INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Robustness of Statistical Voice Conversion based on Direct Waveform Modification against Background Sounds

Yusuke Kurita[1], *Kazuhiro Kobayashi*[1], *Kazuya Takeda*[1], *Tomoki Toda*[1]

[1]Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan

kurita.yusuke@g.sp.m.is.nagoya-u.ac.jp, kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp,
kazuya.takeda@nagoya-u.jp, tomoki@ics.nagoya-u.ac.jp

Abstract

This paper presents an investigation of the robustness of statistical voice conversion (VC) under noisy environments. To develop various VC applications, such as augmented vocal production and augmented speech production, it is necessary to handle noisy input speech because some background sounds, such as external noise and an accompanying sound, usually exist in a real environment. In this paper, we investigate an impact of the background sounds on the conversion performance in singing voice conversion focusing on two main VC frameworks, 1) vocoder-based VC and 2) vocoder-free VC based on direct waveform modification. We conduct a subjective evaluation on the converted singing voice quality under noisy conditions and reveal that the vocoder-free VC is more robust against background sounds compared with the vocoder-based VC. We also analyze the robustness of statistical VC and show that a kurtosis ratio of power spectral components before and after conversion is useful as an objective metric to evaluate it without using any target reference signals.

Index Terms: Statistical voice conversion, background sounds, vocoder, direct waveform modification, kurtosis ratio

1. Introduction

Voice conversion (VC) is a technique that converts speaker characteristics of a source speaker into those of a target speaker while maintaining linguistic contents of the input speech [1, 2]. There are various potential applications based on VC in not only speech processing but also singing voice processing. For instance, although singing voice characteristics produced by each singer are limited by physical constraints of a vocal production mechanism, they are flexibly converted by a VC method [3, 4]. Moreover, it is also possible to convert the singing voice in real-time by using its real-time implementation [5]. Therefore, VC makes it possible for individual singers to sing songs with their desired voice characteristics beyond their physical limitations. The same VC framework can also be used for augmenting speech production function [6].

In a traditional VC framework, the input speech is analyzed to extract some acoustic features, such as F_0, spectral envelope, and aperiodic components [7], and then, they are converted using a previously trained conversion model. Finally, a converted speech waveform is generated from the converted acoustic features using vocoder [8]. Therefore, various factors involved in vocoder processing, such as F_0 estimation error, voiced/unvoiced decision error, spectral envelope modeling error, and source excitation modeling error, directly affect sound quality of the converted speech. Even if using well-known high-quality vocoding methods [7, 9–11], these issues are difficult to completely address. Although the development of high-fidelity vocoder, such as neural vocoder, has attracted attention [12, 13], it is still hard to achieve computationally efficient real-time vocoding process while maintaining sufficiently high sound quality of vocoded speech.

In some VC applications, it is not necessary to convert all of the acoustic features. For instance, in singing voice conversion (SVC), the same melody line is often sung by source and target singers of the same gender, and therefore, the requirement of the source excitation conversion is greatly reduced. For such an application, a conversion method based on modification of log spectral differentials (DIFFVC) has been proposed [14] as a method for converting only a spectral envelope without using a vocoder. In DIFFVC, a log spectral feature sequence, such as a mel-cepstrum sequence, analyzed from the input speech is converted into a sequence of the feature differentials between the target speech and the input speech. Then, a time-varying filter corresponding to it is directly applied to the input speech waveform. Therefore, the vocoder-based waveform generation process causing sound quality degradation can be avoided.

Towards the practical use of the VC applications, it is inevitable that some background sounds, such as external noise or an accompanying sound, are superimposed on the input speech in conversion. Such a noisy input speech easily causes the degradation of conversion performance. To address this issue, there have been studied VC frameworks using noise suppression [15], where noise suppression is applied to the noisy input speech before conversion. This technique is helpful to improve the conversion performance under noisy conditions but its performance improvement is limited because the processing errors or artifacts are usually caused in noise suppression and they also cause adverse effects on the conversion performance. Although the noise suppression performance continues to improve year by year [16–18], it is still difficult to achieve complete noise suppression processing.

Another approach to handle the noisy input speech is to develop a VC method robust against the background sounds. If the noisy input speech could be directly converted into noisy target speech, it would be possible to achieve conversion process while maintaining information of background sounds. Such a conversion process will be effective in various VC applications, e.g., SVC in an accompanying sound where the accompanying sound superimposed on the converted singing voice will not cause any practical issues or VC in telecommunication under noisy conditions where it will be informative to convey not only the converted voice but also background sounds. Moreover, it is expected that such a noisy robust VC processing is also helpful for converting the processed input speech suffering from the noise suppression errors and artifacts.

In this paper, we investigate the effect of background sounds on the conversion performance in two main SVC frameworks, the vocoder-based VC [2] and the vocoder-free VC based on DIFFVC [14]. The subjective evaluation results show

http://dx.doi.org/10.21437/Interspeech.2019-2206

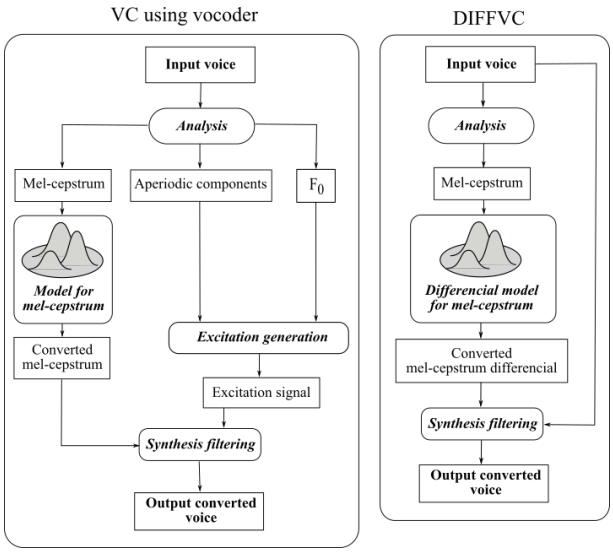


Figure 1: *Conversion processing of vocoder-based VC and vocoder-free VC with DIFFVC*

that the vocoder-free VC is more robust against the background sounds than the vocoder-based VC. We also analyze the robustness of these VC methods and investigate an objective metric to evaluate it without using any target reference signals.

2. Statistical Voice Conversion

Statistical VC consists of training and conversion processes. In the training process, using a parallel data consisting of an utterance pair of the same sentence uttered by the source speaker and the target speaker, a conversion model from the source speech feature to the target speech feature is trained. In the conversion process, an arbitrary utterance of the source speaker is converted into that of the target speaker by using the conversion model. In SVC, a singing voice is used instead of a normal voice in both processes. Figure 1 shows the conversion process of two main SVC frameworks, vocoder-based VC and vocoder-free VC with DIFFVC, in SVC without F_0 conversion.

2.1. Vocoder-based VC

As the acoustic features of the source and target singers, we use 2D-dimensional joint static and dynamic feature vectors of the source mel-cepstrum $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and the target mel-cepstrum $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$. In the training process, the conversion model from $\boldsymbol{X}_t$ to $\boldsymbol{Y}_t$, e.g., the conditional probability density function $P(\boldsymbol{Y}_t|\boldsymbol{X}_t)$, is trained. In the conversion process, the source mel-cepstrum sequence is converted to the target mel-cepstrum sequence using the conversion model. Then, the source excitation signal is generated with a mixed excitation model of vocoder from F_0 and the aperiodic components extracted from the input source speech. Finally, the generated excitation signal is filtered with MLSA filter [19] designed with the converted mel-cepstrum sequence.

2.2. Vocoder-free VC based on DIFFVC

In the training process of DIFFVC [14], a conversion model from $\boldsymbol{X}_t$ to $\boldsymbol{D}_t = \boldsymbol{Y}_t - \boldsymbol{X}_t$, e.g., $P(\boldsymbol{D}_t|\boldsymbol{X}_t)$, is trained. In the conversion process, the source mel-cepstrum sequence is converted into a sequence of the mel-cepstrum differentials. Then, the input source speech waveform is directly filtered with MLSA filter designed with the sequence of the converted mel-cepstrum differentials. In this time-variant filtering process, a spectral envelope sequence of the input source speech waveform is converted into that of the target speech while basically maintaining the input source excitation signal.

2.3. Conversion model

In this paper, a GMM-based conversion model is used. One of the biggest merits of using the GMM-based conversion model is that the same GMM can be shared between the vocoder-based VC and the vocoder-free VC with DIFFVC because the conversion model for DIFFVC $P(\boldsymbol{D}_t|\boldsymbol{X}_t)$ is analytically derived from the conversion model for VC $P(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ by variable transformation from $\boldsymbol{Y}_t$ to $\boldsymbol{D}_t$ [14]. Therefore, the effects of using the different conversion models can be minimized in a comparison between the vocoder-based VC and the vocoder-free VC with DIFFVC. Moreover, it has been reported that the GMM-based DIFFVC framework can achieve sufficiently high conversion performance, which is comparable with the state-of-the-art performance when F_0 transformation is not performed [14, 20, 21]. In this paper, we focus on SVC in the same gender conversion, and therefore, we can develop a high-quality VC system using the DIFFVC framework without F_0 transformation.

3. Investigation of Robustness of VC against Background Sounds

A comparison between the vocoder-based VC and the vocoder-free VC with DIFFVC has been well investigated under clean conditions [14]. On the other hand, their comparison under noisy conditions have never been investigated. To investigate the robustness of VC against background sounds, we performed a comparison between the vocoder-based VC and the vocoder-free VC with DIFFVC under noisy conditions in SVC of the same gender conversion.

3.1. Experimental conditions

We used a singing voice dataset of Japanese songs. 50 phrases randomly selected from it were used as the training data. The same-gender conversion were performed in 12 singer-pairs using six singers (3 females and 3 males) included in the dataset. The GMM-based conversion model was trained in each singer-pair. Clean singing voices were used in training. On the other hand, noisy singing voices were used in conversion, where they were generated by superimposing an environmental noise or an accompaniment sound on the clean singing voices while varying the signal-to-noise ratio (SNR). The same GMM was basically used in both the vocoder-based VC and the vocoder-free VC with DIFFVC.

The sampling rate was 44100 Hz. The bit rate was 16 bits. The shift length was 5 ms. The number of mixture components of the GMM was 32. As the spectral feature, we used the 1st through 40th mel-cepstrum coefficients obtained by WORLD analysis [22]. MLSA filter [19] was used in time-varying filtering based on a given mel-cepstrum sequence. F_0 and aperiodic components were not converted. WORLD [22] was used as vocoder, and sprocket [21] was used as VC software.

Opinion tests were conducted to evaluate naturalness of the converted singing voices.[1] The number of listeners was 11. The

[1] singing voice samples: `https://drive.google.com/drive/folders/1mnN3c5MlJgTWyDRM04ZKiS7_`

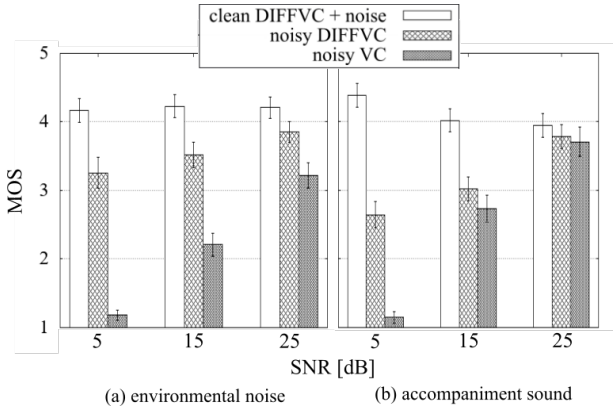


Figure 2: *Results of subjective evaluation on naturalness: (a) under environmental noise, (b) under accompanying sound. Error bars show 95 % confidence intervals.*

SNR of the input source singing voice was set to 5, 15 and 25 dB. The following singing voice samples were created for each noise level:

- clean DIFFVC + noise: singing voices generated by superimposing background sounds after converting the clean singing voice by the vocoder-free VC with DIFFVC,
- noisy DIFFVC: singing voices converted from the noisy singing voices by using the vocoder-free VC with DIFFVC,
- noisy VC: singing voices converted from the noisy singing voices by using the vocoder-based VC,

where clean DIFFVC + noise showed an ideal conversion result. These singing voice samples were presented to each listener in random order. Naturalness of each singing voice sample was evaluated with 5-scaled opinion scores (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Each listener evaluated 144 singing voice samples.

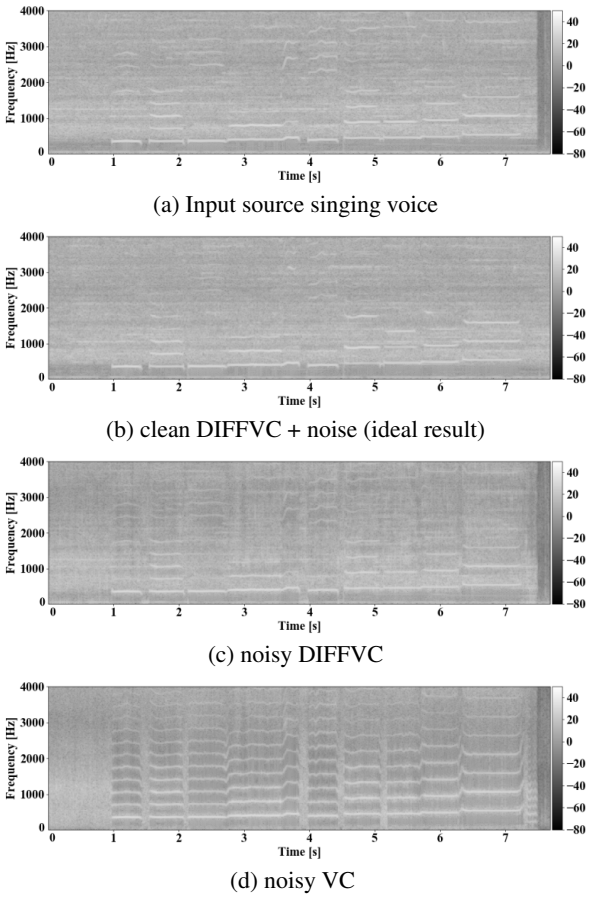
3.2. Experimental results

Figure 2 shows the results of subjective evaluation, where Fig. 2 (a) shows the result under the environmental noise, and Fig. 2 (b) shows the result under the accompanying sound. Under both noisy conditions, the naturalness of converted singing voices greatly decreases as the SNR decreases in the vocoder-based VC. In the vocoder-free VC with DIFFVC, we can also observe the naturalness degradation caused by the decrease of SNR. However, its degradation is much smaller compared to that in the vocoder-based VC. It can be seen from Fig. 2 (a) that the vocoder-free VC at 5 dB of the SNR is comparable to the vocoder-based VC at 25 dB of the SNR. Moreover, it can be seen from Fig. 2 (b) that the vocoder-free VC at 5 dB of the SNR is comparable to the vocoder-based VC at 15 dB of the SNR. These results reveal that the vocoder-free VC with DIFFVC is more robust against background sounds compared to the vocoder-based VC.

4. Robustness Analysis

We analyze the robustness of the VC methods under noisy environments. It is expected that if we develop an objective metric to sensitively capture the robustness of VC against background sounds, it will be helpful to develop a more robust VC method. Moreover, it will be more useful if such an objective metric can be calculated without using any reference signals since it will be available in the conversion process. In this paper, we investigate the effectiveness of kurtosis ratio [23] of power spectrum components before and after conversion as one of the objective metrics to evaluate the robustness of VC.

XC9gGwz0?usp=sharing



(a) Input source singing voice

(b) clean DIFFVC + noise (ideal result)

(c) noisy DIFFVC

(d) noisy VC

Figure 3: *Example of power spectrograms of converted singing voices under noisy conditions.*

4.1. Comparison of Power Spectrograms

Figure 3 shows power spectrograms of the input source singing voice, clean DIFFVC + noise, noisy DIFFVC, and noisy VC when the SNR is 10 dB. We can see that harmonic components are clearly observed in the noisy VC (Fig. 3 (d)) than in the others. This is because the source excitation signal is generated using the traditional mixed excitation model of the vocoder in the vocoder-based VC (i.e., noisy VC). Consequently, the spectral structure of the converted singing voice is significantly different from that by clean DIFFVC + noise or by the input source singing voice. On the other hand, the original source excitation signal is hold in the vocoder-free VC with DIFFVC (i.e., noisy DIFFVC). Therefore, its converted spectral structure is still similar to that of the clean DIFFVC + noise or of the input source singing voice. As a result, the converted singing voices by noisy VC perceptually sound very different from those by clean DIFFVC + noise while those by noisy DIFFVC still sound similar to them; e.g., the background sounds can be perceptually separable from the singing voices in the converted singing voices by

noisy DIFFVC as in the clean DIFFVC + noise but it is hard to distinguish them in the converted singing voices by noisy VC.

4.2. Kurtosis Ratio

Kurtosis is defined as

$$\text{kurt} = \frac{\mu_4}{{\mu_2}^2}, \tag{1}$$

where kurt represents kurtosis and μ_n is the n th moment, which is given by

$$\mu_n = \int_0^\infty x^n P(x) dx, \tag{2}$$

where $P(x)$ is the probability density distribution of the signal x. In this paper, we use power spectrum components in the time frequency domain as the signal. To handle the signals in the power domain, the n-th moment not around the mean value but around the origin is used.

Because the calculation of higher-order statistics easily suffers from lack of stability and accuracy, we use a probability density modeling technique of the power spectrum components with gamma distribution [24, 25]. The probability density function of the gamma distribution is given by

$$P(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} \cdot x^{\alpha-1} e^{-\frac{x}{\theta}}, \tag{3}$$

where $x \geq 0$, $\alpha > 0$, and $\theta > 0$. α is a shape parameter, θ is a scale parameter, and $\Gamma(\alpha)$ is the gamma function. These parameters can be estimated with the maximum likelihood estimation method. Kurtosis of the signal following this probability density function can be calculated as follows [23]:

$$\text{kurt}^{(\text{g})} = \frac{(\alpha+2)(\alpha+3)}{\alpha(\alpha+1)}. \tag{4}$$

Then, the kurtosis ratio between before and after conversion is defined by the following equation.

$$\text{kurtosis ratio} = \frac{\text{kurt}^{(\text{g})}_{\text{con}}}{\text{kurt}^{(\text{g})}_{\text{org}}}, \tag{5}$$

where $\text{kurt}^{(\text{g})}_{\text{org}}$ is kurtosis of the input source singing voice, and $\text{kurt}^{(\text{g})}_{\text{con}}$ is kurtosis of the converted singing voice.

The kurtosis ratio was originally proposed as an objective metric to evaluate artifacts called musical noise, which is often caused by nonlinear noise suppression processing [23]. In this paper, we use it to evaluate changes of the power spectral components caused by the conversion processing as shown in Fig. 3. This metric can be calculated using only the input and the converted waveforms in VC.

4.3. Investigation of Effectiveness of Kurtosis Ratio

We investigated the kurtosis ratio calculated over a 0-2110 Hz frequency band. The SNR of the input source singing voice was set to 5, 10, 15, 20, 25 and 30 dB.

Figure 4 shows the result under the environmental noise. The kurtosis ratio of clean DIFFVC + noise is almost 1. This means that the kurtosis ratio is not significantly affected by a difference of the singers (e.g., Fig. 3 (a) and Fig. 3 (b)). In the vocoder-based VC (noisy VC), the kurtosis ratio is significantly larger than 1. This is because a clear harmonic structure is unnaturally generated by using the vocoder as shown in Fig 3 (d).

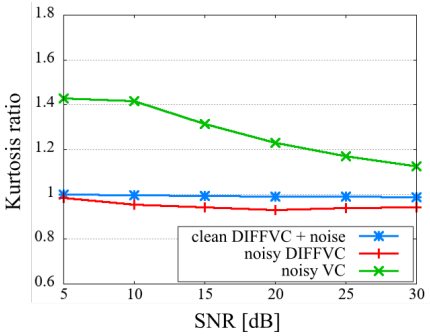


Figure 4: *Kurtosis ratio as a function of SNR under environmental noise.*

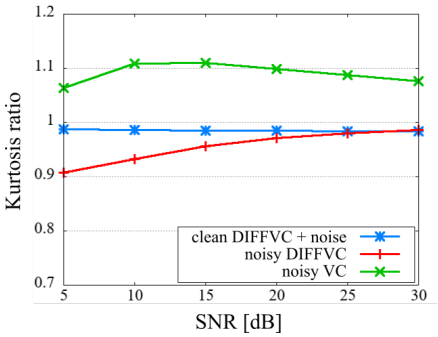


Figure 5: *Kurtosis ratio as a function of SNR under accompanying sound.*

On the other hand, the vocoder-free VC with DIFFVC (noisy DIFFVC) can maintain the Kurtosis ratio around 1 since it can still keep a specific spectral structure under the environmental noise as shown in Fig. 3 (c).

Figure 5 shows the result under the accompanying sound. We can see a similar tendency to Fig. 4. One difference is the result of the vocoder-free VC with DIFFVC (noisy DIFFVC). The kurtosis ratio is maintained around 1 when the SNR is high but it tends to decrease as the SNR decreases. This indicates that DIFFVC can perform the conversion while maintaining a specific spectral structure under the accompanying sound when the SNR is high, but some changes are caused when the SNR is low. It is expected that we further improve the vocoder-free VC with DIFFVC under the accompanying sound by using the kurtosis ratio as an additional constraint in the conversion process.

5. Conclusions

In this paper, we have investigated the robustness of the statistical voice conversion (VC) against background sounds. The results of subjective evaluation have demonstrated that the vocoder-free VC is more robust against the background sounds than the vocoder-based VC. Moreover, we have analyzed the robustness of VC against the background sounds using the kurtosis ratio of power spectrum components. The result has shown that it is an effective metric to partially evaluate the robustness of VC. We plan to incorporate this objective metric for the conversion processing to further improve conversion performance under noisy conditions.

6. Acknowledgements

This work was partly supported by JST, PRESTO Grant Number JPMJPR1657, and JSPS KAKENHI Grant Number 17H0176.

7. References

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

[2] 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, 2007.

[3] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," *Proc. INTERSPEECH*, pp. 2162–2165, 2010.

[4] 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, 2010.

[5] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," *Proc. INTERSPEECH*, pp. 94–97, 2012.

[6] T. Toda, "Augmented speech production based on real-time statistical voice conversion," *IEEE GlobalSIP*, pp. 592–596, 2014.

[7] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

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

[9] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," *Proc. SMAC*, pp. 287–292, 2013.

[10] 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.

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

[12] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." *Proc. INTERSPEECH*, pp. 1118–1122, 2017.

[13] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "Fftnet: A real-time speaker-dependent neural vocoder," *Proc. ICASSP*, pp. 2251–2255, 2018.

[14] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, 2018.

[15] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," *SLT*, pp. 313–317, 2012.

[16] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *Proc. ICASSP*, pp. 5069–5073, 2018.

[17] S. Pascual, M. Park, J. Serrà, A. Bonafonte, and K.-H. Ahn, "Language and noise transfer in speech enhancement generative adversarial network," *Proc. ICASSP*, pp. 5019–5023, 2018.

[18] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," *Proc. ICASSP*, pp. 716–720, 2018.

[19] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[20] K. Kobayashi, T. Toda, and S. Nakamura, "F_0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," *SLT*, pp. 693–700, 2016.

[21] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software," *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pp. 203–210, 2018.

[22] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. & syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[23] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical noise generation analysis for noise reduction methods based on spectral subtraction and mmse stsa estimation," *Proc. ICASSP*, pp. 4433–4436, 2009.

[24] J. W. Shin, J.-H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, 2005.

[25] T. H. Dat, K. Takeda, and F. Itakura, "Gamma modeling of speech power and its on-line estimation for statistical speech enhancement," *IEICE Trans. on Inf. and Syst.*, vol. 89, no. 3, pp. 1040–1049, 2006.