Interspeech 2018
2-6 September 2018, Hyderabad

Frequency domain variants of velvet noise and their application to speech processing and synthesis

Hideki Kawahara[1], *Ken-Ichi Sakakibara*[2], *Masanori morise*[3], *Hideki Banno*[4], *Tomoki Toda*[5], *Toshio Irino*[1]

[1]Wakayama University, Japan
[2]Health Science University of Hokkaido, Japan
[3]University of Yamanashi, Japan
[4]Meijo University, Japan
[5]Nagoya University, Japan

kawahara@sys.wakayama-u.ac.jp, kis@hoku-iryo-u.ac.jp, mmorise@yamanashi.ac.jp, banno@meijo-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

We propose a new excitation source signal for VOCODERs and an all-pass impulse response for post-processing of synthetic sounds and pre-processing of natural sounds for data-augmentation. The proposed signals are variants of velvet noise, which is a sparse discrete signal consisting of a few non-zero (1 or -1) elements and sounds smoother than Gaussian white noise. One of the proposed variants, FVN (Frequency domain Velvet Noise) applies the procedure to generate a velvet noise on the cyclic frequency domain of DFT (Discrete Fourier Transform). Then, by smoothing the generated signal to design the phase of an all-pass filter followed by inverse Fourier transform yields the proposed FVN. Temporally variable frequency weighted mixing of FVN generated by frozen and shuffled random number provides a unified excitation signal which can span from random noise to a repetitive pulse train. The other variant, which is an all-pass impulse response, significantly reduces "buzzy" impression of VOCODER output by filtering. Finally, we will discuss applications of the proposed signal for watermarking and psychoacoustic research.

Index Terms: speech processing, speech synthesis, voice excitation source, all-pass filter, voice quality

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 frequency domain variants of velvet noise provide useful candidates for the excitation source signals of synthetic speech and singing [3]. They can replace excitation source signal models [4–7] for VOCODERs [4, 8, 9] and provide a unified design procedure of mixed-mode excitation signals. The proposed frequency variant of the velvet noise is also an impulse response of an all-pass filter [10]. The all-pass filter use of the frequency domain variant provides an effective and easy way for reducing "buzzy" impression of VOCODER speech sounds. The impulse response of the variant is also a TSP (Time Stretched Pulse) and applicable to information hiding for tampering detection. This article introduces the frequency domain variants of velvet noise and discusses their use in speech signal processing including singing and speech synthesis.

2. Background and related work

How to analyze and generate the random component for synthetic voice has been a difficult problem [5, 7, 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]. The reference suggests that two synthetic speech sounds having 20 dB SNR difference provide perceptually equivalent SNR impression 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.

There have been many studies for solving these quality related problems. Multi-band excitation is useful for improving VOCODED sound quality [16]. However, direct mixing of pulse and colored noise still cannot solve the "buzzy" impression problem. Multi-pulse excitation and CELP [17, 18] are also effective for reducing the "buzzy" impression and voice quality enhancement. However, it is not easy to design appropriate multi-pulse for parameter manipulation, which is necessary for VOCODER-based speech conversion.

One efficient method for reducing the "buzzy" impression is to randomize the phase. Group delay manipulation used in legacy-STRAIGHT was successful for reducing this impression [4]. The log domain pulse model (LDPM) also uses phase manipulation [7]. However, such manipulation results smearing of the signal in the time domain. Although time windowing solves this smearing problem, it introduces other problem, power spectral modification due to the statistical fluctuation of the truncated signal. An element signal of the proposed FVN (Frequency domain Velvet Noise) has an excellent time-frequency localization made possible by a six-term cosine series introduced for antialiasing glottal excitation models [19]. Also, because it is an all-pass filter's impulse response, it is free from statistical fluctuation in power spectrum of the processed signal.

The primary focus of this article is to propose FVN and to introduce its prospective applications. We are planning objective and subjective evaluation of FVN in various applications for the next step. Organization of this article is as follows: The next section briefly introduces the original velvet noise. The following section discusses phase modification using shaping functions which are localized both in the time and the frequency domain. Then, applying similar procedure used in the velvet noise to the phase modification introduced in the preceding section to define FVN. The following section discusses three aspects of FVN useful for applications and introduces several representative examples. Finally, we discuss on prospective applications in speech processing as well

10.21437/Interspeech.2018-43

as application to fundamental research on human auditory processing.

3. Velvet noise

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]

The velvet noise allocates a randomly selected positive or negative unit pulse at a random location in each temporal segment [1, 2]. Let T_d represent the average pulse interval in samples. The following equation determines the location of the m-th pulse $k_{\mathrm{ovn}}(m)$. The subscript "ovn" stands for "Original Velvet Noise." It uses two sequences of random numbers $r_1(m)$, and $r_2(m)$ generated from a uniform distribution in $(0, 1)$.

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

where the rounding function $||\bullet||$ returns the nearest integer. The following equation determines the value of the signal $s_{\mathrm{ovn}}(n)$ at discrete time n.

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

With the pulse density higher than 3,000 pulses per second, OVN sounds like a white Gaussian noise and provides a smoother impression. Supplemental media consists of OVN examples.

4. Frequency domain variant of velvet noise

The discrete Fourier transform of a velvet noise sequence closely approximates a complex Gaussian random sequence. The discrete Fourier transform of the filtered velvet noise provides a complex Gaussian noise on the frequency axis with the filter shape weighting. Using the duality of the frequency and the time of Fourier transform, we apply filtered velvet noise to design phase of the all-pass filter. The impulse response of this all-pass filter is the element of the proposed FVN. The element has the temporally localized envelope and random waveform. The key design issue is the shape of the function to manipulate the phase.

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 [19, 20]. This section investigates relations between phase manipulation and the impulse response of the corresponding all-pass filter. Let $w_p(k, B)$ represent a phase modification function on the discrete frequency domain. The following equation provides the complex-valued impulse response $h(n; k_c, B)$ of the all-pass filter.

$$h(n; k_c, B) = \frac{1}{K}\sum_{k=0}^{K-1} \exp\left(\frac{2kn\pi j}{KN} + j w_p(k - k_c, B)\right), \tag{3}$$

where k_c represents the discrete center frequency, and B defines the support of $w_p(k, B)$ in the frequency domain (i.e. $w_p(k, B) = 0$ for $|k| > B$). 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 [19]. The Nuttall's reference [20] provides a list of coefficients of the first three functions and the design procedure. The following cosine series defines these functions. Let define $B_w = B/M$ as nominal bandwidth.

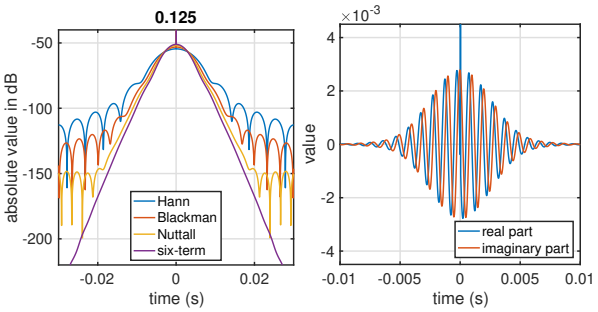

Figure 1: *The absolute value of the impulse response of all-pass filters made from unit phase manipulation using cosine series shapes (left plot). An example impulse response of using the six-term series (right plot).*

$$w_p(k, B) = \sum_{m=0}^{M} a(m)\cos\left(\frac{\pi km}{B}\right), \tag{4}$$

where M represents the highest order of the cosine series.

Figure 1 shows examples of this phase manipulation effects. We found that the six-term cosine series provides the best localization behavior. The six-term series has practically no interference due to sidelobes. We decided to use this six-term series afterward. The coefficients of the six-term series are 0.2624710164, 0.4265335164, 0.2250165621, 0.0726831633, 0.0125124215, and 0.0007833203 from a_0 to a_5. The sidelobes have the highest level of -114 dB and the decay rate of -54 dB/oct.

4.2. Phase manipulation unit allocation using velvet noise

By adding unit phase manipulation $w_p(k - k_c, B)$ on a set of center frequencies k_c obeying the design rule of velvet noise yields the filtered velvet noise in the frequency domain. The following equation defines the allocation index (discrete frequency) $k_c = k_{\mathrm{fvn}}(m)$ where subscript "fvn" stands for Frequency domain Velvet Noise.

$$k_{\mathrm{fvn}}(m) = ||mF_d + r_1(m)(F_d - 1)||, \tag{5}$$

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.

$$\varphi_{\mathrm{fvn}}(k) = \sum_{k_c \in \mathbb{K}} s_{\mathrm{fvn}}(k_c)\left(w_p(k - k_c, B) - w_p(k + k_c, B)\right), \tag{6}$$

$$s_{\mathrm{fvn}}(m) = (2||r_2(m)|| - 1)\,\varphi_{\max} \tag{7}$$

where k spans discrete frequency of a DFT buffer, which has a circular discrete frequency axis and the parameter $\varphi_{\max}$ defines the magnitude of phase manipulation. The second term inside of parentheses of Eq. 6 is to make the phase function have the odd symmetry concerning 0 Hz and $f_s/2$.

The inverse discrete Fourier transform of this all-pass filter provides an impulse response. It is the unit signal $h_{\mathrm{fvn}}(n)$ of the proposed FVN.

$$h_{\mathrm{fvn}}(n) = \frac{1}{K}\sum_{k=0}^{K-1} \exp\left(\frac{2kn\pi j}{KN} + j\varphi_{\mathrm{fvn}}(k)\right). \tag{8}$$

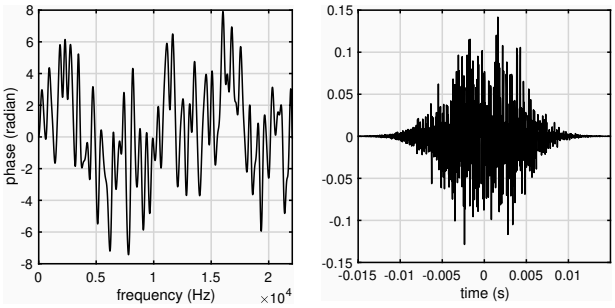

Figure 2: *An example of the phase (left plot) and the corresponding impulse response (right plot) of the designed all-pass filter using the six-term cosine series.*

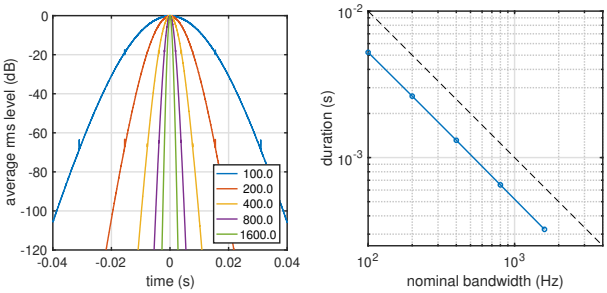

Figure 3: *RMS (Root Mean Squared) value of the impulse response (left plot) and the design parameter B and the duration (right plot).*

Figure 2 shows an example of the designed phase of the all-pass filter and the corresponding impulse response. The impulse response is temporally localized, and the phase behaves like a smoothed random sequence. (This example uses 44,100 Hz sampling frequency, $F_d = 40$ Hz, $B = 200$ Hz, and $\varphi_{\max} = \pi/2$ radian.)

Figure 3 shows the simulation results of generated 5,000 FVN units. The left plot shows the RMS (Root Mean Squared) value of the impulse response. The legend shows the support length B of the smoother $w_p(k, B)$. The right plot shows the relation between the support length B and the duration of the impulse response. These indicate that we can set the desired duration σ_t of the FVN by assigning the B using these simulation results.

$$B = \frac{0.522}{\sigma_t} \quad , \text{ where } \quad \sigma_t^2 = \left\langle t^2 h_{\mathrm{fvn}}^2(t) \right\rangle \tag{9}$$

4.3. Frequency dependent duration control

FVN generated by the procedure in the previous section has a constant temporal duration in each frequency range. It is desirable to introduce frequency-dependent temporal duration, for example, to implement voiced fricatives. This section introduces a variant of FVN which has frequency dependent temporal duration. We call this variant as FFVN (Frequency dependent Frequency domain Velvet Noise).

The duration of the generated FVN is proportional to the frequency width of the smoothing function. It suggests that by locally warping the target frequency axis implements frequency-dependent duration. Let's define a normalized frequency weighting function $g(f) = B_{\mathrm{wtgt}}(f)/B_{\max}$, where $B_{\mathrm{wtgt}}(f)$ represents the target duration at frequency f, and $B_{\max}$ represents the maximum target duration. The following

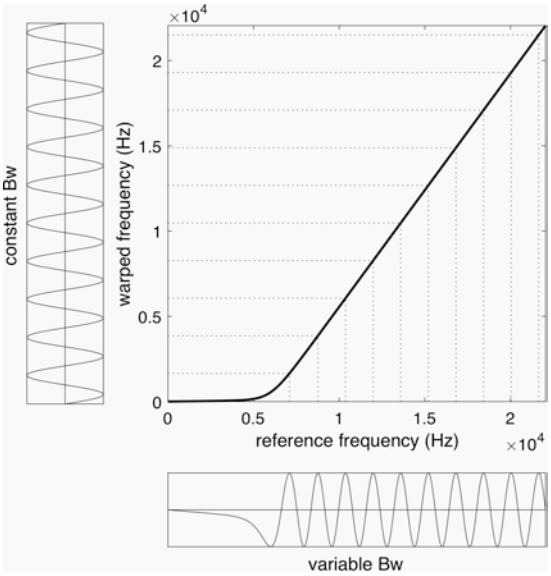

Figure 4: *Implementation of frequency dependent duration control using frequency axis warping. A generated phase function $\varphi_{\mathrm{fvn}}(\nu)$ shown on the left using a constant B is converted to the modified phase function $\varphi_{\mathrm{mod}}(f)$ shown on the bottom, which corresponds to the frequency dependent B.*

equation defines the warped frequency axis $\nu(f)$.

$$\nu(f) = \alpha \int_0^f g(u)du, \tag{10}$$

where α is a calibration coefficient to make $\nu(f_s/2) = f_s/2$ and f_s represents the sampling frequency.

The duration of the FVN on the morphed frequency axis ν is constant under this mapping. The following equation provides the constant duration B_{worg} on this axis.

$$B_{\mathrm{worg}} = B_{\mathrm{wtgt}}(\nu(f_{\max})) \left. \frac{d\nu(f)}{df} \right|_{f=f_{\max}}, \tag{11}$$

Mapping a constant duration FVN's phase function $\varphi_{\mathrm{fvn}}(\nu)$ on the new frequency axis $\nu(f)$ to the original frequency axis f provides the desired variable duration FVN's phase function $\varphi_{\mathrm{mod}}(f)$.

$$\varphi_{\mathrm{mod}}(f) = \varphi_{\mathrm{fvn}}(\nu(f)). \tag{12}$$

Figure 4 illustrates the relations between frequency axes f and ν, and the phase functions $\varphi_{\mathrm{mod}}(f)$ and $\varphi_{\mathrm{fvn}}(\nu)$. The polar form $\exp(j\varphi_{\mathrm{mod}}(f))$ provides the Fourier transform of the unit FVN. This complex exponential function also is the transfer function of an all-pass filter.

The inverse discrete Fourier transform of this all-pass filter with the modified phase function $\varphi_{\mathrm{mod}}(f)$ provides the desired impulse response $h_{\mathrm{mod}}(n)$ of the frequency dependent duration FVN.

$$h_{\mathrm{mod}}(n) = \frac{1}{K} \sum_{k=0}^{K-1} \exp\left(\frac{2kn\pi j}{KN} + j\varphi_{\mathrm{mod}}(k) \right). \tag{13}$$

4.3.1. Implementation

For defining this nonlinear mapping, we introduced a sigmoidal model and a band-wise model. The sigmoidal model $B_{\mathrm{sigm}}(f)$ has two parameters, the transition frequency f_c and transition

width f_{tr}. The band-wise model $B_{\mathrm{band}}(f)$ has two set of parameters, the boundary frequencies $f_b(k)$ $(f_b(0) = 0, \ldots, f_b(K) = f_s/2)$ and the durations $B_k, (k = 1, \ldots, K)$ in the bands. It also has the additional parameter, the width f_w of a raised cosine smoother $s(f; f_w) = (1 + \cos(2\pi f/f_w))/2$ with its support $[-f_w/2, f_w/2]$.

$$B_{\mathrm{sigm}}(f) = \frac{1}{1 + \exp\left(-\frac{f-f_c}{f_{tr}}\right)} \tag{14}$$

$$B_{\mathrm{band}}(f) = s(f; f_w) * \sum_{k=1}^{K} B_k(u_{k-1}(f) - u_k(f)), \tag{15}$$

where "$*$" represents convolution and $u_k(f)$ represents the unit step function starting at $f_b(k)$.

5. Application

This section introduces two applications of an element of FVN and FFVN. The first one is for post-processing of VOCODER output, and the other is data augmentation for training, for example, WaveNet. We also introduce the application of FVN and FFVN to the excitation source of synthetic speech.

5.1. Time invariant all-pass filter

Each element waveform of FVN and FFVN is an impulse response of an all-pass filter. Applying this all-pass filter to the VOCODER outputs reduces their "buzzy" impression significantly. This filtering is simple and effective post-processing for improving the quality of VOCODER outputs. Supplemental media files provide demonstrations of this "buzzy" impression reduction.

Applying this all-pass filter to an original speech sample alters the waveform significantly. However the original and the filtered speech sound similar when the duration of the filter is small (for example about 1 ms: consistent with [21]). This insensitivity suggests that FVN and FFVN can be useful for data-augmentation for training, for example, WaveNet to embed constraints due to human auditory perception.

5.2. Excitation of synthetic speech

Linear interpolation of the phase of two FVN or FFVN elements morphs the generated signal seamlessly. A regularly repeated sequence of an identical FVN or FFVN element, in other words using frozen element, provides clear pitch perception. When the elements are updated using different random number always, and the duration of each element is longer than the repetition period, the sequence sounds like white noise. Linear interpolation of the phase of the elements of these sequences provides an excitation signal which spans from noise to periodic sounds seamlessly. Supplemental media files provide demonstrations of this morphed excitation signal.

The other application of FVN and FFVN is for additive noise component of the excitation source. Allocating an element with relatively short duration (for example, shorter than 5 ms) in each excitation pitch period implements the temporal variation of the random component. This implementation is effective for synthesizing low pitched voices, such as males'. Supplemental media files also provide demonstrations of the temporal variation of the random component. Instead of using a Gaussian noise, substituting it with FVN and FFVN in statistical signal processing [14] is an exciting possibility. Using them in a complex cepstrum-based excitation model [22] is such a prospective example.

5.3. Supplemental media files

Supplemental media files of this article consists of the OVN samples, and FVN and FFVN application examples. The VOCODED speech example uses the file "`slt/arctic_b0436.wav`" of CMU Arctic database [23] and synthesized using Mel-Cepstrum processed envelope. It also consists of the link to MATLAB script and resources [24].

6. Discussion and related future work

The proposed variants have wide potentials in applications. High-quality and wide frequency-range recorded speech sounds have very high kurtosis in amplitude distribution [6]. The proposed all-pass filter reduces the kurtosis of the filtered signal significantly. Since the FVN unit response is one specific type of TSP (Time Stretched Pulse), the convolution of the processed signal with the time-reversed version recovers the original version and consequently the high kurtosis level. This recovery is a useful feature for information hiding and tampering detection [25].

OVN and FVN also provide a set of tools for investigating perceptually equivalent timbre class and fundamental properties of the human auditory system. Effects of phase on timbre were well known [26], but the structure of phase-related timber was only partially investigated [27]. The flexibility of FVN parameter design will open a new systematic research paradigm in auditory processing of signal phase and will provide means to revisit fundamental questions. The apparent contradiction between auditory evoked potential and perception of the optimized chirp signal [28, 29] is an example of such questions. The answer to the question will provide the fundamental solution to the "buzzy" impression of VOCODED sounds. OVN also is useful. For example, the memory span of so-called echoic memory, around 1 to 2 seconds, was tested using random signals [30]. Testing periodicity perception using very sparse repeated OVN samples will shed new light on their information representation and processing. The sparse repeated OVN may serve as a complement test signal to the IRN (Iterated Rippled Noise) [31] used in psychoacoustic experiments.

It is crucially important to design systems dealing with human speech communication based on fundamental understanding of human auditory perception mechanisms and their functions [32] because the end-users of such systems are humans. Data augmentation by introducing physically different while perceptually equivalent preprocessed speech will be one feasible strategy to introduce such understanding built into end-to-end speech applications [15, 33, 34]. All-pass filtering based on FVN provides a prospective tool for required data augmentation.

7. Conclusions

We introduced a flexible excitation source signal which spans from a periodic signal to random signal seamlessly and an all-pass filter which substantially reduces "buzzy" impression of VOCODER outputs. Combination of the well-behaving phase manipulation function and the velvet noise generation procedure in the frequency domain made these important contributions possible. We are planning to introduce this excitation source signal to reformulate perceptually isomorphic VOCODER framework. We also make software of FVN variants and reference applications available as an open-source package on GitHub.

8. Acknowledgements

This work was supported by JSPS KAKENHI (grants in aids for scientific research) Grant Numbers JP15H03207, JP15H02726 and JP16K12464.

9. References

[1] H. Järveläinen and M. Karjalainen, "Reverberation modeling using velvet noise," in *AES 30th International Conference, Saariselkä, Finland*. Audio Engineering Society,, 2007, pp. 15–17.

[2] V. Välimäki, H. M. Lehtonen, and M. Takanen, "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, July 2013.

[3] H. Kawahara, "Application of the velvet noise and its variant for synthetic speech and singing," *IPSJ SIG Technical Report*, vol. 2018-MUS-118, no. 3, 2018.

[4] 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," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

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

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

[7] G. Degottex, P. Lanchantin, and M. Gales, "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, Jan 2018.

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

[9] 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, no. 7, pp. 1877–1884, 2016.

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

[11] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "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, jan 1998.

[12] N. Malyska and T. F. Quatieri, "Spectral representations of nonmodal phonation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 34–46, 2008.

[13] J. Skoglund and W. B. Kleijn, "On time-frequency masking in voiced speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 361–369, 2000.

[14] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

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

[16] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[17] B. Atal and J. Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7. IEEE, 1982, pp. 614–617.

[18] M. Schroeder and B. Atal, "Code-excited linear prediction (celp): High-quality speech at very low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, vol. 10. IEEE, 1985, pp. 937–940.

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

[20] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Audio Speech and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.

[21] J. Blauert and P. Laws, "Group delay distortions in electroacoustical systems," *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1478–1483, 1978.

[22] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4581–4584.

[23] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[24] H. Kawahara, "Resource page for Interspeech 2018," (Last access: 10/June/2018). [Online]. Available: http://www.wakayama-u.ac.jp/%7ekawahara/IS2018/

[25] P. Jayaram, H. Ranganatha, and H. Anupama, "Information hiding using audio steganography–a survey," *The International Journal of Multimedia & Its Applications (IJMA) Vol*, vol. 3, pp. 86–96, 2011.

[26] R. Plomp and H. J. Steeneken, "Effect of phase on the timbre of complex tones," *The Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 409–421, 1969.

[27] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *The Journal of the Acoustical Society of America*, vol. 82, no. 5, pp. 1560–1586, 1987.

[28] T. Dau, O. Wegner, V. Mellert, and B. Kollmeier, "Auditory brainstem responses with optimized chirp signals compensating basilar-membrane dispersion," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1530–1540, 2000.

[29] S. Uppenkamp, S. Fobel, and R. D. Patterson, "The effects of temporal asymmetry on the detection and perception of short chirps," *Hearing research*, vol. 158, no. 1-2, pp. 71–83, 2001.

[30] N. Guttman and B. Julesz, "Lower limits of auditory periodicity analysis," *The Journal of the Acoustical Society of America*, vol. 35, no. 4, pp. 610–610, 1963.

[31] W. A. Yost, R. Patterson, and S. Sheft, "A time domain description for the pitch strength of iterated rippled noise," *The Journal of the Acoustical Society of America*, vol. 99, no. 2, pp. 1066–1078, 1996.

[32] R. F. Lyon, *Human and Machine Hearing: Extracting Meaning from Sound*. Cambridge University Press, 2017.

[33] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast High-Fidelity Speech Synthesis," *ArXiv e-prints*, Nov. 2017.

[34] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *ArXiv e-prints*, Dec. 2017.