high-pass filtered noise excitation signal is time-consuming, and sometimes causes discontinuities or asynchronous conditions with low frequency residual components. Finally, by using this technique, it is assumed that the residual spectrum is perfect, i.e., its spectral tilt is totally flat, which is not in practice due to imperfect inverse filtering. As a result, this method does not work well as reported in [28]. Our proposed method addresses these issues, as described in Section 3.

2.2.2. F0 transformation by direct waveform modification

In this method, rather than performing the complicated chain process (inverse-transform-synthesis) described in Section 2.2, the F0 transformation process is directly applied to the source waveform. The advantages of this process are that first, no approximation errors caused by the processes like inverse filtering are introduced. It can be therefore expected that a high-quality transformed signal can be obtained. Moreover, since this method is simpler, it is more likely to be embedded to real-time VC systems.

However, this direct waveform modification causes a frequency warping issue, so the DIFFGMM needs to be trained with features extracted from the F0 transformed and natural target speech. In other words, a separate DIFFGMM for each F0 transformation ratio needs to be trained because the spectral envelope of the F0 transformed signal depends on the F0 transformation ratio.

3. Proposed Method based on Residual Transformation

Our goal is to extend the vocoder-free DIFFVC framework to any arbitrary VC model, which only knows how to convert normal source features to target features. To impose as few constraints as possible, we only demand the VC model to estimate
the converted spectral features given source spectral features extracted from a normal source speech, regardless of whether a parallel training dataset is available. In this section, we describe our proposed method, as depicted in Figure 2.

### 3.1. Interpretation of F0 transformation based on residual transformation

The key idea of our proposed method resort back to the motivation of using the spectrum differential, which is based on the source-filter model. To obtain the ideal time-domain, spectrally converted speech signal \( y[n] \), consider the following decomposition:

\[
\begin{align*}
\hat{y}[n] &= h_{f \cdot}[n] * x[n] \\
\approx h_{b}[n] * h_{2}[n]^{-1} * x[n],
\end{align*}
\]

where \( x[n] \) is the input time-domain source speech signal. Eq. (2) is realized by filtering \( x[n] \) with a time-variant filter \( h_{f \cdot}[n] \), whose coefficients (the spectrum differential) are estimated by a VC model. This is exactly the DIFFVC method in Section 2. We propose a different approach to reconstruct the high frequency components of the transformed residual when the F0 ratio is smaller than 1. After performing WSOLA, we insert zeroes between every two samples in \( \text{res}^{(\omega x)} \), which is called upsampling by zero stuffing or spectral folding [33]. This process doubles the cutoff frequency and generates a symmetric spectral envelope with respect to the original cutoff frequency. Such technique has been applied to various speech processing fields, such as bandwidth extension [24]. Then, the resampling process discards the spectral envelope with frequency larger than cutoff frequency/F0 ratio. As a result, the missing frequency components are generated by copying a mirrored part of that of \( \text{res}^{(\omega x)} \). Since the reconstructed component is merely a reversed copy, it is expected to be more continuous. Another advantage of this approach is that it is much simpler to implement.

### 3.2. An alternative of high frequency component reconstruction

We propose a different approach to reconstruct the high frequency components of the transformed residual when the F0 ratio is smaller than 1. After performing WSOLA, we insert zeroes between every two samples in \( \text{res}^{(\omega x)} \), which is called upsampling by zero stuffing or spectral folding [33]. This process doubles the cutoff frequency and generates a symmetric spectral envelope with respect to the original cutoff frequency. Such technique has been applied to various speech processing fields, such as bandwidth extension [24]. Then, the resampling process discards the spectral envelope with frequency larger than cutoff frequency/F0 ratio. As a result, the missing frequency components are generated by copying a mirrored part of that of \( \text{res}^{(\omega x)} \). Since the reconstructed component is merely a reversed copy, it is expected to be more continuous. Another advantage of this approach is that it is much simpler to implement. We formulate the aforementioned changes into the following decomposition:

\[
\begin{align*}
x'[n] &= f_w(x[n]) \\
y[n] &= h_{b}[n] * f_r(h_{\nu}[n]^{-1} * x'[n]),
\end{align*}
\]

where we decompose \( f(\cdot) \) into two functions \( f_w(\cdot) \) and \( f_r(\cdot) \), the former being the WSOLA-based duration conversion function (step 1) and the latter being the resampling function (step 4b), preceded by the spectral folding technique (step 4a) if needed. One trick used here is that, since \( \text{mcp} \) extraction (step 2) is time-consuming, we try to utilize \( \text{mcp}^{(\omega x)} \) as much as possible, by using it as the input feature to the conversion process (step 6a). This can be accomplished by applying linear interpolation to \( \text{mcp}^{(\omega x)} \) to restore the original time length (step 5).
3.3. Collapsed waveform detection and feature substitution

In our initial experiments, we often observed collapsed waveform segments in $\text{sig}^{(y)}$. This is a combined result of the frequency axis warping effect caused by resampling, as pointed out in [28][30], and the imperfect residual signal obtained from filtering. During conversion, a GV postfilter could further amplify this effect, leading to greater instability.

As a remedy, we replace the postfiltered features $\text{mcp}^{(y)}_{\text{GV}}$ that cause collapsed segments with the corresponding ones without postfiltering $\text{mcp}^{(y)}$. To detect collapsed waveform intervals, we adopt a modified signal envelope extraction method proposed in [35]. Specifically, the waveform signal is first passed through a Hilbert transform, which is often used to extract the signal envelope. Then, it is divided into non-overlapping slots with a fixed window, and the maximum value of each slot is used to replace every value in that slot. Finally, a low-pass filter is used to smoothen the curve. Figure 3 illustrates the calculated envelopes.

The complete detection and substitution procedures can be summarized as follows. First, we use the WORLD vocoder to generate $\text{sig}^{(y)}_{\text{GV}}$ from $\text{mcp}^{(y)}_{\text{GV}}$ as reference (step 7a). Then, the above process is applied to both $\text{sig}^{(y)}$ and $\text{sig}^{(y)}_{\text{GV}}$ to obtain the corresponding envelopes $\text{env}_{\text{GV}}$ and $\text{env}_{\text{GV}}$ (step 8a). The collapsed interval is then detected by setting an empirically set threshold to the difference between the envelopes. Finally, feature substitution is performed to obtain $\text{mcp}^{(y)}_{\text{SU}}$ (step 8b), which can be used as the coefficients of the final synthesis filtering process.

3.4. Power compensation

It is observed that the power of the signal changes after resampling. We perform two power compensation processes. In step 4c, the signal power is compensated according to the F0 transformation ratio:

$$\text{res}^{(y)} = \text{res}^{(y)} \cdot \sqrt{\frac{1}{\text{F0 ratio}}},$$

while in step 9b, the signal power is normalized to be the same as the input speech:

$$\text{sig}_{\text{SU}}^{(y)} = \text{sig}_{\text{SU}}^{(y)} \cdot \sqrt{\frac{\sum_n \text{sig}^{(y)}_{\text{SU}}[n]^2}{\sum_n \text{sig}^{(y)}_{\text{SU}}[n]^2}}.$$
4.3. Error analysis

We demonstrate why our proposed method stayed comparable in terms of the naturalness with the two spectrogram illustrations shown in Figure 4. Note that although the two illustrated samples come from conversion pairs that are not in the subjective evaluation set, these are indeed problems that exist in our proposed system since they are also occasionally found in the evaluation set. First, the wrapped frequency axes in the F0 transformed residual are very sensitive to imperfect converted spectral estimation. In our internal evaluation, we found that the husky characteristic of male speakers is hard to model due to the disadvantage condition of non-parallel training. Thus, the interaction of the imperfectly estimated converted feature and the F0 transformed residual signal tends to make the converted speech unstable. It could be observed in the bottom left of Figure 4 that the low frequency components tend to be very noisy, causing a seriously deteriorated voice. Informal comments from subjective test participants also confirmed this result.

Another problem worth noticing is that sometimes even by substituting features, the collapsed waveform problem is not completely avoided. As shown in the bottom right of Figure 4, a short collapsed segment still exists even after the substitution process. This is a fundamental issue the DIFFVC is faced with when using a WSOLA plus resampling based F0 transformation process, since this problem exists in not only our proposed framework but also the original DIFFVC framework [27].

5. Conclusions and Future Work

In this paper, we introduced a generalization of the DIFFVC framework to make it applicable to general VC models. The proposed method is based on an F0 transformation in the residual domain, so that synthesis filtering is performed directly using the converted spectral features, thus removing the need for the conversion model to be able to predict the spectral differential, making the entire process free of parallel training data. We also introduced several techniques used in this framework, including 1) an alternative for high frequency component reconstruction based on zero stuffing, 2) collapsed waveform detection and corresponding feature substitution, and 3) power compensation due to resampling. Experimental results confirmed that when applied to a non-parallel VAE-based VC model, our method outperformed the counterpart that used a conventional vocoder in terms of conversion accuracy, yet the naturalness was on par.

The investigation of the effectiveness of individual components proposed in this work, as well as a more in-depth analysis of the experimental results will be of top priority for future work. We also plan to apply our framework to other non-parallel VC models to further validate the effectiveness. Developing a frequency-wrapping robust spectral feature extractor may help solve the various issues discussed in Section 4.3 which will be another important future work.

Acknowledgements: This work was partly supported by JST, PRESTO Grant Number JPMJPR1657 and JSPS KAKENHI Grant Number 17H01763, as well as the MOST-Taiwan Grants 107-2221-E-001-008-MY3 and 108-2634-F-001-004.

6. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, Mar 1998.

[2] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, Nov 2007.
[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectrogram parsing using artificial neural networks for voice conversion,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 954–964, July 2010.

[4] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai, “Voice conversion using deep neural networks with layer-wise generative training,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1859–1872, Dec 2014.

[5] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” in Proc. SLT, 2012, pp. 313–317.

[6] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, “Exemplar-based sparse representation with residual compensation for voice conversion,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1506–1521, Oct 2014.

[7] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, “Locally linear embedding for exemplar-based spectral conversion,” in Proc. Interspeech, 2016, pp. 1652–1656.

[8] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” The Journal of the Acoustical Society of America, vol. 50, no. 2B, pp. 637–655, 1971.

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Re-structuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, no. 3, pp. 187–207, 1999.

[10] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” IEICE Transactions on Information and Systems, vol. 99, pp. 1877–1884, 2016.

[11] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent wavenet vocoder,” in Proc. Interspeech, 2017, pp. 1118–1122.

[12] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, “An investigation of multi-speaker training for wavenet vocoder,” in IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Dec 2017, pp. 712–718.

[13] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stinmburg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” arXiv preprint arXiv:1802.08435, 2018.

[14] S. Mehr, K. Kumar, I. Gutjahr, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, “Samplernn: An unconditional end-to-end neural audio generation model,” arXiv preprint arXiv:1612.07837, 2016.

[15] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, “Fitnet: A real-time speaker-dependent neural vocoder,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 2551–2555.

[16] J.-M. Valin and J. Skoglund, “Lpncnet: Improving neural speech synthesis through linear prediction,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5891–5895.

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

[18] R. Prenger, R. Valle, and B. Catanzaro, “Waveflow: A flow-based generative network for speech synthesis,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3617–3621.

[19] X. Wang, S. Takaki, and J. Yamagishi, “Neural source-filter-based waveform model for statistical parametric speech synthesis,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5916–5920.

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

[21] K. Chen, B. Chen, J. Lai, and K. Yu, “High-quality voice conversion using spectrogram-based wavenet vocoder,” in Proc. Interspeech, 2018, pp. 1993–1997.

[22] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “Zero-shot voice style transfer with only autoencoder loss,” arXiv preprint arXiv:1905.05879, 2019.

[23] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, “Sequence-to-sequence acoustic modeling for voice conversion,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 27, no. 3, pp. 631–644, 2019.

[24] B. Sisman, M. Zhang, S. Sakit, H. Li, and S. Nakamura, “Adaptive wavenet vocoder for residual compensation in gan-based voice conversion,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 282–289.

[25] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, “The nu-nast voice conversion system for the voice conversion challenge 2016,” in Interspeech, 2016, pp. 1667–1671.

[26] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The voice conversion challenge 2016,” in Interspeech 26., 2016, pp. 1632–1636.

[27] K. Kobayashi and T. Toda, “Sprocker: Open-source voice conversion software,” in Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, 2018, pp. 203–210.

[28] K. Kobayashi, T. Toda, and S. Nakamura, “Implementation of f0 transformation for statistical singing voice conversion based on direct waveform modification,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5670–5674, 2016.

[29] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from non-parallel corpora using variational auto-encoder,” in Proc. APISPA ASC, 2016, pp. 1–6.

[30] K. Kobayashi, T. Toda, and S. Nakamura, “F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential,” 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 693–700, 2016.

[31] H. Siln, E. Hel, J. Nurminen, and M. Gabbouj, “Ways to implement global variance in statistical speech synthesis,” in Proc. Interspeech, 2012, pp. 1436–1439.

[32] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (wosla) for high quality time-scale modification of speech,” in ICASSP, 1993.

[33] J. Makhoul and M. Berouti, “High-frequency regeneration in speech coding systems,” in ICASSP 79. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 4, 1979, pp. 428–431.

[34] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, “Speech enhancement via frequency bandwidth extension using line spectral frequencies,” in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No. 01CH37221), vol. 1. IEEE, 2001, pp. 665–668.

[35] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda, “Collapsed speech segment detection and suppression for wavenet vocoder,” in Proc. Interspeech 2018, 2018, pp. 1988–1992.

[36] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” in Proc. Odyssey, 2018, pp. 195–202.

[37] W.-C. Huang, Y.-C. Wu, C.-C. Lo, P. Lomban Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, “Investigation of F0 conditioning and Fully Convolutional Networks in Variational Autoencoder based Voice Conversion,” arXiv e-prints, May 2019.

[38] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Mel-generalized cepstral analysis - a unified approach to speech spectral estimation,” in ICSLP, 1994.