Revisiting spectral envelope recovery from speech sounds generated by periodic excitation

Hideki Kawahara*, Masanori Morise†, and Kanru Hua‡
* Wakayama University, Wakayama, Japan
E-mail: kawahara@sys.wakayama-u.ac.jp
† University of Yamanashi, Kofu, Japan
E-mail: mmorise@yamanashi.ac.jp
‡ University of Illinois at Urbana-Champaign, Urbana, IL, USA
E-mail: khua5@illinois.edu

***Abstract*—We propose a set of new accurate spectral envelope recovery methods for speech sounds generated by periodic excitation based on a set of interference-free power spectrum representations. The proposed methods outperform our previous spectral recovery models used in legacy-STRAIGHT, TANDEM-STRAIGHT and WORLD VOCODERs. We introduce several design procedures of paired time widows which remove interferences caused by signal periodicity in the time domain or in both time and frequency domains. In addition to this interference-free representation, we introduce post and pre-processing to improve recovery accuracy around spectral peak regions. We conducted a set of evaluation tests using voice production simulator and natural speech samples. Finally, we discuss the application of the proposed method on revising high-quality VOCODERs.**

I. Introduction

Spectral envelope recovery (or estimation depending on applications) from observed speech signals is a crucial problem in speech processing applications. We have developed VOCODERs (legacy-STRAIGHT[1], TANDEM-STRAIGHT[2], and WORLD[3]) based on power spectral representations, which remove interferences caused by signal periodicity. We refer to these as "interference-free" representations afterward. We introduce a new set of spectral envelope recovery algorithms by revisiting interference-free power spectrum representations of periodic signals. The proposed algorithms outperform our previous ones for extracting interference-free representations in the accuracy of recovery and tolerance to errors in system parameters.

II. Background and related work

Spectral envelope recovery from observed speech signals has been an important research topic in speech science. The source-filter model of speech production[4] provides a forward model which is a reasonable and straightforward approximation of non-linear and complex mechanism[5]. VOCODERs use this simple model for solving the inverse problem of parametric representation of speech sounds and generate sounds using the forward model. LPC[6], [7] and homomorphic filtering (cepstrum)[8] are representative examples.

Contrary to the general belief on waveform-coding's supremacy over VOCODERs several decades ago[9], [10], legacy-STRAIGHT introduced an example that VOCODER without approximating original waveform can provide high-quality synthetic speech sounds. It interpreted the periodic excitation in voiced sounds a strategy to convey the underlying smooth time-frequency surface by systematic sampling and tried to design procedure to recover the original surface[1]. In a retrospective view, spectral envelope recovery used in legacy-STRAIGHT applied the concept behind the consistent sampling theory[11], [12]. TANDEM-STRAIGHT also uses the same concept for designing the procedure[2], [12]. The interference-free representation of WORLD uses a unique condition associated with Hann window to reduce the computational demand significantly[13].

Sinusoidal models[14], [15] and their recent versions[16], [17] also provide high-quality synthetic speech sounds. Although they are essentially waveform-based (phase conscious) models, spectral envelope recovery is important for flexible manipulations[18]. Usually, the spectral envelope for sinusoidal models uses the true envelope method[19] based on mel-cepstrum representation. However, as far as the spectral model is for flexible manipulation, STRAIGHT, WORLD, LPC, and refined LPC (for example, DAP takes periodicity into account in parameter estimation[20]) spectra apply to sinusoidal models.

Introduction of WaveNet[21] and huge success in various speech processing problem[22], [23], [24] seems to make VOCODERs outdated. Flexible audio manipulation using latent variables based on WaveNet is also a significant contribution[25]. However, we believe that filling the gap between these latent representations and simple source-filter models will significantly facilitate flexibility and visibility of speech processing applications.[1]

Reliable spectral representations have critical importance to make this gap-filling approach successful. The spectral envelope recovery algorithms used in current VOCODERs use heuristics in their infrastructure, algorithms to extract interference-free representations[1], [2], [12], [13]. Strongly parameterized representations, such as LPC, LPC-cepstrum, and true-envelope models[6], [7], [8], [19], introduce errors due to model mismatch.

In this respect, the proposed procedures are hybrid of model-based ones and non-parametric ones. We are going to

[1]The first author learned a lot from making an interactive speech production simulator of an education and research tools for speech science (SparkNG)[26]. He also was inspired by a very intuitive and instructive voice production simulator[27].

revise VOCODER-based flexible speech processing tools[28], [29], [30] by replacing the critical infrastructure, spectral envelope recovery procedures, with the proposed procedures.

III. Problems to be solved

Our goal is to develop algorithms to recover the (preferably smooth) time-frequency surface $P(f,t)$ from generated signals. The signal can be a sum of harmonic sinusoids with instantaneous amplitudes read from the surface. It can also be a response to a repetitive excitation of the time-varying filter made from the surface[1]. A spectrogram calculated using short-term Fourier transform has periodic interferences caused by the periodicity. Removing these interferences is the first goal.

Our second goal is to apply interference-free representations to voiced speech. The underlying surface of voiced speech production has constraints on the spectral shape[4]. It also consists of complex effects caused by non-linear and acoustic-mechanical interactions in phonation mechanisms[5].

We introduce two different methods to attain the first goal. The first one is an interference-free representation both in time and frequency using one staged algorithm. It uses a pair of windowing functions made from series of trigonometric functions and numerical optimization. The second one is a two-stage algorithm. It calculates the temporally interference-free representation and removes interference in the frequency domain. The analysis parameters of this approach do not require numerical optimization. We provided closed-form solutions for parameter setting.

We propose an analysis pipeline to attain the second goal. First, spectrum whitening filter based on autoregressive spectral model equalizes the input signal. Then, the two staged procedure calculates the interference-free representation of the preprocessed input signal. Finally, combining the time-varying whitening filter shape and the calculated interference-free representations yields the time-frequency surface specialized for voiced speech sounds.

The organization of this paper is as follows. First, we introduce interference-free representations which attain the first goal. It describes the one-stage algorithm followed by the two-stage algorithm. It also introduces numerical behavior of these methods. Second, we introduce the pipeline which attains the second goal. It presents examples using natural speech analyses, followed by preliminary evaluation results comparing with our previous methods, legacy-STRAIGHT, TANDEM-STRAIGHT, and WORLD. Finally, we discuss the application of the proposed representations and future issues.

IV. Interference-free representations

In this section, we first simplify the requirement for the interference-free representations, as follows. Let f_o and t_o represent the fundamental frequency and the fundamental period. We impose constraints on the deviation of the estimated/recovered spectral envelope from the truth.

1) Deviation has to be minimum in a time-frequency patch $(nt_\mathrm{o} < t < (n+1)t_\mathrm{o})$ and $(kf_\mathrm{o} < f < (k+1)f_\mathrm{o})$, where t represents time and f represents frequency.
2) Deviation has to be insensitive to relative phase differences between harmonic components.

We do not directly set criteria for side-lobe levels and the time and the frequency resolution of the window functions. The requirement above implicitly embodies these criteria.

A. Interference-free in both time and frequency domain

Power spectra of a periodic pulse train calculated using a Hann window and a window consisting of one cycle of sinusoid yield a constant-valued power spectrum when added together. However, this pair does not behave well for periodic signals with random phase and random level harmonic components. In this section, we introduce a set of paired windowing functions and optimize them based on the requirement mentioned before.

1) Pair of windows: The set of windows for calculating power spectra have the following form:

$$w_r(t) = \sum_{k=0}^{K} a_k \cos(\pi k t) \tag{1}$$

$$w_i(t) = \sum_{k=1}^{K} b_k \sin(\pi k t), \tag{2}$$

where t represents the normalized time and the support of the functions $w_r(t)$ and $w_i(t)$ is $[-1,1]$.[2] We set $a_0 = 1$ to normalize the following derivation. Their frequency domain representations $W_r(f)$ and $W_i(f)$ are as follows:

$$W_r(f) = \mathrm{sinc}(f)a_0 + \frac{1}{2}\sum_{k=1}^{K}(\mathrm{sinc}(f-k) + \mathrm{sinc}(f+k))a_k \tag{3}$$

$$W_i(f) = \frac{1}{2j}\sum_{k=1}^{K}(\mathrm{sinc}(f-k) - \mathrm{sinc}(f+k))b_k, \tag{4}$$

where we use the following definition.

$$\mathrm{sinc}(x) \equiv \frac{\sin(\pi x)}{\pi x} \tag{5}$$

The spectrogram $P(f,t)$ consisting of interference-free power spectrum is the power sum of the spectrograms $P_r(f,t)$ and $P_i(f,t)$ calculated using $w_r(t)$ and $w_i(t)$ for windowing respectively.

$$P(f,t;\Theta_c) = P_r(f,t;\Theta_r) + P_i(f,t;\Theta_i), \tag{6}$$

where $\Theta_c = \Theta_r \bigcup \Theta_i$ represents the set of parameters which define windowing functions.

2) Cost function: For the normalized frequency f and the normalized time t (normalized by the fundamental frequency f_o and the fundamental period t_o respectively), [3] a time-frequency domain $\mathbb{S} = (m < t < m+1,\ k < f < k+1)$ defines the area to evaluate the cost. Without losing generality, we set $m = 0$ and $k = 0$. In the following evaluation, we use

[2] We normalized the time axis to make the window length 2 to simplify the equations. The length of the window in the actual time domain depends on the order K and the fundamental period t_o. The actual window length is $(K+1)t_\mathrm{o}$.

[3] We use f_o instead of using the conventional symbol f_0 according to the recommendation[31].

the squared deviation of the spectrogram of periodic signals from its average.

$$L^2(\Theta) = \int_{(t,f\in\mathbb{S})} \left| P(f,t;\Theta) - \overline{P(f,t;\Theta)} \right|^2 df dt, \quad (7)$$

where the system/signal parameter consists of the following contents.

$$\Theta = \left\{ \{a_k\}_{k=1}^K, \{b_k\}_{k=1}^K, f_o, f_c, \{\varphi_n\}_{n=-N}^{N-1} \right\}, \quad (8)$$

where f_c represents the assumed fundamental frequency in the analysis step. A set of values $\{\varphi_n\}_{n=-N}^{N-1}$ represents the initial phase values of related harmonic components. The constant N represents the number of harmonic components considered in the evaluation of the cost function.

3) Optimization for pulse train: Minimization of the cost function requires numerical optimization. First, the simplest case is to use a periodic pulse train as the input and assumes that the fundamental frequency is known. Then, parameters to be optimized are $\{a_k\}_{k=1}^K, \{b_k\}_{k=1}^K$. We used a nonlinear optimization procedure built-in MATLAB, which does not require explicitly calculating derivatives of the cost function.

Parameters for $K = 1, 2, 3, 4$ are as follows.
$[a_1, b_1] = [1.0000, 0.9998]$,
$[a_1, a_2, b_1, b_2] = [1.5182, 0.4607, 1.0172, 0.6793]$,
$[a_1, a_2, a_3, b_1, b_2, b_3] =$
$[1.7983, 0.9999, 0.1852, 0.7202, 1.0002, 0.4622]$,
$[a_1, \ldots, a_4, b_1, \ldots, b_4] =$
$[1.8470, 1.3733, 0.5980, 0.0770, 0.7385, 1.0896, 0.7539, 0.1925]$,

Figures 1 shows the shape of the optimized windows. In each plot, the solid blue line represents $w_r(t)$, and the solid red line represents $w_i(t)$ the thin yellow line represents the envelope, in other words, $|w_r(t) + jw_i(t)|$.

Figures 2 shows the gain of the optimized windows. In each plot, the solid blue line represents $g_r(f)$, and the solid red line represents $g_i(f)$ the thin yellow line represents the power sum of each gain, in other words, $|g_r(f) + jg_i(f)|$.

The standard deviation from the target spectrum using the optimized pair of windows of order 3 is less than 0.01 dB for randomized initial harmonic phase (uniform distribution in $(0, 2\pi)$ range.). However, the preceding and trailing ripples in the temporal envelope of Fig. 1, especially in order 4 plot, will be disturbing for audio signals. Because out auditory system has a wide dynamic range and the ripple level -20 dB is easily audible. Also, the error in f_o assumption introduces spectral deviations proportional to the amount of error. These motivated the next alternative strategy. It is the two staged procedure for deriving interference-free representations.

(Figure 3 is a placeholder. I will find relevant figure.)

B. Interference-free in the time domain and post processing

We introduce two staged algorithms to calculate other interference-free representations. In the first stage, it calculates temporally interference-free power spectrum. The second stage removes periodic variation in the frequency domain. This method uses one windowing function and frequency derivative of the short-term Fourier transform using the windowing function.

In the first stage, the temporally interference-free power spectrum $P_{\mathrm{TIF}}(\omega)$ is a weighted sum of a power spectrum

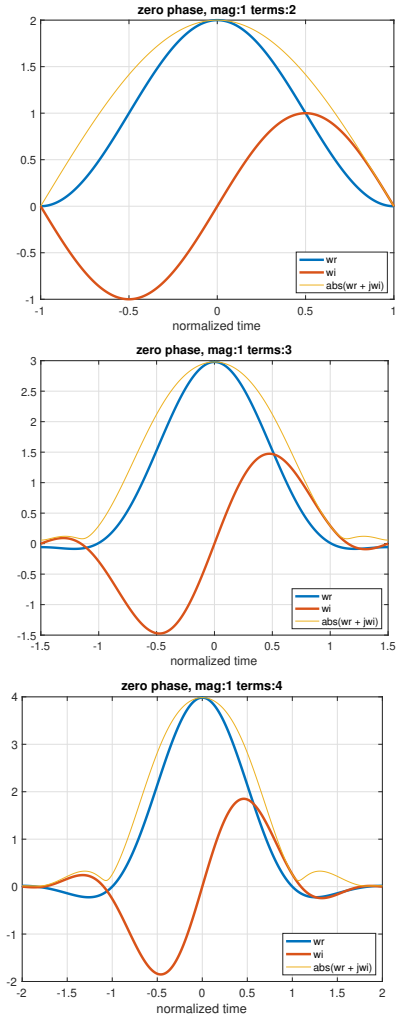

Fig. 1. The optimized pair of window shapes and the root of the squared sum. The orders are 2, 3, and 4. The time axis is normalized by the fundamental period t_o

$P(\omega)$ and a newly introduced power spectrum $P_a(\omega)$. The newly introduced $P_a(\omega)$ (we call it associated power spectrum) behaves complementary to the first power spectrum. Based on preliminary investigations, we found that the squared absolute value of (angular) frequency derivative of a short-term Fourier transform $S(\omega)$ provides the desired behavior. The following equations summarize this:

$$P_{\mathrm{TIF}}(\omega) = P(\omega) + c_f^2 P_a(\omega) \quad (9)$$

$$= |S(\omega)|^2 + c_f^2 \left| \frac{dS(\omega)}{d\omega} \right|^2, \quad (10)$$

where c_f^2 represents the mixing weight of power spectra. Note that this mixing coefficient does not need tuning. We derived closed-form representations for determining c_f for each type of windowing function.

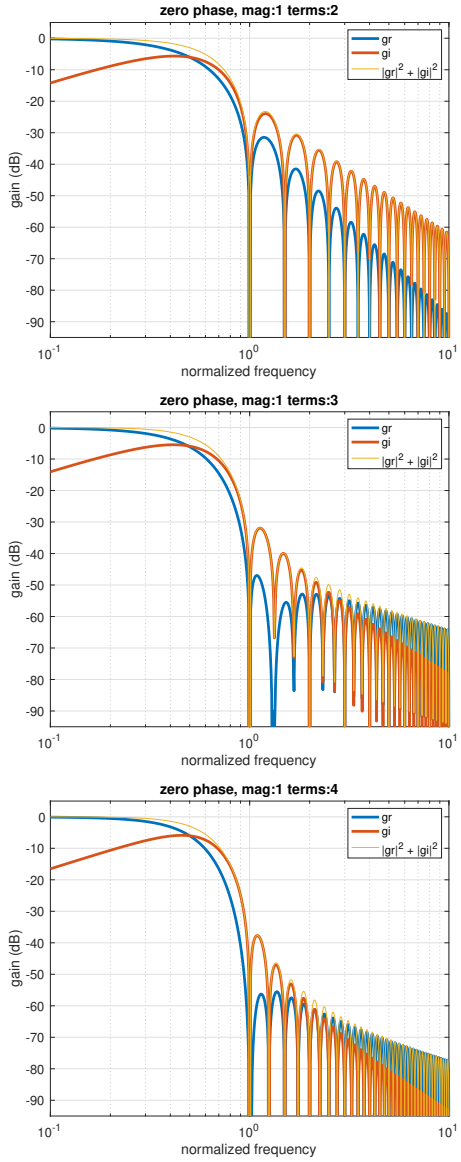

Fig. 2. The gain functions of the optimized pari of windows. The orders are 2, 3, and 4.

We selected two types of windowing functions for this approach. One is B-spline functions. The other is a combination of a trigonometric function and B-spline functions. We designed these functions to have zeros at harmonic frequencies.

1) B-spline based functions[32]: The B-spline expansion of $s(x)$ is defined below:

$$s(x) = \sum_{k \in \mathbb{N}} c(k) \beta^n (x - k), \tag{11}$$

where $\mathbb{N}$ represents the set of natural number and $\beta^{(n)}(x)$ represents the n-th B-splines.

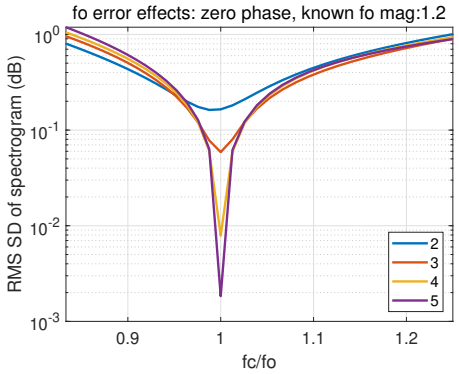

Fig. 3. RMS standard deviation error on assumed f_o error. Left plot shows the optimized parameters for regular pulse train analysis. The right plot shows that for fully scanned harmonic signals.

The n-th B-spline is defined as an extension of the 0-th spline.

$$\beta^0(x) = \begin{cases} 1, & -\frac{1}{2} < x < \frac{1}{2} \\ \frac{1}{2}, & |x| = \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$\beta^{(n)}(x) = \underbrace{\beta^0 * \beta^0 * \cdots * \beta^0(x)}_{(n+1) \text{ times}}. \tag{13}$$

The Fourier transform of B-splines are as follows:

$$\hat{\beta}^{(n)}(\omega) = \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}} \right)^{n+1} \tag{14}$$

In the first order approximation, the interfering component is from the neighboring harmonic component. This interference in the power spectrum varies sinusoidally. Using the identity $\sin^2\theta + \cos^2\theta = 1$ removes this temporal variation. Because the level in the middle of harmonic component is the sum of both contribution, in other words, the sum of complex exponentials, the slope at that point provides orthogonal variation. The frequency derivative of the gain function is what we need. By using Fourier transform theories, it is equivalent using an associated window which is the product of the original window and the time axis t.

$$\frac{d\mathcal{F}[g(t)]}{d\omega} = tg(t) \tag{15}$$

It also has the following form:

$$\frac{d}{d\omega}\left[\left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}} \right)^{n+1} \right] = \frac{(n+1)(\frac{\omega}{2}\cot(\frac{\omega}{2}) - 1)\left(\frac{\sin(\frac{\omega}{2})}{\frac{\omega}{2}} \right)^{n+1}}{\omega} \tag{16}$$

By adjusting the gain at $\omega/2 = \pi/2$ to that of the original B-splines, the sum of power spectra calculated using these windows temporally static.

Table I provides the gain and the calibration coefficients to make the sum of the power spectra temporally static. Note that the frequency derivative can be calculated numerically from the Fourier transform of the original B-splines.

2) Combination of a trigonometric function and B-spline functions: The next useful set of functions starts from a half-cycle of the trigonometric function.

$$v^{(0)}(x) = \begin{cases} \cos(\pi x) & -\frac{1}{2} \le x \le \frac{1}{2} \\ 0, & |x| > \frac{1}{2} \end{cases} \tag{17}$$

$$v^{(n)}(x) = v^{(0)}(x) * \underbrace{\beta^0 * \beta^0 * \cdots * \beta^0(x)}_{n \text{ times}} \tag{18}$$

The frequency domain representations of these are as follows.

$$\hat{v}^{(n)}(\omega) = \frac{2\pi \cos(\frac{\omega}{2})}{\pi^2 - \omega^2} \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}} \right)^n \tag{19}$$

For later use, it is convenient to have gain 1 at frequency zero. It yields the following:

$$\hat{v}^{(n)}(\omega) = \frac{\pi^2 \cos(\frac{\omega}{2})}{\pi^2 - \omega^2} \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}} \right)^n \tag{20}$$

The sum of power spectra calculated using the original and the associated windowing functions provides a temporally static power spectral representation. This summation needs the appropriate mixing coefficients.

The following equation shows the frequency derivative of the gain of the associated windowing function:

$$\begin{aligned} \frac{d\hat{v}^{(n)}(\omega)}{d\omega} &= \frac{d}{d\omega}\left[\frac{\pi^2 \cos(\frac{\omega}{2})}{\pi^2 - \omega^2} \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}} \right)^n \right] \\ &= \frac{\pi^2 2^{n-1} \left(\frac{\sin\left(\frac{\omega}{2}\right)}{\omega} \right)^{n+1} \left(n\left(\pi^2 - \omega^2\right)\omega \cot^2\left(\frac{\omega}{2}\right) + \left(2n\omega^2 - 2\pi^2 n + 4\omega^2\right)\cot\left(\frac{\omega}{2}\right) + \omega^3 - \pi^2\omega \right)}{\left(\pi^2 - \omega^2\right)^2} \end{aligned} \tag{21}$$

TABLE I
CALIBRATION TABLE FOR MAKING SUM OF POWER SPECTRA TEMPORALLY STATIC. THE GAINS ARE AT $\omega = \pi$. NOTE THAT THE ORDER n REPRESENTS THE ORDER OF B-SPLINE AND DOES NOT REPRESENTS THE EXPONENT.

order(n)	$\beta^{(n)}(\pi)$	$\left.\frac{d\beta^{(n)}(\omega)}{d\omega}\right\|_{\omega=\pi}$	$c_f(n)$
0	-3.9224	-13.8654	9.9430
1	-7.8448	-11.7672	3.9224
2	-11.7672	-12.1678	0.4006
3	-15.6896	-13.5914	-2.0982

(gain: dB)

TABLE II
CALIBRATION TABLE FOR MAKING SUM OF POWER SPECTRA TEMPORALLY STATIC FOR THE CONVOLUTION OF TRIGONOMETRIC FUNCTION AND 0-TH B-SPLINES. THE GAINS ARE AT $\omega = \pi$.

order(n)	$\hat{v}^{(n)}(\pi)$	$\left.\frac{d\hat{v}^{(n)}(\omega)}{d\omega}\right\|_{\omega=\pi}$	$c_f(n)$
0	-2.0982	-18.0618	15.9636
1	-6.0207	-12.4418	6.4211
2	-9.9431	-11.9273	1.9841
3	-13.8656	-12.9271	-0.9384

(gain: dB)

This equation provides the calibration coefficients to make the sum of the power spectra temporally static. Table II shows the resulted coefficients.

3) Removing frequency domain interference: These temporally static power spectral representations have periodic variation in the frequency domain. Using convolution with the 0-th spline of the width of the fundamental frequency f_o removes this periodic variation in the frequency domain.

$$P^{(n)}_{\mathrm{TFIF}}(\omega) = \frac{1}{\omega_o} \int_{-\omega_o/2}^{\omega_o/2} P^{(n)}_{\mathrm{TIF}}(\omega + \nu) d\nu \tag{22}$$

$$= \frac{1}{\omega_o} \left(U^{(n)}_{\mathrm{TIF}}(\omega + \omega_o/2) - U^{(n)}_{\mathrm{TIF}}(\omega - \omega_o/2) \right) \tag{23}$$

$$U^{(n)}_{\mathrm{TIF}}(\omega) \equiv \int_{-\omega_o}^{\omega} P^{(n)}_{\mathrm{TIF}}(\nu) d\nu \tag{24}$$

This smoothing operation has a side effect, over-smoothing. A digital filter on the discrete frequency axis compensates for this over-smoothing, based on the consistent sampling theory[11].

4) Numerical results: This section illustrates the behavior of the algorithms for B-spline based method and the method using the combination of a trigonometric function and B-spline functions (trigo-splines afterward)

Figure 4 shows the standard deviation of the spectrogram from the target spectrogram. The test signal is a periodic signal with initial phase randomization of each harmonic component. The horizontal axis shows the order of the function. The upper plot shows the results without post-processing of frequency domain smoothing. The lower plot shows that with the post-processing. These plots reveal the needs and effectiveness of the post-processing. It is also interesting to note that for the order 1, the trigo-spline function provides good performance even without the post-processing.

Figure 5 shows the standard deviation caused by the error of the assumed f_o. These results indicate that the B-spline-based method and the convolution of trigonometric and B-splines are relatively tolerant using order 2 functions.

Figure 6 shows the systematic error due to the harmonic level difference. In this simulation, a uniform distribution (in dB) gave each harmonic level. The results suggest

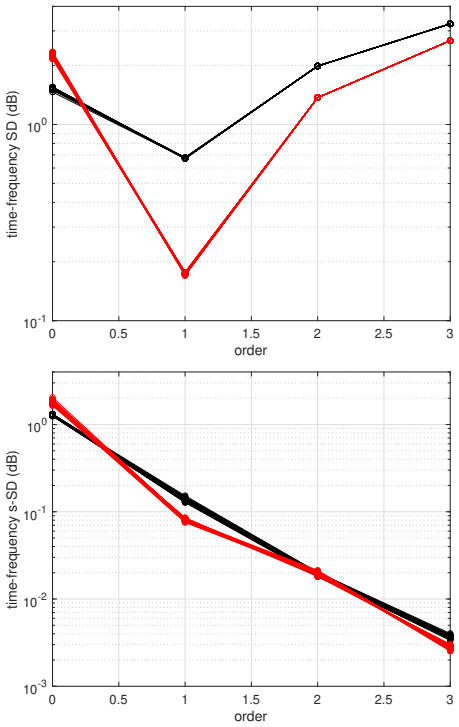

Fig. 4. Standard deviation of the spectrograms in the time-frequency plane, as a function of the order n. The test signal is a periodic signal with initial phase randomization of each harmonic component. The black lines represent the results of B-splines and the red lines represent the results of trigo-splines. The upper plot shows the results without frequency domain processing. The lower plot shows that with post processing.

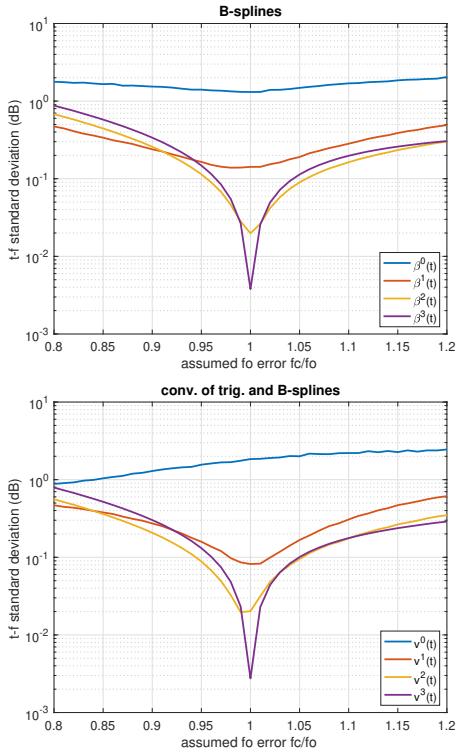

Fig. 5. Tolerance to assumed fo error. Left plot shows results using B-splines. Right plot shows results using the convolution of trigonometric function and B-splines.

that by equalizing harmonic levels before calculating the interference-free representations, the digital filter for compensating over-smoothing effect is not needed at least for the 2nd and 3rd order windowing functions. By adding back the amount of equalization to the resulted interference-free representations yields the desired representation. This equalization and adding back for pre- and post-processing is the main idea to attain the second goal mentioned in the beginning.

V. Interference-free representation for voiced sounds

Application of these interference-free representations for voiced sounds needs further refinement. The one-stage method behaves as interpolation of harmonic levels using sinc function and its exponentials. The two-stage method behaves like the nearest neighbor interpolation of harmonic levels with smoothing. When the underlying target representation is flat in the time, and the frequency domain, procedures introduced in the previous sections section behave similarly well. However, the underlying spectral envelope of voiced sounds has sharp peaks at vocal tract resonance frequencies. Direct application of the procedures for calculating interference-free representations smears out these peak shapes and this spectral smearing results in quality degradation when used in VOCODERs[12], [33].

As mentioned in the last section, pre-whitening of the speech signal before calculating interference-free representations and adding back afterward alleviates this discrepancy. This pipeline procedure is the answer to the second goal. We introduce the lattice type pre-whitening filter to reduce disturbing transient caused by parameter update for handling time-varying systems. The lattice filter parameters are reflection coefficients based on LPC analysis.

VI. Natural speech examples and simulation

This section shows analysis examples using natural speech sounds followed by simulation tests to verify the findings. The speech spectral envelope recovery from natural speech example consists of steps described in Appendix A

The test script also shows the magnified spectrograms and the spectral slice views. The slice consists of the raw power spectrum, the time-frequency interference-free power spectrum, and the recovered power spectrum. The NDF fundamental frequency extractor[34] provided the assumed f_{o} for each frame.

Figure 7 shows the spectrogram recovered using the pipeline mentioned before. The signal is a Japanese vowel sequence /aiueo/ spoken by a male speaker. The sampling frequency is 22,050 Hz and 16 bit quantization.

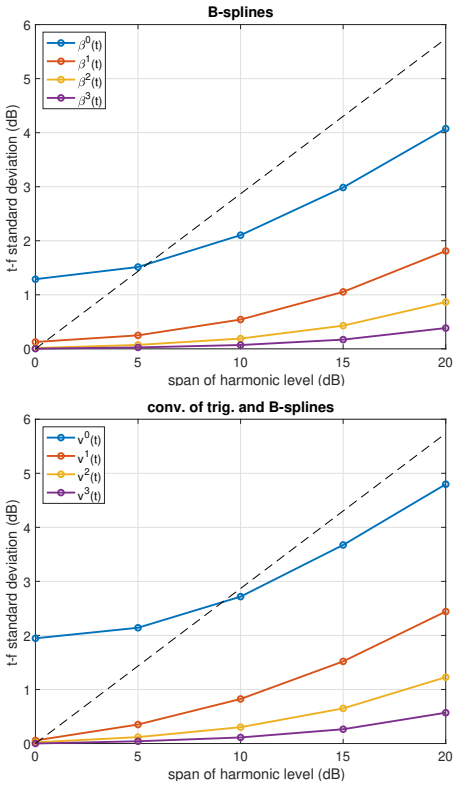

Fig. 6. Standard deviation of harmonic level error as a function of level distribution. Each harmonic level (in dB) distributes in uniform distribution with the width represented in the horizontal axis. The dashed line represents the standard deviation of the uniform distribution. Left plot shows results using B-splines. Right plot shows results using the convolution of trigonometric function and B-splines.

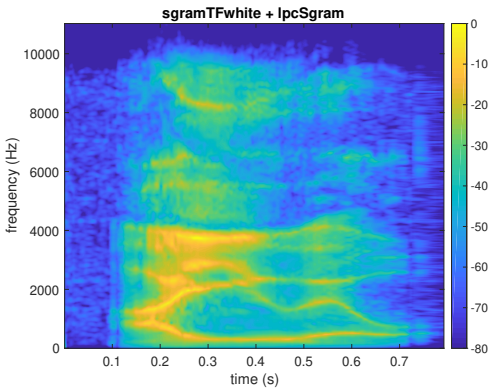

Fig. 7. Recovered spectrogram using the pipeline.

To illustrate the role of each procedure, we prepared magnified views of the beginning part of the spectrogram. Figure 8 shows the magnified view of the recovered spectrogram.

Figure 9 shows several illustrative representations with

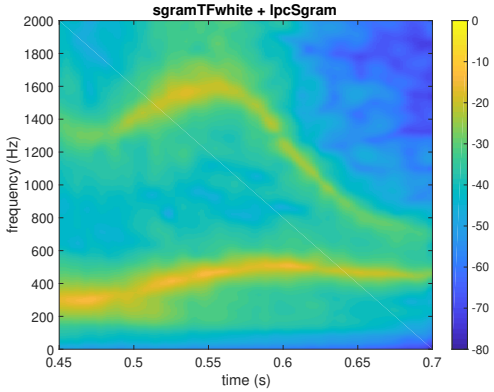

Fig. 8. Magnified view of the recovered spectrogram. The image shows the initial low frequency part.

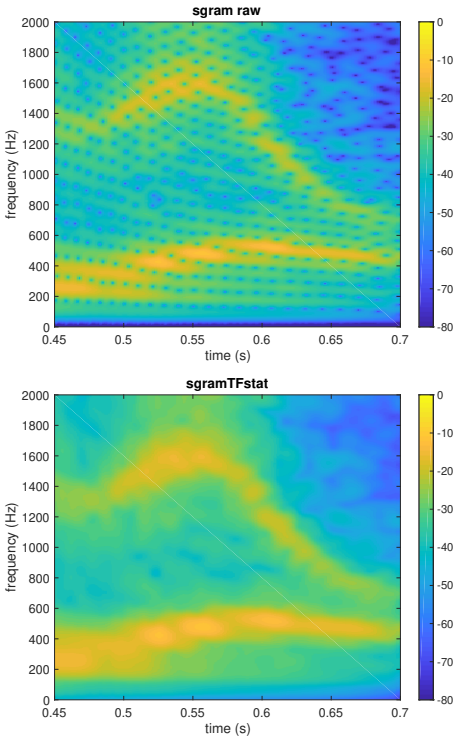

Fig. 9. Illustrative spectrogram with magnified view.

magnified view. The top image shows the raw spectrogram calculated using only the main window. The regular dots are the result of interferences between adjacent harmonic components. The second image shows the directly derived interference-free representation. Note that the formant trajectories look broader by smearing caused by the mismatch of interpolation behavior.

Figure 10 shows spectral slices of these representations. The raw power spectrum is shown using the blue line. The finally recovered spectral envelope is shown using the orange

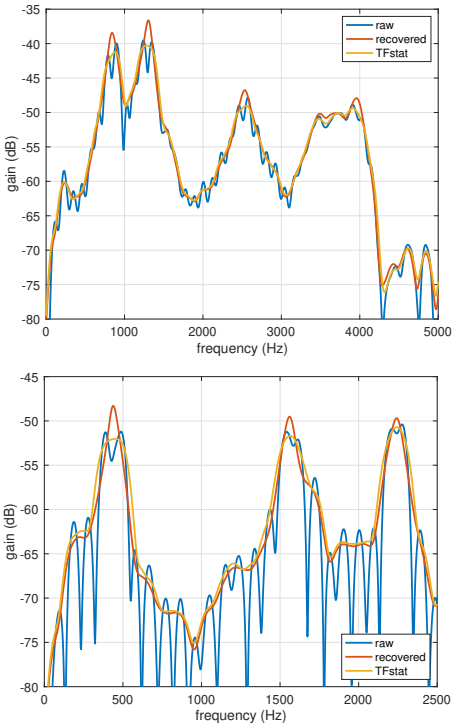

Fig. 10. Spectral slice of several representations. Slice positions from top-left, top-right to bottom-right: 180 ms, and 540 ms.

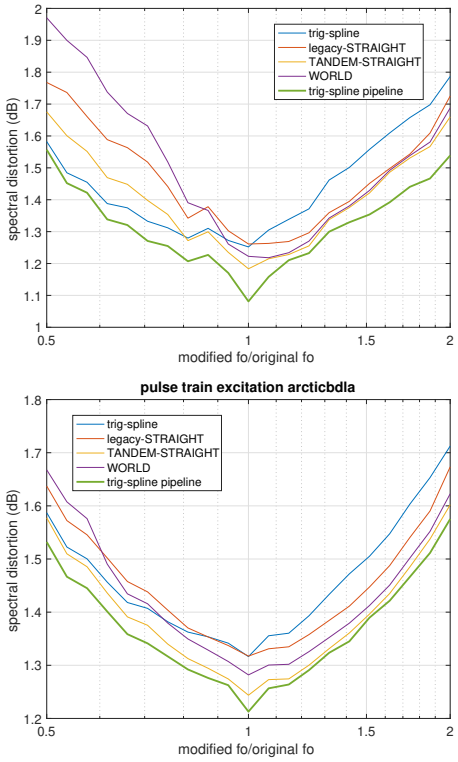

Fig. 11. Simulated fo conversion and spectral recovery accuracy.

line. The directly applied interference-free representation is shown using the yellow line. These examples are worst-case examples. In these spectral slices, the first and the second formant frequencies are between adjacent harmonic components. It made the direct interference-free representation have flattened peak shape. The pre-whitening procedure and the recovery post-processing recover the proper spectral peak shape.

A. Simulated evaluation by f_o modification

We conducted recovery accuracy evaluation by using the recovered interference-free representation as the ground truth surface. We generated synthetic speech signals using converted f_o trajectories by multiplying the modification coefficient c_m. We set the instantaneous amplitude of each modified harmonic component by reading from the surface and added together to generate the test signals. We tested the proposed method using the trigo-spline function with and without pipeline procedure and compared recovered spectral envelope using legacy-STRAIGHT, TANDEM-STRAIGHT, and WORLD. Figure 11 shows the results. The horizontal axis represents the amount of f_o modification c_m. The vertical axis represents the RMS average of the dB distance between the ground truth and the calculated spectral envelope. The upper plot shows the results using the Japanese vowel sequence used in the previous section. The lower plot shows the results using a file (`arctic_a0023.wav` of speaker bdl) in the CMU arctic database[35]. Note that the latter sample was downsampled to 16 kHz before analysis. These results illustrates that the interference-free representation calculated using pipeline processing outperforms in the spectral envelope recovery accuracy.

B. Evaluation by speech production simulator

We tested spectral envelope recovery accuracy using a speech production simulator in SparkNG[26]. The simulator consists of a one-dimensional vocal tract and the anti-aliased L-F model[36] for the glottal excitation source. This setting enables to evaluate spectral distortion objectively because the simulator provides the ground truth. We prepared twelve Swedish vowels as the reference points using the first three formant frequencies and band widths[37]. Then introduced perturbation to the formant frequencies and bandwidths and added higher formant information by assuming the sampling frequency 22,050 Hz with 17.0 cm vocal tract length.

Figure 12 shows the spectral distortion as the function of the fundamental frequency. Evaluation of the spectral distortion used the frequency range from the fundamental frequency to 4,000 Hz The results indicate that the proposed pipeline procedure provides the best performance. We tested using three voice quality types (modal, breathy, creaky)[38]. For all conditions, the proposed pipeline procedure provided the best performance.

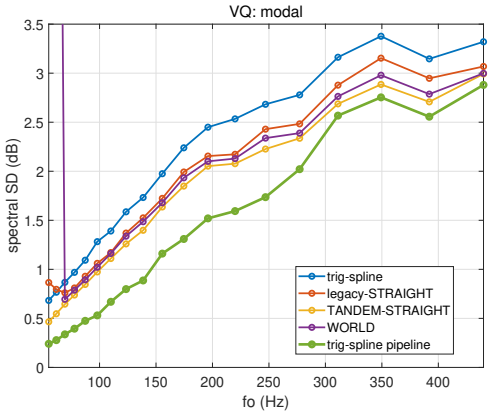

Fig. 12. Spectral distortion of legacy-STRAIGHT, TANDEM-STRAIGHT, WORLD and the proposed method.

VII. Discussion and Future Work

We introduced two types of procedures for calculating interference-free representations of periodic sounds. They are generally applicable to periodic sound analysis. The one-stage procedure yields the best performance when the actual value of the fundamental frequency is known. When the assumed fundamental frequency consists of errors, the two-stage procedure is more tolerant to the errors. This tolerance to errors suggests that the one-stage procedure is suitable for measurements where the fundamental frequency of the test signal is known. We recently introduced a signal called frequency domain velvet noise (FVN)[39], which is a variant of the velvet noise[40], [41]. Application of FVN enables a novel set of procedures for acoustic impulse response measurement[42] which includes musically entertaining test signals for impulse response. Combination of FVN with this interference-free representation also introduces an attractive real-time acoustic measurement application.

It also suggests that the two-stage procedure is preferable for speech analysis where the estimate of the fundamental frequency usually consists of errors. In addition to the two-stage procedure, we proposed a pipeline procedure specialized for voiced speech analysis. It uses an LPC-based analysis for pre-whitening equalization of the input signal. By mixing the LPC-based envelope and the interference-free representation of the whitened signal yields the refined spectral envelope. One technical issue of this pipeline procedure is the pre-processing for LPC-based analysis. The current implementation uses a simple differentiation for this pre-processing. This simple procedure has to be revised by a relevant adaptive procedure for processing a variety of natural speech materials. This pipeline-based method can revise the infrastructure of our VOCODERs and may provide better sounding resynthesized sounds. The proposed representations may also useful for improving speech manipulation procedures. However, these are the topics for further research.

VIII. Conclusion

We proposed a set of new accurate spectral envelope recovery methods for speech sounds generated by periodic excitation based on a set of interference-free power spectrum representations. The proposed methods outperform our previous spectral recovery models used in legacy-STRAIGHT, TANDEM-STRAIGHT, and WORLD VOCODERs. Our first goal was to calculate interference-free representations of periodic sounds. We introduce two methods to attain this goal. The first one is the one-stage algorithm by using paired time windows. The other one is the two-stage algorithm using B-spline functions and a composite of a trigonometric function and B-spline functions. Our second goal is to apply these interference-free representations to voiced speech analysis. We introduce a post and pre-processing to improve recovery accuracy around spectral peak regions to attain this goal. We conducted a set of evaluation tests using voice production simulator and natural speech samples. Finally, we discuss the application of the proposed method for revising high-quality VOCODERs. We will make these procedures open-source because they are useful for analyzing and processing general periodic signals.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 16H01734, and 15H03207.

References

[1] 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.

[2] 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.

[3] 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, pp. 1877–1884, 2016.

[4] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[5] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, may 2008.

[6] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencie," *Electro. Comm. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.

[7] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[8] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America*, vol. 45, no. 2, 1969.

[9] M. Schroeder and B. S. 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.

[10] R. Salami, C. Laflamme, J. P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 116–130, 1998.

[11] M. Unser, "Sampling-50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, apr 2000.

[12] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA*, vol. 36, no. 5, pp. 713–727, 2011.

[13] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[14] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech and Signal Processing,*, vol. 34, no. 4, pp. 744–754, aug 1986.
[15] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. Eurospeech 95*, Madrid, 1995, pp. 451–454.
[16] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.
[17] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085–2095, Oct 2013.
[18] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, mar 2007.
[19] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, Apr 1983, pp. 93–96.
[20] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, Feb 1991.
[21] A. v. d. 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*, 2016.
[22] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6309–6318.
[23] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel WaveNet: fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.
[24] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech 2017*, 2017, pp. 1118–1122.
[25] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with WaveNet autoencoders," *arXiv preprint arXiv:1704.01279*, 2017.
[26] H. Kawahara, "SparkNG: MATLAB realtime speech tools and voice production tools," (Accessed:2018-08-11). [Online]. Available: https://github.com/HidekiKawahara/SparkNG
[27] N. Thapen, "PINK TROMBONE: bare-handed procedural speech synthesis," (Accessed:2017-09-20). [Online]. Available: https://dood.al/pinktrombone/
[28] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *ICASSP 2003*, vol. I, 2003, pp. 256–259.
[29] H. Kawahara, M. Morise, Banno, and V. G. Skuk, "Temporally variable multi-aspect N-way morphing based on interference-free speech representations," in *ASPIPA ASC 2013*, 2013, p. 0S28.02.
[30] S. Popham, D. Boebinger, D. P. W. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature Communications*, vol. 9, no. 1, p. 2122, dec 2018.
[31] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D. G. Miller, H. Noé, R. C. Scherer, J. R. Smith, B. H. Story, J. G. Švec, S. Ternström, and J. Wolfe, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015.
[32] M. Unser, "Splines: a perfect fit for signal and image processing," *IEEE Signal Processing Magazine*, vol. 16, no. 6, pp. 22–38, Nov 1999.
[33] M. Morise and Y. Watanabe, "Sound quality comparison among high-quality vocoders by using re-synthesized speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 263–265, 2018.
[34] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT." in *Interspeech 2005*, 2005, pp. 537–540.
[35] J. Kominek and A. Black, "The CMU Arctic databases for speech synthesis," *Proc. ISCA Workshop on Speech Synthesis*, pp. 223–224, 2004.
[36] 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 *Interspeech 2017*, vol. 2017-Augus. ISCA: ISCA, aug 2017, pp. 1358–1362.
[37] J. W. Hawks and J. D. Miller, "A formant bandwidth estimation procedure for vowel synthesis," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1343–1344, 1995.
[38] D. G. Childers and C. Ahn, "Modeling the glottal volume- velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.
[39] H. Kawahara, K.-I. Sakakibara, M. Morise, H. Banno, T. Toda, and T. Irino, "Frequency domain variants of velvet noise and their application to speech processing and synthesis," in *Proc. Interspeech 2018*, 2018, [in print].
[40] H. Järveläinen and M. Karjalainen, "Reverberation modeling using velvet noise," in *AES 30th International Conference, Saariselkä, Finland*. Audio Engineering Society,, 2007, pp. 15–17.
[41] 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.
[42] H. Kawahara and K.-I. Sakakibara, "Acoustic measurements using a frequency domain velvet noise and interference-free power spectral representations of periodic sounds," *TECHNICAL REPORT OF IEICE*, vol. IEICE, no. EA, Aug. 2018, [in Japanese].

APPENDIX

Steps to calculate the interference-free representation of natural speech examples.

Source analysis:
: The fundamental frequency extraction is the first step of the analysis.

Spectrogram: first scan:
: The time-frequency interference-free spectrogram is extracted. This test script also calculates the usual spectrogram, and temporally static spectrogram together. Note that the speech signal is pre-emphasized to compensate the global spectral tilt. This time, differentiation is applied (as the first order approximation).

Conversion to the interference-free auto-correlation:
: The inverse Fourier transform of the interference-free spectrogram (power spectrum dimension) yields autocorrelation. The 0-lag component corresponds to the power.

LPC analysis: seek for the appropriate order
: The linear prediction errors from order 1 to 30 provide the clue to decide the relevant analysis order. LPC-spectrograms with and without power matching are calculated.

LPC analysis:
: LPC analysis provides the reflection coefficients. The analysis order was 26, based on the visual inspection of the prediction errors.

Sample-wise interpolation of the reflection coefficients:
: The update rate of the reflection coefficients was 1 ms. This process interpolates the reflection coefficients at the audio sampling rate.

Whitening of the signal
: The LPC-lattice inverse filter whitens the speech signal and yields the whitened speech signal.

Interference-free spectrogram of the whitened speech signal:
: This is the second scan of calculating speech spectrogram without time-frequency interferences.

Recovery of details:
: The LPC-spectrogram and the interference-free spectrogram of the whitened speech are combined to yield the recovered spectrogram.