LETTER

Modification of Velvet Noise for Speech Waveform Generation by Using Vocoder-Based Speech Synthesizer

Masanori MORISE[†,††a)], ***Member***

SUMMARY This paper introduces a new noise generation algorithm for vocoder-based speech waveform generation. White noise is generally used for generating an aperiodic component. Since short-term white noise includes a zero-frequency component (ZFC) and inaudible components below 20 Hz, they are reduced in advance when synthesizing. We propose a new noise generation algorithm based on that for velvet noise to overcome the problem. The objective evaluation demonstrated that the proposed algorithm can reduce the unwanted components.

key words: *noise generation, white noise, velvet noise, zero-frequency component, speech analysis/synthesis*

1. Introduction

Vocoders [1] are widely used as speech analyzers for decomposing a speech signal into fundamental frequency, spectral envelope, and aperiodicity information. Several applications such as voice morphing [2] and real-time voice conversion [3] require a high-quality vocoder such as STRAIGHT [4] and WORLD [5] (D4C edition [6]). Recent vocoders can synthesize speech as naturally as that of its input, so we can control speech parameters without degradation.

In waveform generation, white noise is generally used for generating the aperiodic components such as unvoiced phonemes and the aperiodic noise in voiced speech. The pitch synchronous overlap and add (PSOLA) technique [7] is used for waveform synthesis, and it requires short-term white noise. The short-term white noise contains the zero-frequency component (ZFC) and inaudible components below 20 Hz. These components must be reduced in advance when synthesizing to prevent degradation in the synthesized speech. The spectral envelope reflects the power of a short-term waveform, so the short-term power of the noise should be stable to approximate the aperiodic component. White noise with an unstable short-term power is inadequate for this purpose.

To overcome the problem, we propose a new noise generation algorithm based on that for velvet noise [8], [9]. Velvet noise is superior to white noise in the perceived smoothness and stability of short-term power. We propose a modification to the original algorithm to reduce the unwanted components of short-term velvet noise while maintaining the original's advantage. In this paper, the velvet noise generated by the proposed algorithm is defined as the modified velvet noise (MVN). The objective evaluation was carried out to verify the effectiveness of the MVN. The results demonstrated that the MVN can reduce the unwanted components compared with the white and original velvet noises.

Manuscript received August 22, 2018.
Manuscript revised November 3, 2018.
Manuscript publicized December 5, 2018.

[†]The author is with Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Koufu-shi, 400–8511 Japan.
[††]The author is with JST PRESTO, Kawaguchi-shi, 332–0012 Japan.
a) E-mail: mmorise@yamanashi.ac.jp

DOI: 10.1587/transinf.2018EDL8179

2. Proposal of Modified Velvet Noise

In this section, we introduce the original algorithm for generating velvet noise. We then compare its characteristic in the short-term power with that of white noise. Finally, we explain how to generate the MVN.

2.1 Velvet Noise

Velvet noise has several advantages for the waveform generation. We explain how to generate it and then show the advantages.

First, the pulse locations are determined by the following equation.

$$k_{\mathrm{ovn}}(m) = \|mT_{\mathrm{d}} + r_1(m)(T_{\mathrm{d}} - 1)\|, \tag{1}$$

where, m represents the pulse counter ($m = 0, 1, 2, \ldots$), $\|x\|$ represents the rounding function for the input x, $r_1(m)$ represents a sequence of random numbers uniformly distributed from 0.0 to 1.0, and T_d represents the average distance of pulses. The velvet noise $s_{\mathrm{ovn}}(n)$ is given by

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

where, $r_2(m)$ represents a second sequence of random numbers in the same range as above. The amplitude of $s_{\mathrm{ovn}}(n)$ consists of -1, 0, and 1, and its power spectrum is similar to that of white noise.

2.2 Comparison of Short-Term Power

Velvet noise has the characteristic that one pulse is contained in the period from kT_d to $(k + 1)T_d$ (k: integer). Since the powers of all pulses are the same, the short-term power of velvet noise is stable.

Figure 1 shows the relationship between the signal length and standard deviation in the short-term power of

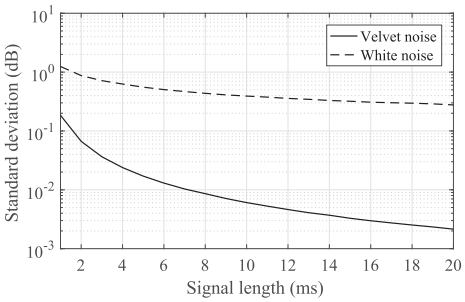

Fig. 1 Relationship between signal length and standard deviation in short-term power of white and velvet noises.

white and velvet noises. In this simulation, T_d was set to four. To calculate the standard deviation, 10,000 noises generated from different random seeds were used, and the long-term powers of both noises were normalized. This result clearly shows that velvet noise is superior to white noise in terms of short-term power stability.

The duration of velvet noise's short-term signal is also stable. However, the ZFC of velvet noise depends on the difference between the number of positive and negative pulses. The ZFC can be removed by unifying the number of both pulses. In this paper, we propose a modification to generate the MVN on the basis of this idea.

2.3 Modified Velvet Noise

To generate the MVN, we first define the sample N as the basic length of $s_{\mathrm{ovn}}(n)$, and the parameter T_d is set to four. In cases where N is set to a multiple of eight, the number of pulses is fixed to an even number. We unify the number of pulses with amplitudes of -1 and 1 by controlling the random number. This modification guarantees that there is no ZFC in the period N. When a long length greater than N is required, N sample velvet noises generated with different random seeds are concatenated and extracted with the required length.

The ZFC can be removed by using short-term N. However, the noise generated with this approach has a frequency peak at f_s/N Hz, where f_s represents the sampling frequency. It is suggested that the peak is observed in cases where the signal has the periodicity by concatenating short-term velvet noises that have no ZFC. We hypothesized that the peak would be reduced by breaking the periodicity and attempted to modify the algorithm to concatenate short-term velvet noises generated with the different samples. We use two samples as the basic parameter, and two short-term velvet noises generated with their samples are randomly selected and concatenated. In this paper, two samples (400- and 152-sample) were determined as the parameter to obtain an approximately flat power spectrum (±0.6 dB) from 100 Hz to the Nyquist frequency. The sampling frequency f_s was fixed to 48 kHz. In cases where the amplitudes of pulse are set to -2 and 2, the long-term power of the MVN equals that of white noise.

Figure 2 illustrates the power spectra of three examples. Power spectra were calculated 10,000 times by using different random seeds, and their averages were calculated. The signal length was set to 65,536. The top and middle graphs show the results of the 400- and 152-sample, respectively. The bottom shows the result of the MVN using both samples. Peaks are visible at f_s/N Hz in the top and middle graphs, but these peaks are reduced by using the proposed modification as shown in the bottom graph. This result shows that the hypothesis is supported. The bottom graph also shows that the inaudible noise below 20 Hz can also be reduced by using the MVN.

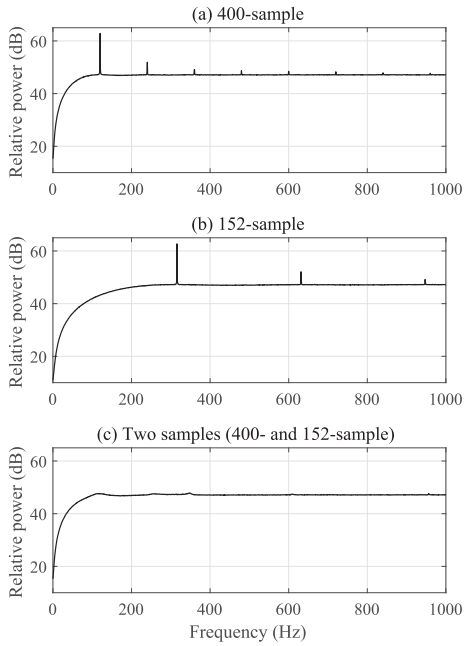

Fig. 2 Power spectra of each noise. Top and middle show noises generated by 400- and 152-sample, respectively. Bottom shows MVN by using both samples.

Since the difference between the original and modified velvet noises is only the order of $r_2(m)$, stability of the short-term power is maintained. The result suggests that the MVN has a characteristic similar to that of a high-pass filter. This has the advantage of generating speech waveforms because recent high-quality vocoders [4], [5] reduce not only ZFCs but also lower-frequency components in advance.

3. Evaluation

We evaluated each noise in the power ratio between the ZFC and whole power in short-term noise. The effectiveness of the MVN is then discussed on the basis of the result.

3.1 Comparison of Power Ratio

In the evaluation, we first generated a long-term noise and then extracted a short-term noise at a random position. A power ratio P was used as the evaluation index, and it is defined by

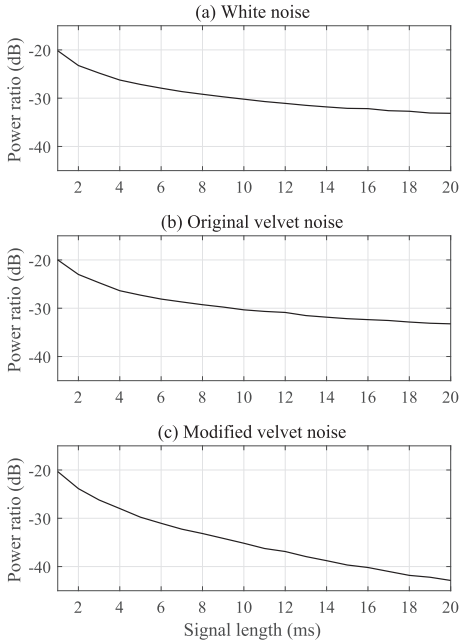

Fig. 3 Relationship between signal length and average power ratios in each noise.

$$P = \frac{\frac{1}{N}\left(\sum_{n=0}^{N-1} x(n)\right)^2}{\sum_{n=0}^{N-1} x^2(n)}, \tag{3}$$

where, $x(n)$ and N represent the short-term noise and the signal length, respectively. The denominator represents the whole power of the signal, and the numerator represents the power of ZFC. Therefore, the index P represents the power ratio between the ZFC and whole signal. When the signal contains no ZFC, the power ratio P indicates 0. We calculated 10,000 power ratios from short-term noises extracted at random positions, and then their median values were calculated as the evaluation index. The white noise, the original velvet noise, and the MVN were compared in the evaluation. In the MVN, two different samples (400- and 152-sample) were used as with the case shown in Fig. 2.

Figure 3 illustrates the results. Those of the white and original velvet noises were almost the same. On the other hand, that of the MVN demonstrated that it could reduce the ZFC compared with the others. In cases where the signal lengths were 10 and 20 ms, the differences of power ratios between the original and modified velvet noises were around 4.9 and 9.6 dB, respectively.

3.2 Discussion

The experimental result shows that the MVN can reduce the ZFC and inaudible components, which also suggests that the proposed idea that controls the $r_2(m)$ was effective. On the other hand, since parameters were determined to guarantee that the power above 100 Hz is approximately flat, parameter optimization for speech synthesis is important for future work. This optimization should be carried out on the basis of not only the power spectrum but also the sound quality of synthesized speech. Therefore, a subjective evaluation by using synthesized speech is also important.

4. Conclusion

We proposed a new noise generation algorithm based on velvet noise. The proposed algorithm can generate a MVN that can reduce unwanted components while maintaining the advantage of the original one. The objective evaluation demonstrated that the MVN was superior to others in the amount of ZFCs.

The next step of this study is to optimize the parameters to implement a speech analysis/synthesis system. A subjective evaluation for comparison between the MVN and white noise is required for this purpose.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H01734 and JP16H05899, and JST PRESTO Grant Number JPMJPR18J8, Japan.

References

[1] H. Dudley, "Remaking speech," J. Acoust. Soc. Am., vol.11, no.2, pp.169–177, 1939.
[2] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," Proc. ICASSP2009, pp.3905–3908, 2009.
[3] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime straight speech manipulation system," Acoust. Science & Technology, vol.28, no.3, pp.140–146, 2007.
[4] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring 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.3-4, pp.187–207, 1999.
[5] 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.E99-D, no.7, pp.1877–1884, July 2016.
[6] M. Morise, "D4C, A band-aperiodicity estimator for high-quality speech synthesis," Speech Communication, vol.84, pp.57–65, 2016.
[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, vol.9, no.5-6, pp.453–467, 1990.
[8] M. Karjalainen and H. Järveläinen, "Reverberation modeling using velvet noise," Proc. AES 30th International Conference, 9-page, 2007.
[9] V. Välimäki, H.-M. Lehtonen, and M. Takanen, "A perceptual study on velvet noise and its variants at different pulse densities," IEEE Trans. Audio, Speech, Language Process., vol.21, no.7, pp.1481–1488, 2013.