Application of velvet noise and its variants for synthetic speech and singing (Revised and extended version with appendices)

(Compiled: 12:35 A.M., 20 February, 2018)

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 variants for speech and singing synthesis. A new set of variants uses the symmetry of time and frequency in Fourier transform to design the desired signal. These variants 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. This version provides error corrections and detailed design procedures to the technical report presented at 118th SIGMUS meeting.

1. Introduction

The Velvet noise is a sparse discrete signal which consists of fewer than 20% of non-zero (1 or -1) elements. The name "velvet" represents its perceptual impression. It 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. In addition, the proposed frequency variant of the velvet noise is also an impulse response of an all-pass filter [9]. It provides an effective and easy way for reducing "buzzy" impression of VOCODER speech sounds. This article introduces the velvet noise and its time domain and frequency domain variants and discusses its use in singing and speech synthesis. This version is a revision and extension of the article [10] presented at SIGMUS/SLP meeting held at Tsukuba Japan on 20, 21 February 2018.

2. Background

How to analyze and generate the random component for synthetic voice has been a difficult problem [4, 6, 11, 12]. In addition to this difficulty in analysis and synthesis, auditory perception introduces another difficulty. It is the significant variation of the masking level of a burst sounds within one pitch period [13]. Two synthetic speech sounds having 20 dB SNR difference are perceptually equal in a specific condition. The characteristic buzziness also has been a source of severe degradations in analysis-and-synthesis type VOCODERs. This degradation is made worse in statistical text-to-speech systems [14]. Although WaveNet [15] 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, 16, 17]. 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 temporal smearing of the random component [18].

It is time to reconsider revising new lines of VOCODER [8, 19] because patents which prevented the 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 have not been 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

---

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

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 variants. Then, after discussions on their behavior, we introduce the frequency-domain variants of the velvet noise. These frequency domain variants are the main contribution of this article.

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."

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

where T_d represents the average pulse interval in samples. The following equation determines the value of the signal $s_{\mathrm{ovn}}(n)$ at discrete time n.

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

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, $s_{\mathrm{uvn}}(n)$ at a discrete time n.

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

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, $s_{\mathrm{uvn}}(n)$ at a discrete time n.

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

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

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, $s_{\mathrm{upvn}}(n)$ at a discrete time n.

$$s_{\mathrm{upvn}}(n; T_p, T_w) = \begin{cases} 1 & Q(m; T_p, T_w) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$
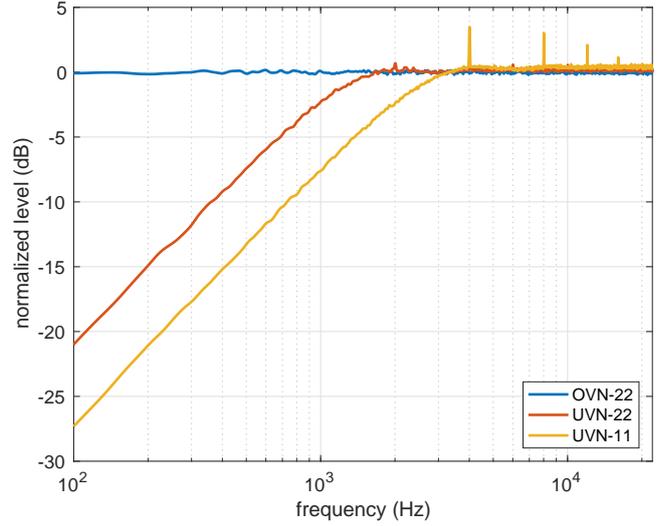
Fig. 1 Long time average of the power spectrum of OVN and UVNs. `OVN-22` and `UVN-22` used $T_d = 22$ samples and `UVN-11` used $T_d = 11$.
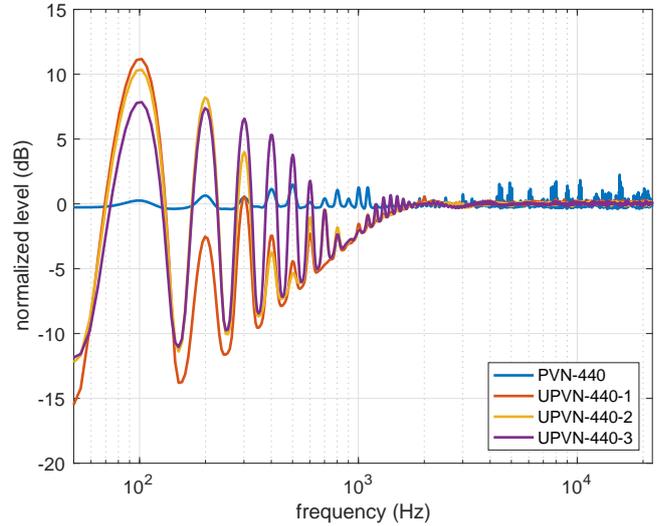

Fig. 2 Power spectrum of PVN and PUVNs All signals used $T_d = 22$ samples and $T_p = 440$ samples. `UPVN-440-1`, `UPVN-440-2` and `UPVN-440-3` used $193, 193/2$ and $193/4$ samples for T_w.

3.3 Frequency domain characteristics

OVN with a pulse density higher than 2,000 pulses per second for 44,100 Hz sampling rate sounds smoother than Gaussian white 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.

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 found in `UVN-11` 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 UPVNs is
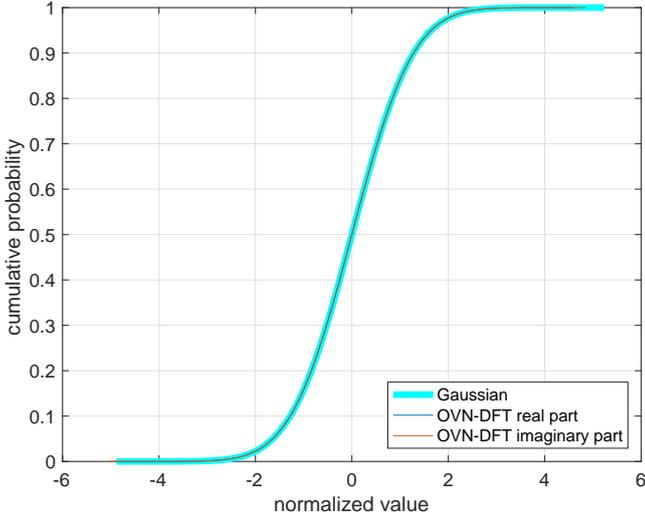
Fig. 3 Cumulative distribution of DFT sequences of OVN. Thick cyan plot shows the cumulative Gaussian distribution.
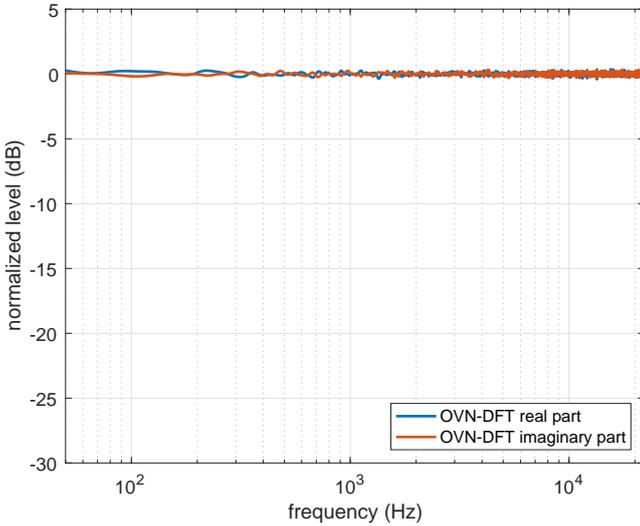

Fig. 4 Long time average power spectrum of DFT sequences of OVN.

$1/T_w$. 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 the odd symmetry.

Figure 3 shows the simulation results. The tested OVN has $T_d = 16$ and the length of 2^{21} samples. The first 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 [20].

Figure 4 shows the long-time average power spectrum of the real and imaginary part of DFT of the OVN sequence. This plot used each DFT sequence 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 shaping using an FIR filter is the underlying idea of the frequency domain variant of velvet noise.

4. Frequency domain variant of velvet noise

By exchanging the time and the frequency, we design an all-pass filter based on velvet noise procedure. We call the impulse response of the all-pass filter as FVN (Frequency domain Velvet Noise). FVN uses the FIR-filtered velvet noise for the phase characteristics of the all-pass filter.

An all-pass filter has a constant gain with (usually) nonlinear phase characteristics. A causal all-pass filter using pole-zero pairs has an exponentially decaying impulse response [9]. The legacy-STRAIGHT used smoothed group delay for designing all-pass filters and used them for the excitation source [3]. Their impulse responses are not localized. We propose to use the velvet noise procedure to design all-pass filters. Using velvet noise procedure for designing phase of all-pass filters makes their impulse responses localized.

4.1 Unit of phase manipulation

We use a set of cosine series functions for manipulating the phase because it is easy to implement well behaving localization [21, 22]. This section investigates relations between phase manipulation and the impulse response of the corresponding all-pass 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 all-pass filter.[*1]

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

where k_c represents the discrete center frequency, and B_k defines the support of $w_p(k, B_k)$ in the frequency domain (i.e. $w_p(k, B_k) = 0$ for $|k| > B_k$). The symbol of the imaginary unit is $j = \sqrt{-1}$ and N represents the number of DFT bins.

We tested four types of cosine series. They are Hann, Blackman, Nuttall, and the six-term cosine series used in [22]. The Nuttall's reference [21] provides a list of coefficients of the first three functions and the design procedure. The following cosine series defines these windows.

$$w_p(k, B_k) = \sum_{m=0}^{M} a(m) \cos\left(\frac{\pi km}{B_k}\right), \quad (7)$$

where M represents the highest order of the cosine series. Let define $B_w = B_k/M$ as nominal bandwidth.

Figure 5 shows the absolute value of each impulse response. In this simulation, the center frequency was 1,000 Hz, and The nominal bandwidth was 100 Hz. The maximum phase deviations are $\pi/2, \pi/4, \pi/8$, and $\pi/16$ from top left, top right, bottom left, and bottom right respectively. Note that the six-term cosine series has side lobes lower than -200 dB to the peak level.

Figure 6 shows an example impulse response using the

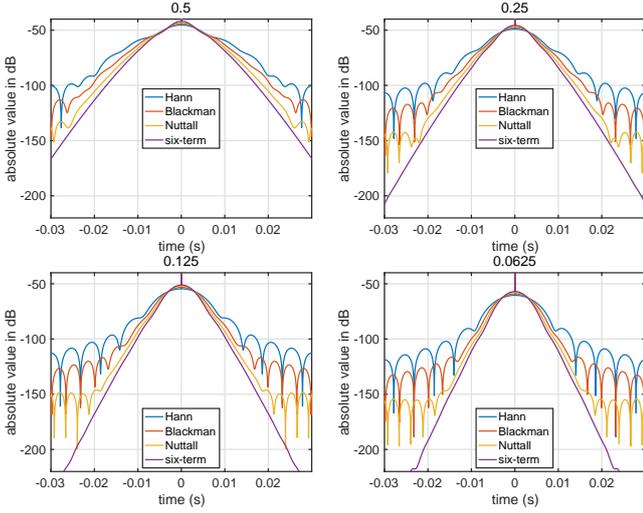[*1] Equation (6) in the reference [10] is a mistake. Equation (6) in this article is correct.

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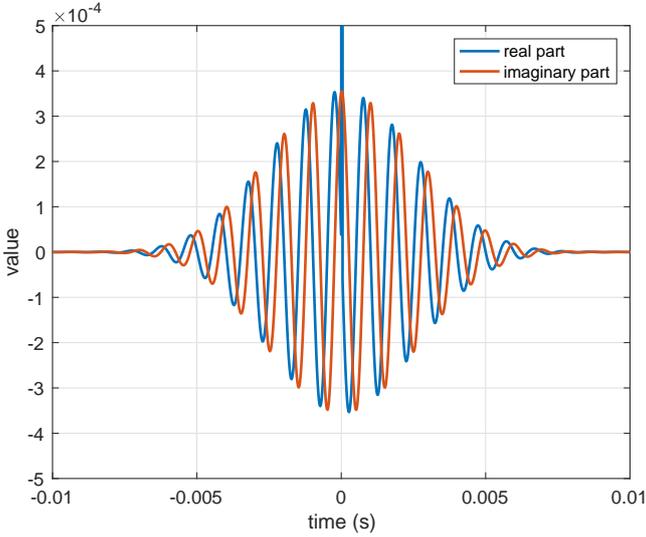

Fig. 6 Impulse response example of the designed all-pass filter using the six-term cosine series.

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

six-term cosine series. 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 Velvet noise-based phase design

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.

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

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.

$$\varphi_{\mathrm{fvn}}(k) = \sum_{k_c \in \mathbb{K}} \varphi_{\max} r_f(k_c) \Big(w_p(k - k_c, B_k) - w_p(k + k_c, B_k) \Big), \quad (9)$$

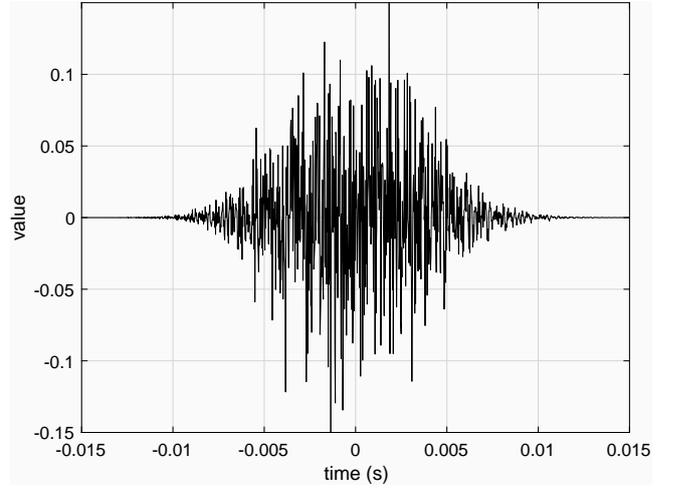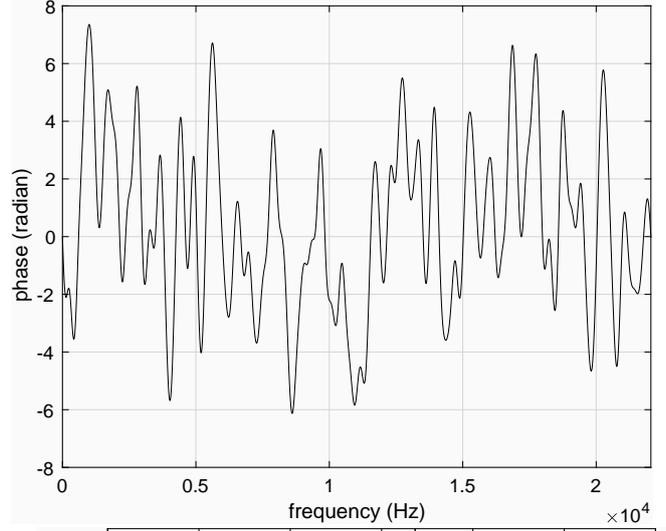where k spans discrete frequency of a DFT buffer, which has a

circular discrete frequency axis. The term $r_f(k_c)$ represents a sample from a random sequence of 1 or -1. The second term inside of parentheses is to make the phase function have the odd symmetry. This symmetry assuring procedure removed the ad hoc shaping function introduced in our article [23] presented at the annual meeting of ASJ.

The inverse discrete Fourier transform provides the impulse response of the frequency domain velvet noise. Note that the corresponding equation in SIGMUS/SLP version [10] is erroneous.

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

4.3 Behavior of frequency domain variant

A series of simulations were conducted to test the behavior of FVN. The sampling frequency was 44,100 Hz in this section. FVN consists of five design parameters. Appendix A.1 describes these parameters and shows test results. Based on these tuning results, examples in this section use the following setting; six-term cosine series DFT buffer size $K = 2^{13}$, average frequency interval $F_d = B_w/6$, and unit phase modification $\varphi_{\max} = \pi/2$.

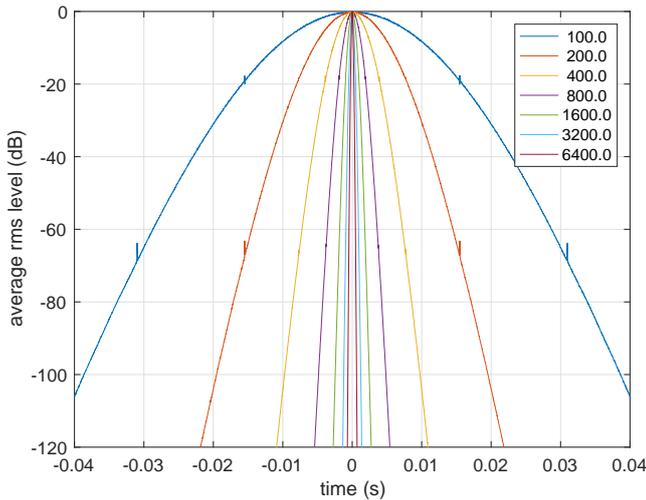Figure 7 shows an example of the frequency domain velvet

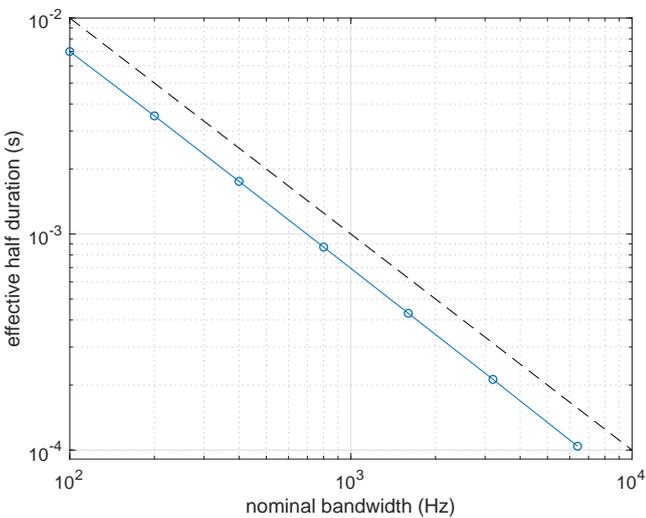**Fig. 8** Average RMS (root mean square) value of FVN samples.

Fig. 9 Nominal bandwidth and effective duration of FVN.

noise. The nominal bandwidth B_w is 200 Hz. The average frequency interval F_d is 66.7 Hz.

Figure 8 shows the averaged RMS (root mean square) value of FVN samples. The number of iterations was 5,000. The legend represents the nominal bandwidth.

Figure 9 shows the relation between the nominal bandwidth and the effective duration d_{50}, which is defined as the duration between 25% location to 75% location of the cumulative power. The dashed line shows the reciprocal of the nominal bandwidth. The effective duration is parallel to the dashed line. These Figures indicate that FVN is highly localized and the effective duration is designed easily by the nominal bandwidth using $B_w = 0.79/d_{50}$.

5. Application to speech and singing synthesis

The 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 the further research topic.

The MATLAB codes are linked from the author's page. They will be 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.

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

[10] Kawahara, H.: Application of the velvet noise and its variant for synthetic speech and singing, *SIGMUS Tech. Report. IPSJ*, Vol. 118, No. 8 (2018).

[11] Yegnanarayana, B., d'Alessandro, C. and Darsinos, V.: An iterative algorithm for decomposition of speech signals into periodic and aperiodic components, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 1, pp. 1–11 (online), DOI: 10.1109/89.650304 (1998).

[12] Malyska, N. and Quatieri, T. F.: Spectral representations of nonmodal phonation, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16, No. 1, pp. 34–46 (online), DOI: 10.1109/TASL.2007.911063 (2008).

[13] 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).

[14] Zen, H., Tokuda, K. and Black, A. W.: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).

[15] 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).

[16] 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).

[17] 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).

[18] 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].

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

[20] Lyon, R. H.: Statistics of Combined Sine Waves, *The Journal of the Acoustical Society of America*, Vol. 48, No. 1B, pp. 145–149 (1970).

[21] 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).

[22] Kawahara, H., Sakakibara, K.-I., Morise, M., Banno, H., Toda, T. and Irino, T.: A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis, *Proc. Interspeech 2017*, Stocholm, pp. 1358–1362 (2017).

[23] Kawahara, H. and Sakakibara, K.-I.: Extending glottal source models using logarithmic domain pulse model, *Proc. Acoustical Society of Japan Spring Meeting*, Saitama, Japan, pp. 2–6–6 (2018).

Appendix

A.1 Tuning of FVN

The seemingly relevant behavior of FVN is a result of trial and error. This section investigates effects of design parameters of FVN and recommends a useful setting.

A.1.1 Design parameters

FVN has following design parameters.

Window shape $w_p(k)$ The six-term cosine series proposed in [22] is the recommended shape. Figure 5 shows that other windows have higher side lobe effects. Nuttall's four-term cosine series is the second choice.

DFT buffer size K Larger the better. This factor will be more significant when using group delay than using phase in the design process.

Average frequency interval F_d This parameter and the following two parameters, the normalized bandwidth, and the unit phase modification are dependent.

Nominal bandwidth B_k The actual bandwidth B_w, which is the width of the support, is a fixed multiple of this width. For Hann, Blackman and Nuttall, their orders are 2, 3, and 4 respectively. For the six-term cosine series, the order is 6.

Unit phase modification $\varphi_{\max}$ This parameter defines the phase modulation depth at its peak.

A.1.2 Effect of Average frequency interval F_d

Figure A·1 shows effects of average frequency interval F_d on the average RMS value of the response. The nominal bandwidth B_k is 200 Hz, and the unit phase modification is $\pi/2$ radian. The larger F_d makes the average RMS value have spikes at the center and $\pm 1/F_d$. These glitches appear when the frequency interval
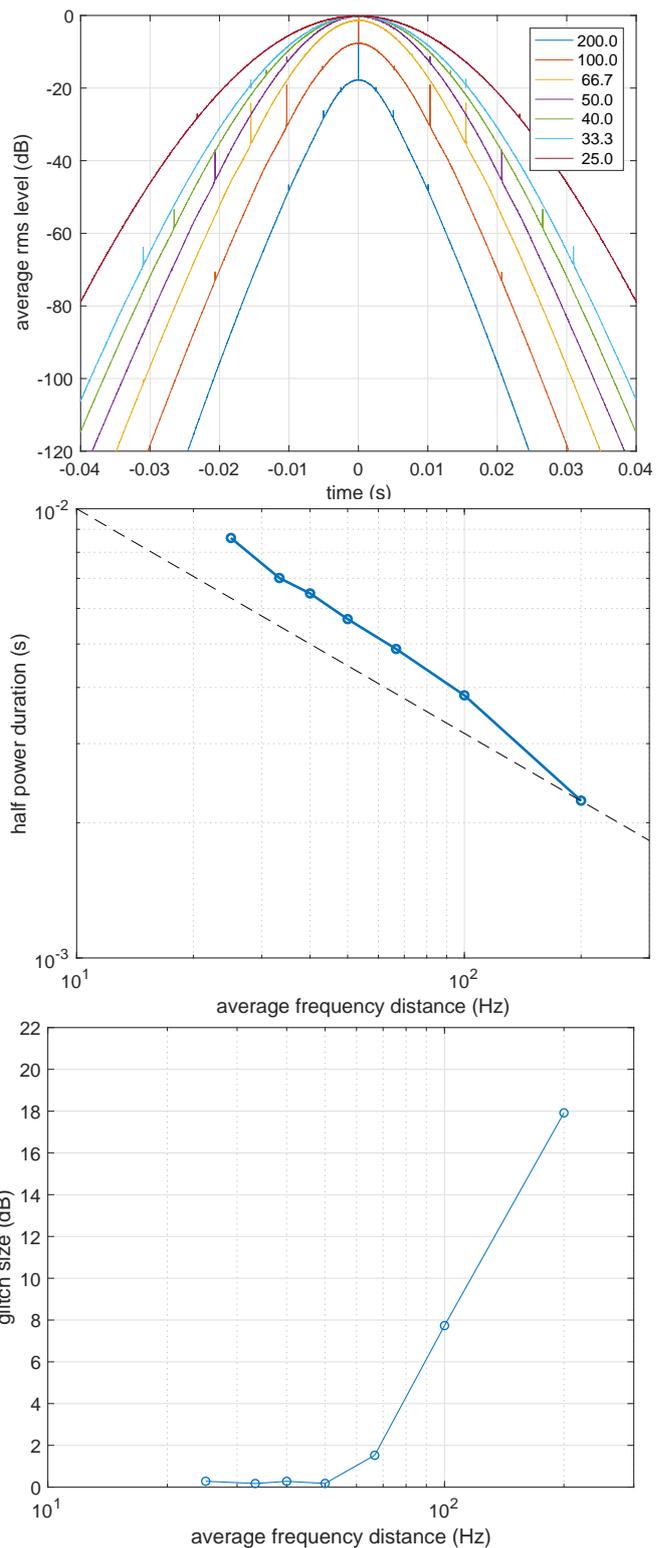

Fig. A·1   Effects of average frequency interval F_d on the average RMS value of the response, effective half duration and the glitch size, from top to bottom respectively

exceeds $B_w/6$. Note that the effective half duration is inversely proportional to the average frequency distance. This square root is because modification of each frequency segment is random and independent.

A.1.3 Unit phase modification $\varphi_{\max}$

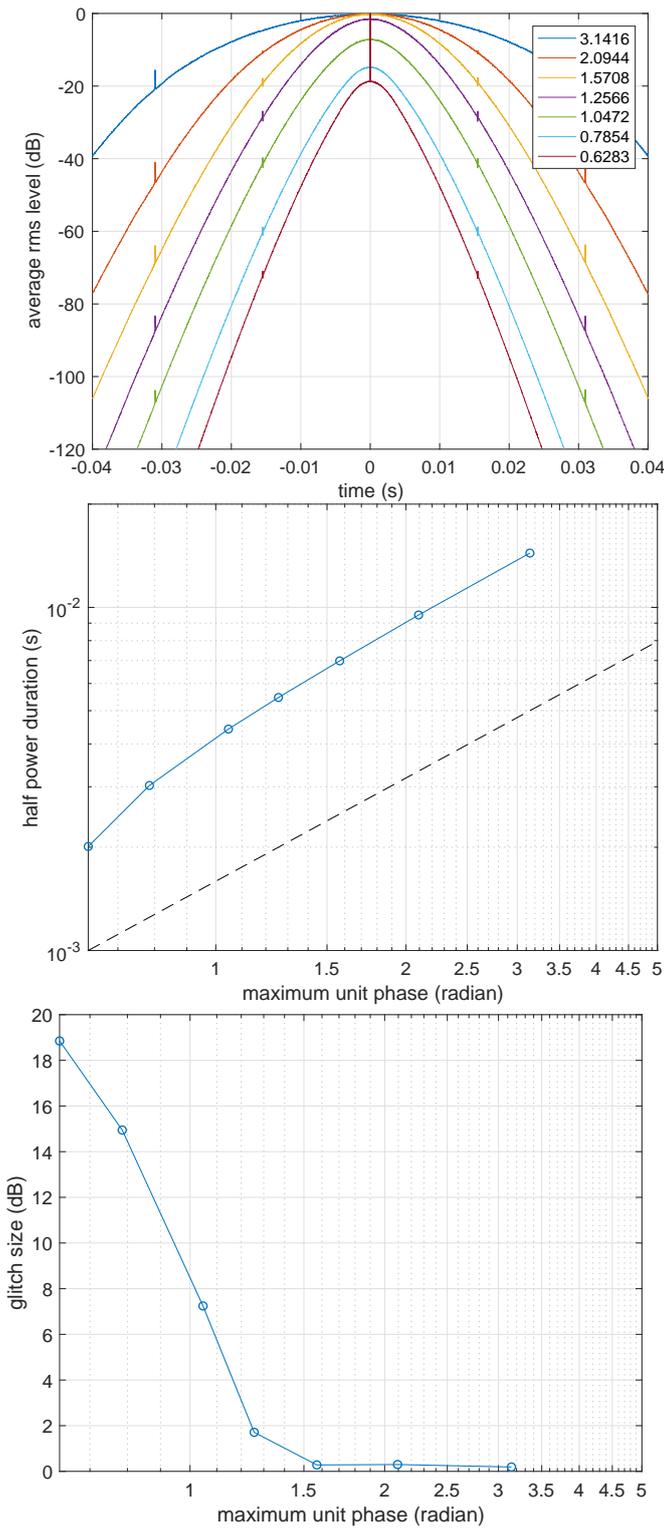Figure A·2 shows effects of unit phase modification $\varphi_{\max}$ on

A.1.4 Recommended parameter setting

These test results suggest the following setting is practically useful; The six-term cosine series for shaping, DFT buffer size $K = 2^{13}$, average frequency interval $F_d = B_w/6$, and unit phase modification $\varphi_{max} = \pi/2$. The other options, which were used to prepare SIGMUS 118 article; Nuttall's four-term window for shaping, DFT buffer size $K = 2^{14}$, average frequency interval $F_d = B_w/5$, and unit phase modification $\varphi_{max} = \pi/2.5$.

A.2 MATLAB implementation

This appendix explains implementation details of MATLAB scripts and functions used in writing this article. (To be avilable)

Fig. A·2 Effects of unit phase modification φ_{max} on the average RMS value of the response, effective half duration and the glitch size, from top to bottom respectively.

the average RMS value of the response, effective half duration and the glitch size. The nominal bandwidth B_k is 200 Hz and the average frequency interval F_d is 33.3 Hz. The glitch at the center appears when the unit phase modification is smaller than $\pi/2$. Note that the effective half duration is proportional to the unit phase modification.