Vol.2018-MUS-118 No.3
Vol.2018-SLP-120 No.3
2018/2/20

Application of the Velvet noise and its variant for synthetic speech and singing

Hideki Kawahara[1,a)]

Abstract: The Velvet noise is a sparse signal which sounds smoother than Gaussian white noise. We propose the direct use of the velvet noise and application of its variant for speech and singing synthesis. An interesting variant uses the symmetry of time and frequency in Fourier transform to design the desired signal. This variant can replace the logarithmic domain pulse model, mixed excitation source signals and, a group delay-manipulated excitation pulse which is the excitation source signal of legacy-STRAIGHT.

1. Introduction

The Velvet noise is a sparse signal which sounds smoother than Gaussian white noise [1, 2]. We found that the Velvet noise itself and its variants provide useful candidates for the excitation source signals of synthetic speech and singing. They can replace excitation source signal models [3–6] for VOCODERs [3, 7, 8] and provide a unified design procedure of mixed-mode excitation signals.

2. Background

How to analyze and generate the random component for synthetic voice has been a difficult problem. The significant variation of the masking level of a burst sounds within one pitch period [9] made this problem harder. The characteristic buzziness also has been a source of severe degradation in analysis-and-synthesis type VOCODERs. This degradation is made worse in statistical text-to-speech systems. Although WaveNet [10] effectively made this problem disappear, a flexible and general purpose excitation signal will be beneficial for interactive and compact applications.

One successful implementation of a less-buzzy source signal is a group delay manipulated pulse introduced in legacy-STRAIGHT [3]. The source signal uses a smoothed random noise for designing the group delay in higher (typically 3 kHz) frequency region. The smoothing parameter and the magnitude of group delay variation were pre-determined based on trial-and-error tests. Even with several investigations [4], the source model failed to be coupled with relevant analysis procedures to determine these parameters. The revised STRAIGHT (TANDEM-STRAIGHT [7]) also failed to formulate a unified, flexible framework for the excitation source signal, after several trials [5, 11, 12]. The recent introduction of (LDPM: log-domain pulse model) seems to provide a unified framework consisting of relevant analysis procedure [6]. We tried a variant of the LDPM. Although the signal showed desirable behavior, it introduced smearing of the random component [13].

It is time to reconsider revising new lines of VOCODER [8, 14] because patents which prevented use of simple procedures for improving synthetic voice quality are expired. The group delay manipulated pulse and other quality improvement procedures used in legacy-STRAIGHT were not used in TANDEM-STRAIGHT to prevent infringement of the patents. Because of this issue and other minor factors, the synthesized speech quality using legacy-STRAIGHT was better than TANDEM-STRAIGHT [8]. These quality-related patents of legacy-STRAIGHT were expired before 2018 and free to use them now.

The velvet noise and its variants provide the key for this revision of excitation signals. In the following section, we introduce the original velvet noise and its time-domain variant. Then, after discussions on their behavior, we introduce the frequency-domain variant of the velvet noise.

3. Velvet noise and time domain variants

The velvet noise was designed for artificial reverberation algorithms. It is a randomly allocated unit impulse sequence with minimal impulse density vs. maximal smoothness of the noise-like characteristics. Because such sequence can sound smoother than the Gaussian noise, it is named “velvet noise.” [1]

3.1 Original velvet noise

The velvet noise allocates a randomly selected positive or negative unit pulse at a random location in each temporal segment [1]. The following equation determines the location of the m-th pulse $k_{\mathrm{ovn}}(m)$. The subscript “ovn” stands for “Original Velvet Noise.”

$$k_{\mathrm{ovn}}(m) = \|mTd + r_1(m)(T_d - 1)\|, \tag{1}$$

where T_d represents the average pulse interval in samples. The

[1] Wakayama University, Wakayama, Wakayama 640–8510, Japan
[a)] kawahara@sys.wakayama-u.ac.jp

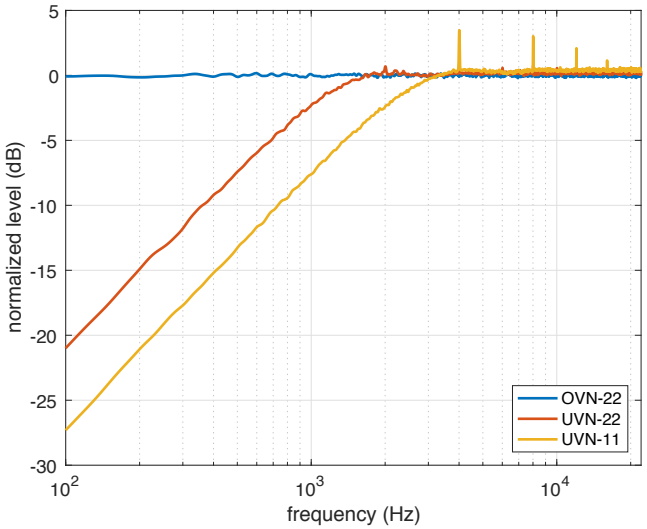

Fig. 1 Long time average of the power spectrum of OVN and UVNs. `OVN-22` and `UVN-22` used $T_d = 22$ samples and `UVN-11` used $T_d = 11$.

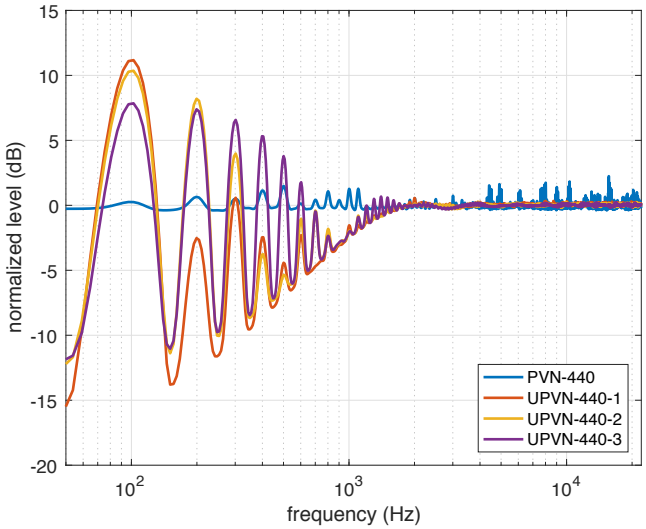

Fig. 2 Power spectrum of PVN and PUVNs All signals used $T_d = 22$ samples and $T_p = 440$ samples. `UPVN-440-1`, `UPVN-440-2` and `UPVN-440-3` used 193, 193/2 and 193/4 samples for T_w.

following equation determines the value of the signal $s_{\text{ovn}}(n)$ at discrete time n.

$$s_{\text{ovn}}(n) = \begin{cases} 2\|r_2(m)\| - 1 & n = k_{\text{ovn}}(m) \\ 0 & \text{otherwise} \end{cases}. \qquad (2)$$

3.2 Time domain variants of velvet noise

We introduce three variants of velvet noise; a unipolar velvet noise (UVN), a periodic velvet noise (PVN), and their combination, a unipolar periodic velvet noise (UPVN). The UVN modifies the value in Eq. (2). The following equation provides the value of UVN, $s_{\text{uvn}}(n)$ at a discrete time n.

$$s_{\text{uvn}}(n) = \begin{cases} 1 & n = k_{\text{ovn}}(m) \\ 0 & \text{otherwise} \end{cases}, \qquad (3)$$

The PVN modifies the time index in Eq. (2). The PVN has additional two factors; the fundamental period T_p and the duty cycle $D = T_w/T_p$. The following equation provides the value of UVN, $s_{\text{uvn}}(n)$ at a discrete time n.

$$s_{\text{pvn}}(n; T_p, T_w) = \begin{cases} 2\|r_2(m)\| - 1 & Q(m; T_p, T_w) \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

$$Q(m; T_p, T_w) = \left(n \bmod T_p = k_{\text{ovn}}(m)\right) \wedge \left(n \bmod T_p \le T_w\right),$$

where $Q(m; T_p, T_w)$ is a mathematical predicate representing the condition and "mod" represents the modulo operator.

The following equation provides the value of UPVN, $s_{\text{upvn}}(n)$ at a discrete time n.

$$s_{\text{upvn}}(n; T_p, T_w) = \begin{cases} 1 & Q(m; T_p, T_w) \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

3.3 Frequency domain characteristics

OVN with a pulse density higher than 2,000 pulses per second sounds smoother than Gaussian noise [1, 2]. This section illustrates numerical examples of the OVN and the variants in this pulse density region. The sampling frequency is 44,100 Hz in the following examples.

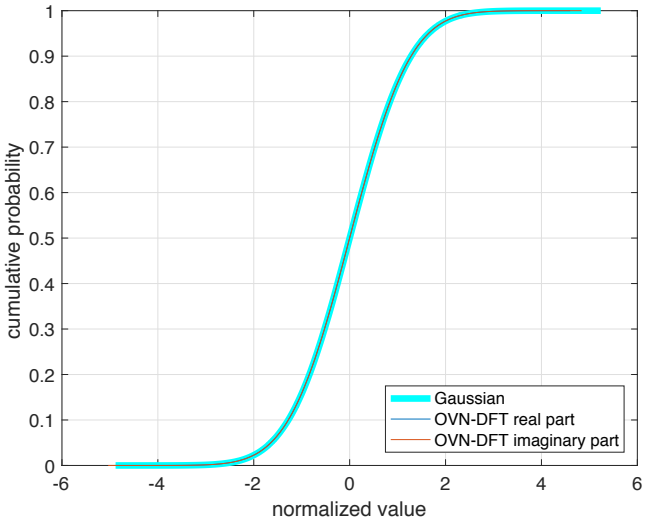

Fig. 3 Cumulative distribution of DFT sequences of OVN. Thick cyan plot shows the cumulative Gaussian distribution.

Figure 1 shows average power spectra of OVN and UVNs. The segment length was $T_d = 22$ samples for `OVN-22` and `UVN-22`. `UVN-11` used $T_d = 11$. The signal duration was 100 s. The power spectra used the Blackman window with 50 ms length and 50% overlap. Note that the average value of UVN was subtracted. Spectral peaks correspond to integer multiples of $1/T_d$.

Figure 2 shows average power spectra of PVN and PUVNs. All signals used $T_d = 22$ samples and $T_p = 440$ samples. `UPVN-440-1`, `UPVN-440-2` and `UPVN-440-3` used 193, 193/2 and 193/4 samples for T_w. The signal duration was 100 s. The fundamental frequency of the harmonic structure of `UPVN`s is $1/T_w$. The spectrum envelope in the lower frequency region is sinc function.

3.4 DFT sequence characteristics of OVN

Discrete Fourier Transform (DFT) converts a periodic time-domain sequence to a periodic frequency-domain complex sequence. The real part of the sequence has even symmetry and the imaginary part has odd symmetry.

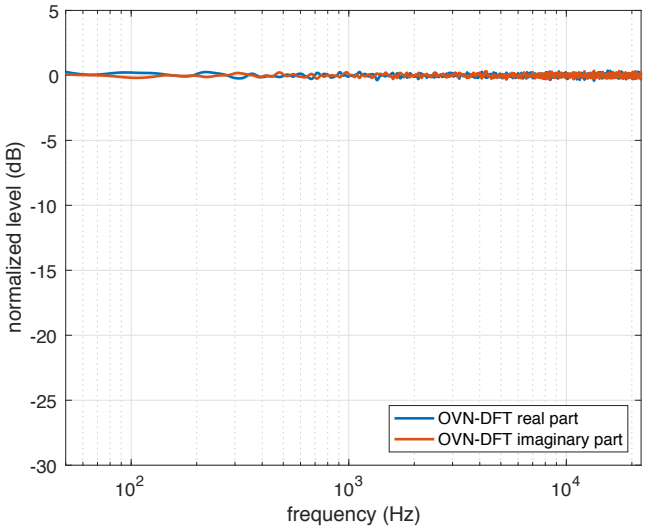

Fig. 4 Long time average power spectrum of DFT sequences of OVN.

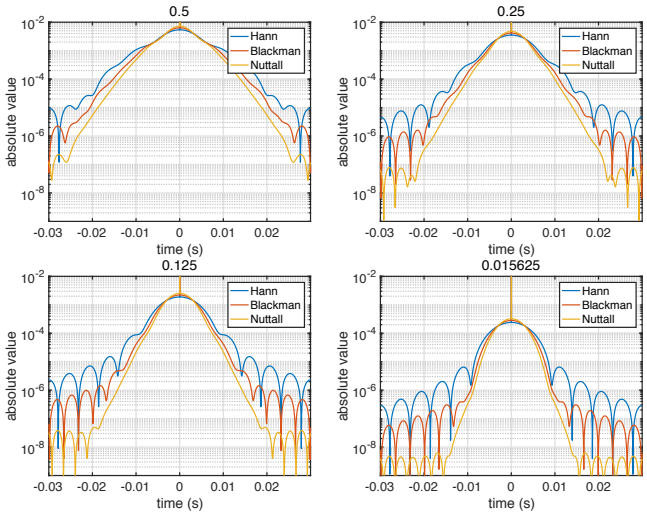

Fig. 5 Absolute value of unit phase manipulation. The title of each plot represents the maximum value of $w_p(k)$.

Figure 3 shows the simulation results. The tested OVN has $T_d = 16$ and the length of 2^{21} samples. The initial half bins of the real and imaginary part of DFT of the OVN sequence are used to calculate this distribution. It is safe to state that value distribution of the real and the imaginary part of the DFT sequence of OVN sequence is Gaussian.

Figure 4 shows the long time average power spectrum of the real and imaginary part of DFT of the OVN sequence. In this plot, each DFT sequence is considered as a time series. Figures 3 and 4 suggest that each DFT sequence is a Gaussian random sequence.

Applying a time invariant (linear phase) FIR filter to OVN shapes the DFT sequences with the filter's spectral shape. In other words, it yields shaped Gaussian random sequences. This is the underlying idea of the frequency domain variants of velvet noise.

4. Velvet noise and frequency domain variants

Allpass filter has a constant gain with (usually) nonlinear phase characteristics. A causal allpass filter using pole-zero pairs has an exponentially decaying impulse response [15]. The legacy-STRAIGHT used smoothed group delay for designing allpass filters and used them for the excitation source [3]. We propose to use velvet noise procedure to design allpass filters. Using velvet noise procedure for designing phase of allpass filters makes their impulse responses localized.

4.1 Unit phase manipulation

This section investigates relations between phase manipulation and the impulse response of corresponding alllpass filter. Let $w_p(k, B_k)$ represent a phase modification function on the discrete frequency domain. The following equation provides the complex valued impulse response $h(n; k_c, B_k)$ of the allpass filter.

$$h(n; k_c, B_k) = \frac{1}{K} \sum_{k=0}^{K-1} w_p(k - k_c, B_k) \exp\left(\frac{2kn\pi j}{KN}\right), \qquad (6)$$

where k_c represents the discrete center frequency and B_k represents the nominal band width of $w_p(k, B_k)$.

Figure 5 shows the absolute value of each impulse response. Three time window function shaped the phase response. The

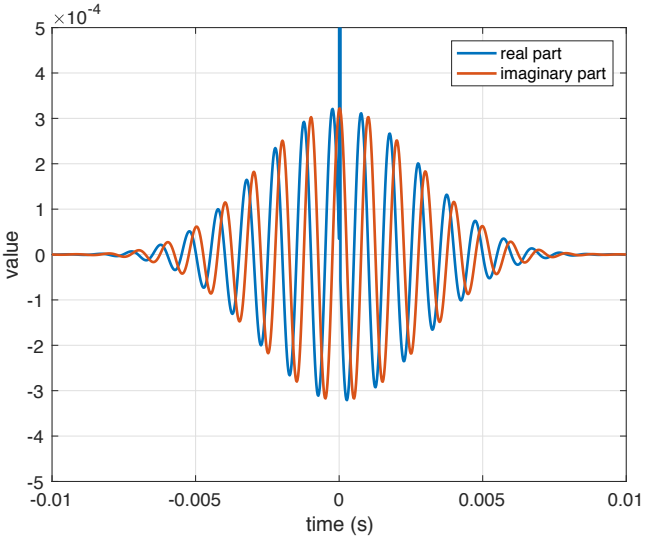

Fig. 6 Impulse response example of the designed allpass filter.

following cosine series defines these windows.

$$w_p(k, B_k) = \sum_{n=0}^{N} a(n) \cos\left(\frac{\pi kn}{NB_k}\right). \qquad (7)$$

Note that outside of the support $-B_k < k < B_k$, the shape is $w_p(k, B_k) = 0$. Three windows are Hann, Blackman and Nuttall. The coefficients of each function is listed in [16]. In this simulation, the center frequency was 1,000 Hz and the bandwidth corresponds to 100, 150, and 200 Hz respectively.

Figure 6 shows an example impulse response. This example corresponds to the bottom right plot of Fig. 5. Note that the maximum value at time 0 is close to 1.

4.2 Center frequency allocation after velvet noise

By adding unit phase manipulation $w_p(k - k_c, B_k)$ on randomly allocated center frequency k_c yields the filtered velvet noise on the frequency domain. The following equation defined the allocation index (discrete frequency) $k_c = k_{\mathrm{fvn}}(m)$ where subscript "fvn" stands for Frequency domain Velvet Noise.

$$k_{\mathrm{fvn}}(m) = \|mF_d + r_1(m)(F_d - 1)\|, \qquad (8)$$

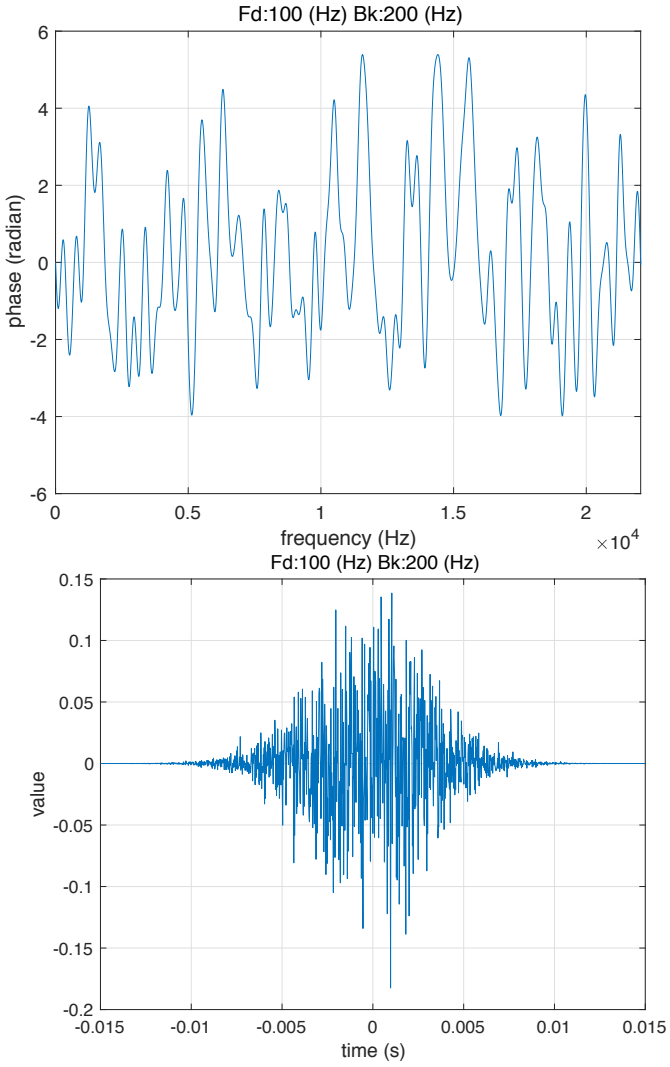

Fig. 7 Frequency domain velvet noise example. Upper plot shows the phase and the lower plot shows the waveform.

where F_d represents the average frequency segment length. Each location spans from 0 Hz to $f_s/2$. Let $\mathbb{K}$ represent a set of allocation indices $k_{\mathrm{fvn}}(m)$. The following equation provides the phase $\varphi_{\mathrm{fvn}}(k)$ of this frequency variant of velvet noise.

$$\varphi_{\mathrm{fvn}}(k) = \sum_{k_c \in \mathbb{K}} \varphi_{\max} \left(w_p(k - k_c, B_k) - w_p(k + k_c, B_k) \right), \qquad (9)$$

where k spans discrete frequency of a DFT buffer, which has a circular discrete frequency axis.

The inverse discrete Fourier transform provides the impulse response of the frequency domain velvet noise.

$$h_{\mathrm{fvn}}(n) = \frac{1}{K} \sum_{k=0}^{K-1} \varphi_{\mathrm{fvn}}(k) \exp\left(\frac{2kn\pi j}{KN} \right). \qquad (10)$$

4.3 Behavior of frequency domain variant

A series of simulations were conducted to test behavior of FVN. The sampling frequency was 44,100 Hz in this section.

Figure 7 shows an example of the frequency domain velvet noise. The width of the frequency segment F_d is 100 Hz and the base band width B_k is 200 Hz. This calculation used 8,192 bins for the DFT buffer length.

Figure 8 shows the averaged RMS (root mean square) value of FVN samples. The number of iterations was 1,000. The legend represents the nominal band width. This plot indicates that FVN is highly localized and designed easily.

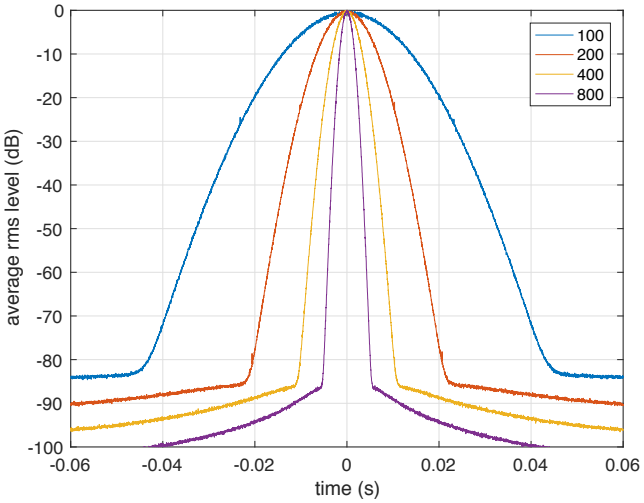

Fig. 8 Average RMS (root mean square) value of FVN samples.

5. Application to speech and singing synthesis

Sparseness of OVN is useful for efficient implementation of unvoiced sounds in speech and singing synthesis. FVN has two applications. By allocating each FVN with the same temporal separation and generating it using different random sequence, it provides an excitation signal spanning from random signal to a purely periodic pulse. The other application is to use one FVN for a filter for reducing buzziness of synthetic voices. Nonlinear frequency axis warping with the group delay representation provides flexible excitation source design procedure. It will be further research topic.

The MATLAB codes are linked from the author's page. They are placed on GitHub and open to everyone.

6. Conclusion

This article introduced the velvet noise and its variants for speech and singing synthesis application. The original velvet noise is useful for efficient implementation. The frequency domain variant is useful for a unified flexible excitation signal and for a buzziness reduction filter. Perceptual evaluation of these applications are further research topics.

Acknowledgments This work was supported by JSPS KAKENHI Grant Numbers JP15H03207, JP15H02726 and JP16K12464.

References

[1] Järveläinen, H. and Karjalainen, M.: Reverberation Modeling Using Velvet Noise, *AES 30th International Conference, Saariselkä, Finland*, Audio Engineering Society,, pp. 15–17 (2007).

[2] Välimäki, V., Lehtonen, H. M. and Takanen, M.: A Perceptual Study on Velvet Noise and Its Variants at Different Pulse Densities, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 7, pp. 1481–1488 (online), DOI: 10.1109/TASL.2013.2255281 (2013).

[3] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207 (1999).

[4] Kawahara, H., Estill, J. and Fujimura, O.: Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT, *Proceedings of MAVEBA*, Firentze Italy, pp. 59–64 (2001).

[5] Kawahara, H., Morise, M., Takahashi, T., Banno, H., Nisimura, R. and Irino, T.: Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems, *Interspeech 2010*, Makuhari Japan, pp. 38–41 (2010).

[6] Degottex, G., Lanchantin, P. and Gales, M.: A Log Domain Pulse Model for Parametric Speech Synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 1, pp. 57–70 (online), DOI: 10.1109/TASLP.2017.2761546 (2018).

[7] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, *ICASSP 2008*, Las Vegas, pp. 3933–3936 (2008).

[8] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: A vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).

[9] Skoglund, J. and Kleijn, W. B.: On time-frequency masking in voiced speech, *Speech and Audio Processing, IEEE Transactions on*, Vol. 8, No. 4, pp. 361–369 (online), DOI: 10.1109/89.848218 (2000).

[10] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499*, pp. 1–15 (2016).

[11] Kawahara, H., Irino, T. and Morise, M.: An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction, *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, pp. 5420–5423 (2011).

[12] Kawahara, H., Morise, M., Toda, T., Banno, H., Nisimura, R. and Irino, T.: Excitation source analysis for high-quality speech manipulation systems based on an interference-free representation of group delay with minimum phase response compensation, *Interspeech 2014*, Singapore, pp. 2243–2247 (2014).

[13] Kawahara, H. and Sakakibara, K.-I.: An extended log domain pulse model for VOCODERs, *IEICE Technical Report*, No. SP2017-66, pp. 1–4 (2018). [In Japanese].

[14] Kawahara, H., Agiomyrgiannakis, Y. and Zen, H.: YANG vocoder, Google (online), available from ⟨https://github.com/google/yang_vocoder⟩ (accessed 2017-01-17).

[15] Oppenheim, A. V. and Schafer, R. W.: *Discrete-time signal processing: Pearson new International Edition*, Pearson Higher Ed. (2013).

[16] Nuttall, A. H.: Some windows with very good sidelobe behavior, *IEEE Trans. Audio Speech and Signal Processing*, Vol. 29, No. 1, pp. 84–91 (1981).