Excitation source design for high-quality speech manipulation systems based on a temporally static group delay representation of periodic signals

Hideki Kawahara*, Masanori Morise†, Tomoki Toda‡, Hideki Banno§, Ryuichi Nisimura* and Toshio Irino*
*Faculty of Systems Engineering, Wakayama University, Wakayama, Wakayama, Japan
E-mail: {kawahara, nisimura, irino}@sys.wakayama-u.ac.jp Tel: +73-457-8461
†Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, Yamanashi, Japan
E-mail: mmorise@yamanashi.ac.jp
‡Nara Advanced Institute of Science and Technology, Ikoma, Nara, Japan
E-mail: tomoki@is.naist.jp
§Graduate School of Science and Technology, Meijo University, Nagoya, Japan
E-mail: banno@meijo-u.ac.jp

***Abstract*—A new group delay representation, which yields value zero for periodic signals irrespective to the initial phase and the relative level of each harmonic component. This new group delay representation provides a unified basis for defining "aperiodicity" in speech sounds. For example, the periodic to noise ratio or harmonic to noise ratio is directly derived from the deviation of this group delay representation from value zero, after removing FM effects of harmonic frequencies and removing AM effects of harmonic component level. The derived deviation is combined with estimated excitation duration information and used to design aperiodic components of excitation source for high-quality synthetic speech. The proposed group delay representation is based on F0-adaptive weighted average of frequency shifted versions and temporally shifted versions of group delays with power spectral weighting.**

I. Introduction

Combination of the new group delay representation [1] and group delay-based compensation [2] provides a unified basis for analyzing aperiodic aspects of speech sounds. Deviation from pure periodicity in voiced sounds plays important roles in speech communication. Temporal variation of F0 (fundamental frequency) is the primary carrier of prosodic information. Expressive voices in singing or theatrical performances use aperiodic aspects very effectively [3]. Speakers' emotional states also affect voice apperiodicity and are directly (sometimes unconciously) perceived by listeners. However, despite of the importance, it has been very difficult to analysis, represent and design the voice aperiodicity in a unified and mathematically well defined framework.

The new group delay representation enables to introduce a simple and powerful strategy, the null method, because the representation yields value zero for periodic signals irrespective to the initial phase and the level of each harmonic component. The magnitude of deviation from zero of this group delay representation, after removing known biasing factors such as AM and FM by fine tuning parameters of these modulations to minimize the deviation, provides the magnitude of aperiodicity which are not represented by these modulations. This magnitude of deviation is directly corresponds to the power ratio of the periodic component to the random component, in other words, the harmonic to noise ratio.

Since this measure is not affected by the initial phase and the level of each harmonic component, a complementary measure which represents temporal distribution of aperiodic component is necessary for representing and designing the excitation source signals. Duration of the windowed signal with minimum phase group delay compensation [2], [4] provides this information. Note that temporal distribution of aperiodic component has significant perceptual effects, especially for male voices, in terms of temporal masking level (sometimes the effect exceeds 20 dB) [5].

The primary motivation of this investigation is to revise the representation of the aperiodic component of TANDEM-STRAIGHT [6], a speech analysis, modification and resynthesis framework, based on a solid conceptual as well as methodological ground. The framework is based on temporally static representations of periodic signals, such as power spectrum and instantaneous frequency [7]. Introduction of this new group delay representation makes all modules of TANDEM-STRAIGHT temporally static.

This article mainly focuses on the new temporally static group delay, since the idea and the formulation are novel and fundamental. Temporal distribution of the aperiodic power and its application are briefly discussed and their details are left for future investigations.

II. Target system of the proposed representations

TANDEM-STRAIGHT is a speech analysis, modification and synthesis framework primarily designed for providing flexible tools for speech perception research [8]. Input speech

978-616-361-823-8

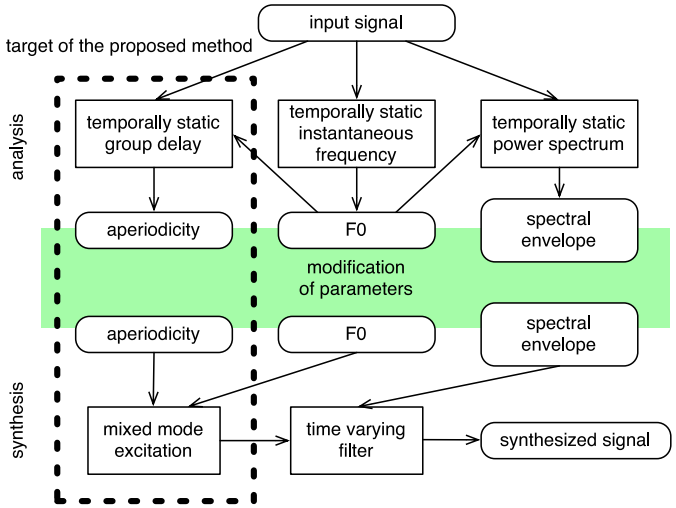

Fig. 1. Schematic diagram of TANDEM-STRAIGHT structure. The portion sorrounded by dashed square indicates the target of this manuscript

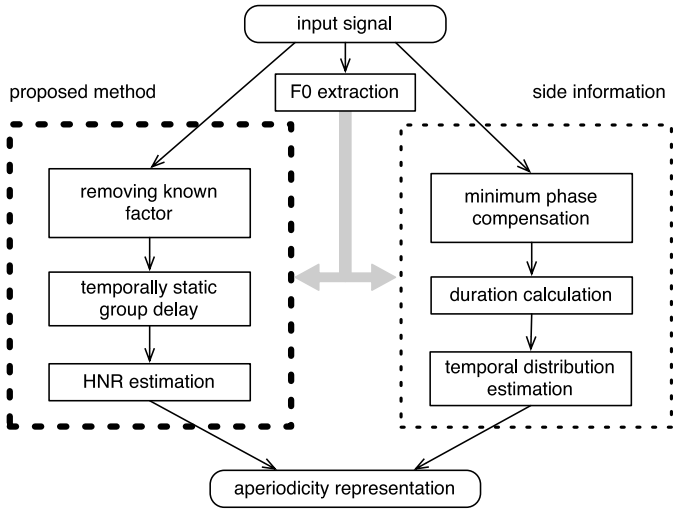

Fig. 2. Overview of the aperiodicity extraction and the proposed method

signals are analyzed to yield the source and spectral representations. The source representations consist of F0 and aperiodic information, which is the target topic of this manuscript. Figure 1 shows the schematic diagram of TANDEM-STRAIGHT and the target.

Continuing expansion of TANDEM-STRAIGHT-based applications, such as morphing [9], [10], [11], [12], [13], made requirement on speech quality of the manipulated sounds more demanding and clarified weakness of the current representations used. The most crucial issue is excitation source representations, especially non-periodic components [4], [3], [14].

Figure 2 shows overview of the revised aperiodicity extraction system for TANDEM-STRAIGHT. HNR value is calculated by the procedures in the left box using the proposed group delay representation. Details of the procedure in the box are illustrated in Figure 8.

III. Background and Related Works

A number of high-quality speech analysis, modification and synthesis frameworks have been introduced [15], [16], [17], [6], [18]. Discarding phase information makes such systems more flexible usually with a cost of quality degradation. Flexibility centered design of STRAIGHT[1] makes it more vulnerable to this issue than the other systems.

Modular structure of STRAIGHT allows using different types of excitation representations to generate output signals. Harmonic plus noise with phase control extension [19] and a cross synthesis VOCODER application [20] are such examples. Other source representations [15], [17], [21], [22], [23], [24], [18] based on other systems can also be used as the input to synthesis subsystem of STRAIGHT, since it is implemented as an approximate time varying filter in those examples [19], [20]. Such STRAIGHT-based hybrid systems may make synthesized sounds sound better possibly with a cost of reduced flexibility. However, instead of seeking such possibilities, this article tries to explore flexibility enhancement by introducing unified model of excitation source based on interference-free representations and reliability bounds posed by TB (time bandwidth) product [25].

For highly flexible manipulations, for example morphing, simple parameterized signal models are desirable. At first glance, quality and flexibility are in trade-off. However, taking into account of perception of temporal fine structures [26], [5], [27], a simple pulse plus time-frequency shaped noise model may provide a counter example, based on the proposed new group delay representation and temporal shaping of aperiodic energy. The proposed representation is applicable to both pulse or epoch [22] based models and sinusoid based models.

IV. Static Representations of Periodic Signals

This section briefly summarizes three interference-free representations. Interference-free representation of power spectra of periodic signals [28] enabled separation of filter information and source information of speech sounds and provided the foundation of STRAIGHT. Interference-free representation of instantaneous frequency of periodic signals [7] provided F0 refinement procedure with fine temporal resolution and high-fidelity trajectory tracking [29]. Interference-free representation of group delay of repetitive signals [30], was introduced but was not been effectively used.

This article extends this group delay representation to be dually inteference-free, in other words, it does not have periodic variations both in the time and the frequency domain. Moreover, this extended representation yields constant zero for all frequency range, when the signal is periodic. Since all these representations share the same strategy, power spectral representation is discussed first.

[1]STRAIGHT represents both STRAIGHT [16] and TANDEM-STRAIGHT [6] afterwards. When distinction is necessary, they are represented as legacy-STRAIGHT and TANDEM-STRAIGHT respectively.

A. Power spectrum

Let T_0 represent fundamental period of a periodic signal, the following equation provides power spectral representation $P_T(\omega,t)$, which does not have temporally varying component: [28], [6]

$$P_T(\omega,t)=\frac{P\left(\omega,t+\frac{T_0}{4}\right)+P\left(\omega,t-\frac{T_0}{4}\right)}{2}, \quad (1)$$

where $P(\omega,t)$ represent the short term power spectrum using a time window centered at time t. The main idea behind this is that the temporal variation of power spectra caused by the interference of adjacent harmonic components is sinusoid (cosine) of period T_0 and can be cancelled out by the component having the opposite polarity [28].

This temporally static representation of power spectra still has periodic variations on the frequency domain reflecting harmonic structure. A F0-adaptive smoothing and compensating operation based on consistent sampling [31] is introduced to remove this variations while preserving levels at harmonic frequencies unaltered. The following approximate implementation based on cepstral liftering effectivey perform the desired function and yields the time-frequency representation $P_{ST}(\omega)$. This power spectral representation $P_{ST}(\omega)$ is called STRAIGHT-spectrum. (Variable t is not shown here for visual simplicity.)

$$P_{ST}(\omega)=\exp\left(\mathcal{F}^{-1}\left[\left(\tilde{q}_0+2\tilde{q}_1\cos\left(\frac{2\pi\tau}{T_0}\right)\right)g(\tau)C(\tau)\right]\right), \quad (2)$$

where $C(\tau)$ represents the cepstrum of TANDEM-spectrum $P_T(\omega,t)$. One of the following lifters are used for $g(\tau)$.

$$g_1(\tau)=\frac{\sin(\pi f_0\tau)}{\pi f_0\tau}=\mathcal{F}[h_1(\omega)] \quad (3)$$

$$g_2(\tau)=\left(\frac{\sin(\pi f_0\tau)}{\pi f_0\tau}\right)^2=\mathcal{F}[h_2(\omega)], \quad (4)$$

where $g_1(\tau)$ corresponds to the rectangular smoother ($h_1(\omega)$; width is $2\pi f_0$) used in TANDEM-STRAIGHT and $f_2(\tau)$ corresponds to the triangular smoother ($h_2(\omega)$; base width is $4\pi f_0$) used in legacy-STRAIGHT.

B. Instantaneous frequency

The following average of instantaneous frequencies $\omega_i(\omega,t)$ weighted by power spectra provides an instantaneous frequency representation $\omega_{iT}(\omega,t)$, which does not have temporally varying component: [7]

$$\omega_{iT}(\omega,t)=\frac{P^{(+)}\omega_i\left(\omega,t+\frac{T_0}{4}\right)+P^{(-)}\omega_i\left(\omega,t-\frac{T_0}{4}\right)}{P^{(+)}+P^{(-)}} \quad (5)$$

where $P^{(+)}$ represents $P\left(\omega,t+\frac{T_0}{4}\right)$ and $P^{(-)}$ represents $P\left(\omega,t-\frac{T_0}{4}\right)$. Note that the denominator of (5) is the interference-free power spectrum $P_T(\omega,t)$ defined by (1) multiplied by 2. Interference-free behavior is proven [7] by using Flanagan's instantaneous frequency equation [32].

C. Group delay: removing frequency interference

Group delay $\tau_d(\omega,t)$ is complementary to instantaneous frequency (for example [33]). This duality led to the following representation of group delay $\tau_{dF}(\omega,t)$, which does not have interferences in the frequency domain caused by multiple (this time two) events: [30]

$$\tau_{dF}(\omega,t)=\frac{P^{(U)}\tau_d\left(\omega+\frac{\omega_0}{4},t\right)+P^{(D)}\tau_d\left(\omega-\frac{\omega_0}{4},t\right)}{P^{(U)}+P^{(D)}}, \quad (6)$$

where $P^{(U)}$ represents $P\left(\omega+\frac{\omega_0}{4},t\right)$ and $P^{(D)}$ represents $P\left(\omega-\frac{\omega_0}{4},t\right)$. Periodicity interval $\omega_0=2\pi/T_0$ on the frequency axis is determined by the temporal interval T_0 between the events. Lengthy derivation of interference-free behavior of $\tau_{dF}(\omega,t)$ is given in [30]. Since group delay is the main topic of this article, outline of the derivation is given below.

The group delay is defined by the negative frequency derivative of the phase of $X(\omega,t)$, the short term Fourier transform of a signal. It is equivalent to calculate the derivative of the imaginary part of the log-converted short term spectrum $\log(X(\omega,t))$.

$$\begin{aligned}-\tau_g&=\frac{d\,\Im\left[\log(X(\omega,t))\right]}{d\omega}=\Im\left[\frac{1}{X(\omega,t)}\frac{dX(\omega,t)}{d\omega}\right]\\&=\frac{\Re[X(\omega,t)]\Im\left[\frac{dX(\omega,t)}{d\omega}\right]-\Im[X(\omega,t)]\Re\left[\frac{dX(\omega,t)}{d\omega}\right]}{|X(\omega,t)|^2},\end{aligned} \quad (7)$$

where $|X(\omega,t)|^2$ is also the power spectrum $P(\omega,t)$. This equation is the counterpart of the Flanagan's equation, in case of group delay. Substituting $X(\omega,t)$ and $X_d(\omega,t)$ defined below:

$$X(\omega,t)=\int_{-\infty}^{\infty}w(\tau)x(\tau-t)e^{-j\omega\tau}d\tau \quad (8)$$

$$X_d(\omega,t)=\frac{dX(\omega,t)}{d\omega}=-j\int_{-\infty}^{\infty}\tau w(\tau)x(\tau-t)e^{-j\omega\tau}d\tau, \quad (9)$$

into (7) yields efficient calculation of group delay by: It leads to the following computationally efficient equation:

$$-\tau_g(\omega,t)=\frac{\Re[X(\omega,t)]\Im[X_d(\omega,t)]-\Im[X(\omega,t)]\Re[X_d(\omega,t)]}{|X(\omega,t)|^2}, \quad (10)$$

where $X(\omega,t)$ and $X_d(\omega,t)$ are defined below:

$$X(\omega,t)=\int_{-\infty}^{\infty}w(\tau)x(\tau-t)e^{-j\omega\tau}d\tau \quad (11)$$

$$X_d(\omega,t)=\frac{dX(\omega,t)}{d\omega}=-j\int_{-\infty}^{\infty}\tau w(\tau)x(\tau-t)e^{-j\omega\tau}d\tau. \quad (12)$$

Note that the weights $P^{(U)}$ and $P^{(D)}$ in (6) cancel out with the denominator of (10) and that the denominator of (6) does not have periodic variation on the frequency axis. These make inspection on the denominator unnecessary. Substituting (10) to (6) and using the identity ($\sin^2\theta+\cos^2\theta=1$) shows that the periodic variation of group delay on the frequency axis caused by multiple excitation effectively vanishes [30]. However, unlike power spectrum and instantaneous frequency,

the proposed interference-free representation of group delay $\tau_{dF}(\omega,t)$ was not very successful in speech applications [30]. This inefficacy is caused by the huge dynamic range of speech spectra, because interference suppression requires that the denominator $P^{(U)}+P^{(D)}$ is changing smoothly and gradually in terms of ω. This is not the case for vowels.

D. Group delay: removing time-frequency interference

The interference-free representation of group delay $\tau_{dF}(\omega,t)$ defined by (6) still has periodic interference in the time domain when periodic signals are analyzed. Similar to the interference-free power spectra and instantaneous frequencies, calculating weighted average of $\tau_{dF}(\omega,t)$ calculated at two points $T_0/2$ apart may suppress the temporal interferences in $\tau_{dF}(\omega,t)$. A group delay representation $\tau_{dD}(\omega,t)$ that is interference-free in the both time and frequency domains is defined below:

$$\tau_{dD}(\omega,t)=\frac{P^{B+}\tau_{dF}\left(\omega,t+\frac{T_0}{4}\right)+P^{B-}\tau_{dF}\left(\omega,t-\frac{T_0}{4}\right)}{P^{B+}+P^{B-}}, \quad (13)$$

where P^{B+} represents $P\left(\omega+\frac{\omega_0}{4},t+\frac{T_0}{4}\right)+P\left(\omega-\frac{\omega_0}{4},t+\frac{T_0}{4}\right)$ and P^{B-} represents $P\left(\omega+\frac{\omega_0}{4},t-\frac{T_0}{4}\right)+P\left(\omega-\frac{\omega_0}{4},t-\frac{T_0}{4}\right)$. When the signal is periodic, $\tau_{dD}(\omega,t)=0$ effectively holds. This equation is conceptually simple and computationally efficient.

E. Determination of windowing function and parameters

Unfortunately, this dually interference-free representation $\tau_{dD}(\omega,t)$ does not suppress both interferences perfectly. Numerical optimization was conducted for determining the time windowing function and related parameters. The cost function L for this tuning is defined below:

$$L^2=\frac{1}{S(\Omega,T)}\int_\Omega\int_T|\tau_{dD}(\omega,t)|^2dtd\omega, \quad (14)$$

where $S(\Omega,T)$ represents the measure defined by the set of temporal observation T and the frequency region Ω. Note that the cost L represents spread of the calculated group delay in time (duration).

The periodic component $x_p(t)$ of the test signals were generated by using following equation.

$$x_p(t)=\sum_{k=0}^{\left\lfloor\frac{f_s}{2f_0}\right\rfloor}a_k\cos\left(2\pi kf_0t+\varphi_k\right), \quad (15)$$

where f_s represents the sampling frequency, f_0 represents the fundamental frequency, a_k represents the amplitude of the k-th harmonic component, and φ_k represents the initial phase of the k-th harmonic component. A test signal $x(t)$ is prepared by mixing a periodic component and a Gaussian white noise $x_n(t)$ by assigning mixing weight for each component.

$$x(t)=c_px_p(t)+c_nx_n(t), \quad (16)$$

where c_p and c_n represent mixing weights for the periodic component and the random component respectively. In this simulation $T_0=0.01$ s ($f_0=100$ Hz) is used. For the frequency range, $\Omega=[0,\ f_s/4]$ was used in this simulation.

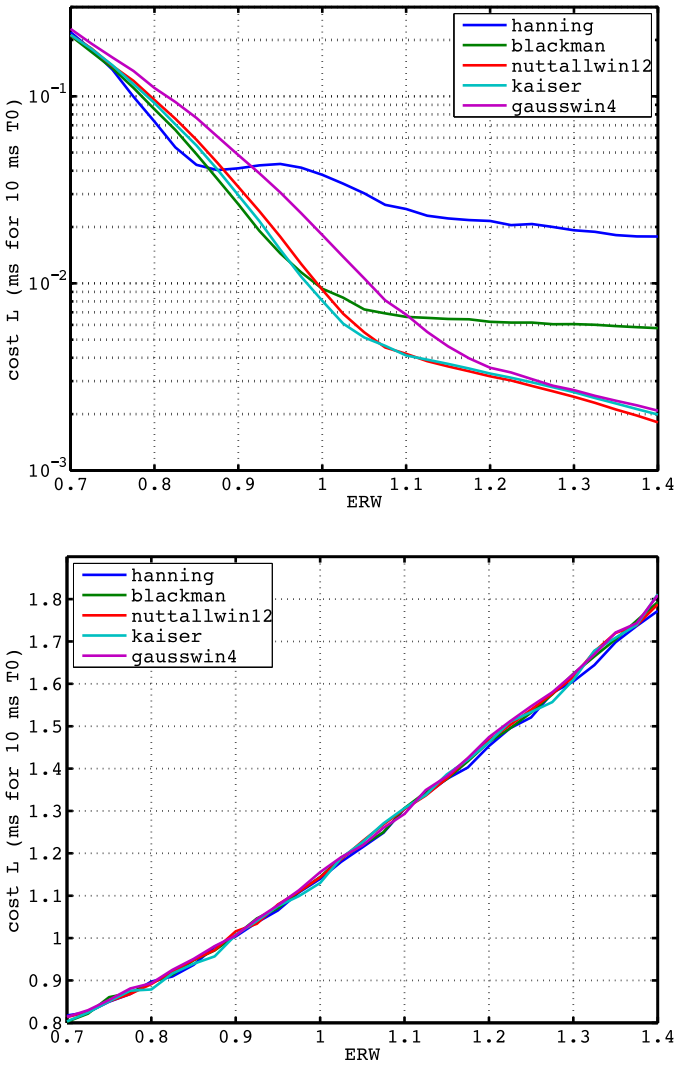

Fig. 3. Window size and cost L for different windows. Upper plot represents the results for 100 Hz periodic signal with random initial component phases which uniformly distribute in $[0,2\pi)$. The window size is represented in terms of the effective rectangular window duration. The lower plot shows results for Gaussian random input.

Fig. 3 shows the cost function values for Hann [34], Blackman [34], Nuttall [35][2], Kaiser [34], [36] ($\alpha=10$) and Gaussian (width is 4σ) windows in terms of the effective rectangular window length ERW defined below.

$$ERW=\left(\frac{\int_{-T_W/2}^{T_W/2}t^2w^2(t)dt}{\int_{-T_0/2}^{T_0/2}t^2dt\int_{-T_W/2}^{T_W/2}w^2(t)dt}\right)^{\frac{1}{2}}, \quad (17)$$

where $T_0=1/f_0$ represents the fundamental period and T_W represents the nominal window length of the windowing function $w(t)$.

[2]The 12th item in Table II of this reference is used here. It is different from the Matlab function `nuttallwin`.

The upper plot of Fig. 3 shows the results for $c_n = 0$ and the lower plot shows the results for $c_p = 0$. The initial phase φ_k of each harmonic component is sampled from the uniform distribution in $[0, 2\pi)$. For the observation set T, 50 observations (10 locations in one cycle for 5 different initial phase settings) were used for upper plot and 200 independent observations were used for lower plot. Note that at $ERW = 1.1$, the cost function value for periodic signals is about 300 times smaller than that for random signals when Nuttall or Kaiser windowing function is used. At $ERW = 1$, Kaiser window provides the best cost for periodic signals, which is about 150 times smaller than that for random signals. These cost differences between periodic signals and random signals are large enough to evaluate deviation from pure periodicity accurately and can be applicable to design aperiodic components in excitation signals. This is a significant improvement from our previous report [1] on a temporally static group delay representation, where only Hann window was evaluated. (The cost for periodic signal is only 25 times smaller than that for random signal when Hann window is used.)

It is important to note that to attain the same performance at $ERW = 1.1$, Kaiser window needs 10% shorter window length than that of Nuttall window. It reflects the fact that Kaiser window [34], [36] is an approximation of prolate spheroidal wave function, which provides the best time-frequency uncertainty when support of the function is bounded [37]. Based on these factors, we decided to rely on Kaiser window in the following sections.

F. Behavior of the static group delay

An example snapshot of a visualization movies is shown in Fig. 4. The movie which is the source of this snap shot is designed to illustrate behavior of the proposed group delay.

In the following subsections, this type of snapshots are extensively used to introduce behaviors of the proposed method for different types of input signals. The snapshot consists of the following panels to display intermediate representations and the proposed static group delay representation.

Waveform and windowing functions:
The top left panel shows the input signal and time windows. The thick green line represents the windowing function which is used to calculate the phase spectrogram below. The other two windows represented using thin green and red lines represent windows actually used to calculate the static group delay.

Phase spectrogram:
The bottom left panel shows the phase spectrogram. Phase values are represented using pseudo color scheme. In this example, the color continuously changes in the following order; red, yellow, green, cyan, blue, violet and red according to the phase value. The first red corresponds to the phase value 0 and the last red corresponds to the phase value 2π. The horizontal time axis is aligned with the waveform panel so that the phase calculated by using the time window displayed on the left top panel is pasted on the center of this phase spectrogram. The vertical frequency axis is aligned with the power spectra and the group delay panels placed on the right side.

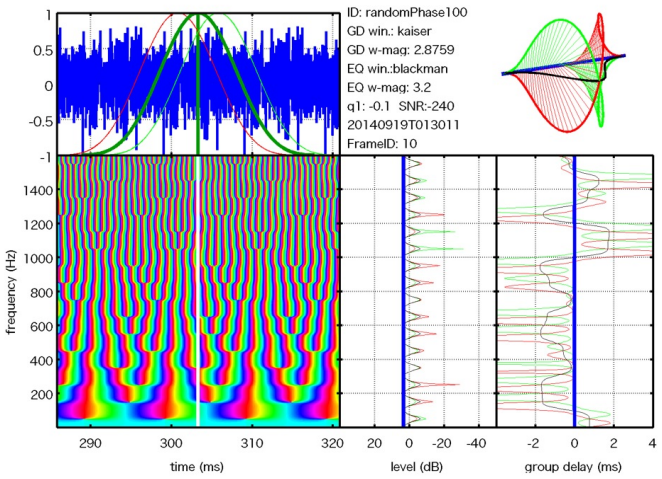

Fig. 4. Integrated display of the static group delay with additional intermediate information. The test signal is a periodic signal consisting of harmonically related sinusoids with random initial phase ($f_0 = 100$ Hz) and the same amplitude.

Power spectra:
The bottom center panel shows two power spectra (thin green and red lines) and the TANDEM spectrum (thin black line) and the STRAIGHT-spectrum (Thick blue line) calculated using the two time windows in the waveform panel.

Group delay representations:
The bottom right panel shows two group delays (thin green and red lines) which are calculated using the center window shown in the waveform panel for illustration purpose. It also displays the averaged group delay (thin black line) using frequency shifted versions of power spectra. The static group delay is represented using a thick blue line. Note that it visually matches to the vertical line located on the center.

Analysis conditions:
The top center panel lists parameter settings used to calculate displayed results.

Windowing function for frequency shifted group delay:
The top right panel displays shape of windowing functions used to calculated the frequency shifted group delay shown in the green and red thin lines in the group delay panel.

The source movie of Fig. 4 illustrates that the proposed group delay (thick blue line in the bottom right panel) does not move and stays at 0 ms. This represents that the input signal is locally highly periodic.

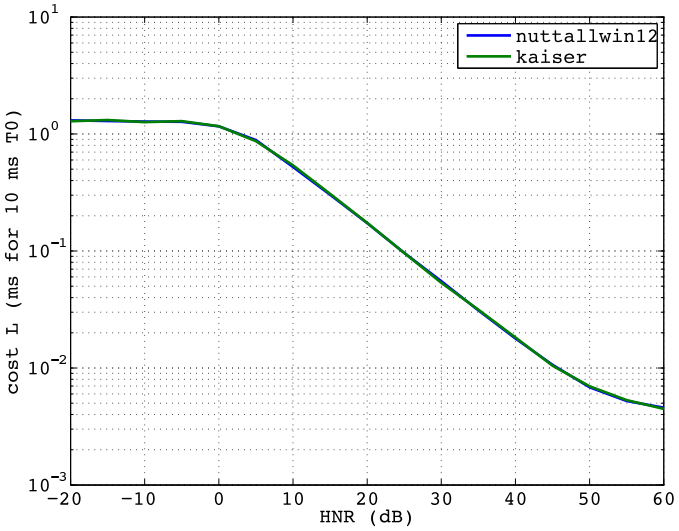

Fig. 5. Costs for HNR conditions using the Nuttall window with the nominal length $3.1629T_0$ (1.1 in ERW) and Kaiser window with the nominal length $2.8759T_0$ (1.1 in ERW).

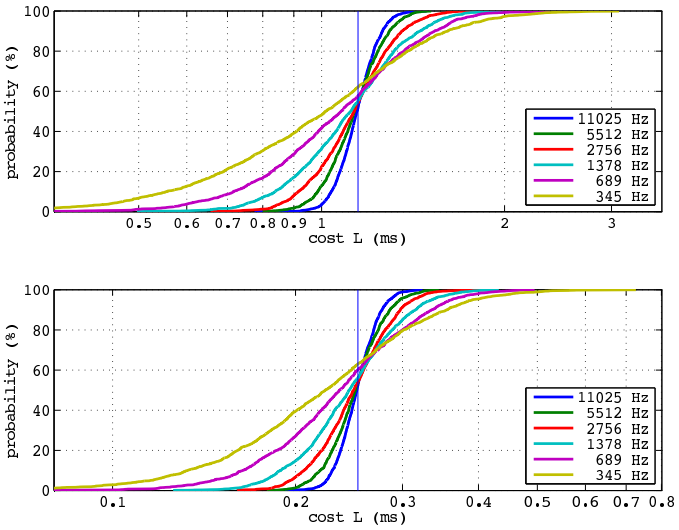

Fig. 6. Costs cumulative distribution as a function of bandwidth. Input signals are Gaussian noise. Kaiser window with $1.1ERW$ is used ($T_0 = 0.01$ s).

1) Insensitivity to the initial phase: Figure 3 shows that the proposed group delay representation is effectively independent on the initial phase of each harmonic component when the level of each harmonic component is constant. Fig. 4 shows a snapshot for the input periodic signal with random initial phase. The signal was generated by setting the initial phase of harmonic components $\{\varphi_k\}_{k\in Z}, (Z = \{0, \ldots, \left\lfloor \frac{f_s}{2f_0} \right\rfloor\})$ in (15) using samples from the uniform distribution in $[0, 2\pi)$. The movie shows that the proposed group delay (thick blue vertical line in the bottom right panel) does not move and stays at 0 ms while signal looks random due to phase randomization and the thin black line in the group delay display moves periodically. This illustrates insensitivity of the proposed group delay to the initial phase of harmonic components. These results suggest that deviation from 0 in the proposed group delay can be used as an objective measure of aperiodic components. This idea is explored in the following section for designing excitation source aperiodicity.

V. Excitation source design

In this section, a design procedure of the aperiodic component is introduced based on simulation of each constituent functions. The most important function is HNR (harmonic to noise ratio) design based on the observed cost.

Fig. 5 shows the relation between HNR and the cost function for a Nutall window and a Kaiser window with the same effective window length ($ERW = 1.1$). They are closely overlapped and virtually parallel to -20 dB/oct log-linear decay. This indicates that HNR can be directly obtained from the cost L using a simple linear conversion for a reasonably wide HNR range. The nominal window length of Kaiser is about 9% shorter than that of Nuttall window. It implies that Kaiser window is preferable because it provides equivalent performance using fewer samples of data. Note that these results are averaged value based on many observations. Application to excitation design requires reliability in a temporally single observation.

Fig. 6 shows cumulative distribution of the cost L and the modified cost function L_d as functions of frequency bandwidth (width of S) in case of single observation in time. The modified cost function L_d is defined below.

$$L_d^2 = \frac{1}{S(\Omega, T)} \int_\Omega \int_T \left| \frac{d\tau_{dD}(\omega, t)}{d\omega} \right|^2 dt d\omega, \qquad (18)$$

where the frequency range Ω was selected from one of octave bands prepared by halving whole frequency range recursively. ($[f_s/4,\ f_s/2],\ [f_s/8,\ f_s/4], \ldots, [f_s/128,\ f_s/64]$) Note that for the widest band, about 90% of observations yield the cost value L within $\pm 10\%$ around the averaged value, which is represented using a thin blue vertical line in the plot. Distributions of L and L_d are close to each other. Only major difference is the average value.

Fig. 7 shows the standard deviation and average of the cost L and the modified cost L_d. These figures show the test results of 1579 independent single observations. Note that the average value of costs L and L_d are independent from the bandwidth and equal to those in Fig. 5.

A. Processing strucuture

Fig. 8 illustrates the schematic diagram of the proposed method for designing aperiodic component of the excitation source. The procedure consists of the preprocessing, static group delay calculation, and post processing.

The band-wise processing in Fig. 8 calculates effective durations of aperiodic components using $L_{OCT}(\omega, t)$ and $L_{dOCT}(\omega, t)$, which are defined by the following equations based on the static group delay and its frequency derivative,

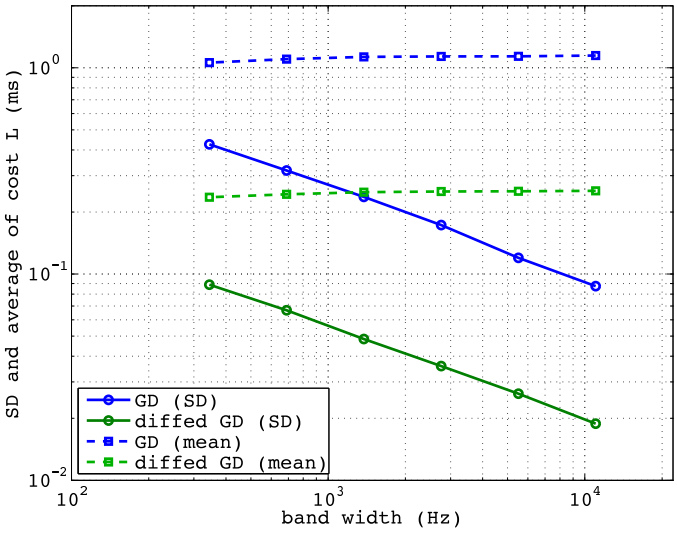

Fig. 7. Standard deviation and average of cost L as a function of bandwidth. Input signals are Gaussian noise. Kaiser window with $1.1ERW$ is used ($T_0 = 0.01$ s).

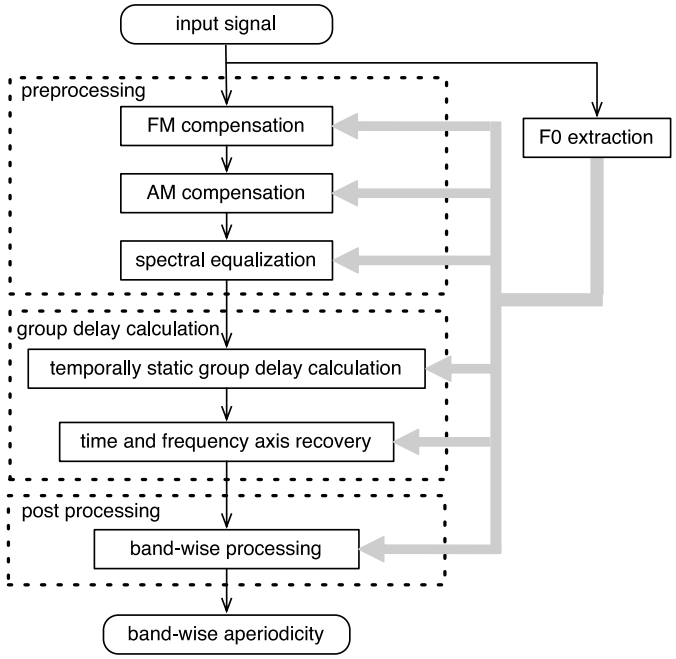

Fig. 8. Schematic diagram of the processing structure.

respectively.

$$L^2_{OCT}(\omega,t) = \frac{\int_{\omega_L}^{\omega_H} P_{ST}(\nu,t)\tau^2_{dD}(\nu,t)d\nu}{\int_{\omega_L}^{\omega_H} P_{ST}(\nu,t)d\nu}, \quad (19)$$

$$L^2_{dOCT}(\omega,t) = \frac{\int_{\omega_L}^{\omega_H} P_{ST}(\nu,t)\left(\frac{d\tau_{dD}(\nu,t)}{d\nu}\right)^2 d\nu}{\int_{\omega_L}^{\omega_H} P_{ST}(\nu,t)d\nu}, \quad (20)$$

$$\omega_L = \frac{\omega}{\sqrt{2}}, \quad \omega_H = \omega\sqrt{2},$$

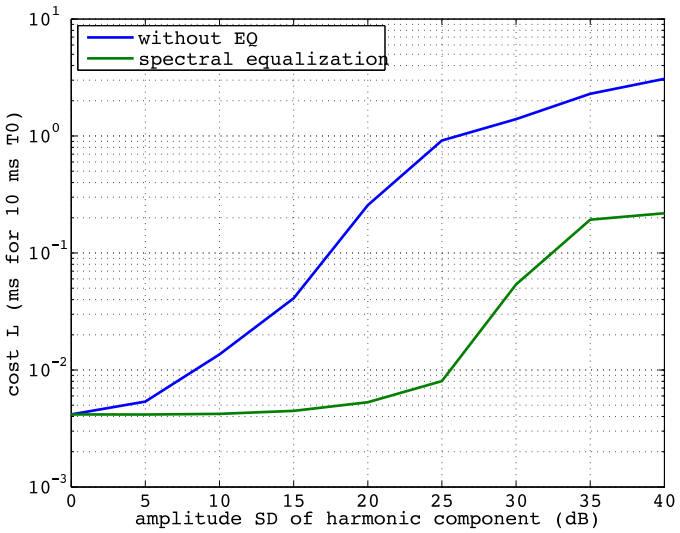

Fig. 9. Cost L to harmonic amplitude variations. The horizontal axis represents standard deviation of harmonic amplitude variations in terms of dB. The upper line represents the results without spectral equalization. The lower line represents the results with spectral equalization based on STRAIGHT spectrum.

B. Preprocessing for parameter extraction

The derivation of the proposed group delay representation assumes that there exist no AM or FM and all harmonic components have the same amplitude. These do not hold for speech. A set of preprocessing procedures were introduced to modify the input signals to reduce these discrepancies. The following subsections provides descriptions of each required preprocessing procedure.

1) Spectral equalizaton of the harmonic amplitudes: Fig. 9 shows the dependency of the cost function L to the amplitude variations of harmonic components of periodic signals defined by (15). The horizontal axis of Fig. 9 represents the amplitude variation in terms of dB. Gaussian distribution was used to randomize the amplitudes of harmonic components. The initial phase distribution is the same to Fig. 4. For each amplitude condition, 600 independent observations were simulated.

Upper line in Fig. 9 represents the results without spectral equalization. It illustrated that the cost L deteriorates by introducing amplitude variation of harmonic components. Lower line represents the results with spectral equalization using the inverse filter designed based on the STRAIGHT-spectrum of the input signal. The lifter coefficient $\tilde{q}_1$ is numerally adjusted to minimize the cost L using the cepstrum liftering in (3).

The results indicates this equalization effectively suppresses this deterioration up to 25 dB amplitude variations of harmonic component. Maximum suppression level of L,1/100 is observed at this point.

Fig. 10 shows a snapshot of the movie with the amplitude and phase randomized input. The thick blue line of the bottom center panel shows the STRAIGHT-spectrum, which is used to design the preprocessing equalizer. The final result, the proposed group delay, also does not move and stays at 0 ms.

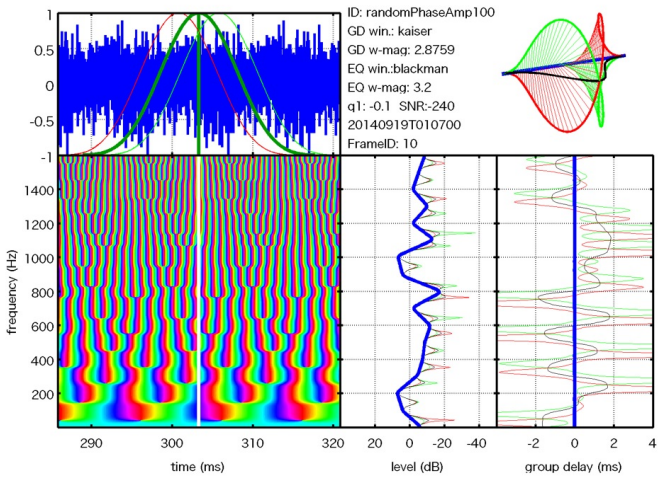

Fig. 10. Integrated display of the static group delay with additional intermediate information. The test signal is a periodic signal consisting of harmonically related sinusoids with random initial phase and random amplitude ($f_0 = 100$ Hz).

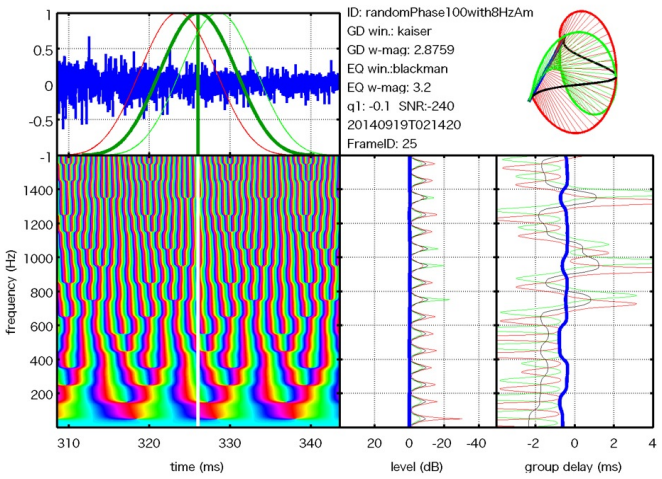

Fig. 11. Integrated display of the static group delay with additional intermediate information. The test signal is a periodic signal consisting of harmonically related sinusoids with random initial phase and applied AM with the following parameters ($f_m = 8$ Hz, $c_{AM} = 0.5$).

This illustrates effective insensitivity of the proposed group delay with relevant preprocessing, STRAIGHT-spectrum-based spectral equalization.

2) Suppression of AM effects: Amplitude variation also make the cost L deteriorate. The following equation is used to generate test signals $x_{AM}(t)$ with amplitude modulation.

$$x_{AM}(t) = a(t) \sum_{k=0}^{\left\lfloor \frac{f_s}{2f_0} \right\rfloor} \cos\left(2\pi k f_0 t + \varphi_k\right), \tag{21}$$

$$a(t) = \left(1 + c_{AM} \sin\left(2\pi f_m t\right)\right), \tag{22}$$

where c_{AM} represents the amplitude modulation depth and f_m represents the frequency of the amplitude modulation.

Fig. 11 shows a snapshot of a visualization movie of AM signal input. It is a periodic signal with random initial phase setting of harmonic components. The modulation frequency f_m was 8 Hz and the modulation depth c_{AM} was 0.5. The waveform display of the snapshot clearly indicates rapid amplitude decay. The group delay display shows that the final static representation is shifted left (Energy centroid at each frequency, in other word, group delay, is biased backward because of the amplitude decay.).

3) Suppression of FM effects: Temporal variation of the fundamental frequency of the test signal also make cost L deteriorate. The following equation was used to generate test signals $x_{FM}(t)$ with frequency modulation of the fundamental frequency.

$$x_{FM}(t) = \sum_{k=0}^{\left\lfloor \frac{f_s}{2f_0} \right\rfloor} a_k \cos\left(\varphi_k + k\theta(t)\right), \tag{23}$$

$$\theta(t) = 2\pi \int_0^t \exp\left[\left(1 + c_{FM} \sin(2\pi f_m \tau)\right) \log(f_0)\right] d\tau, \tag{24}$$

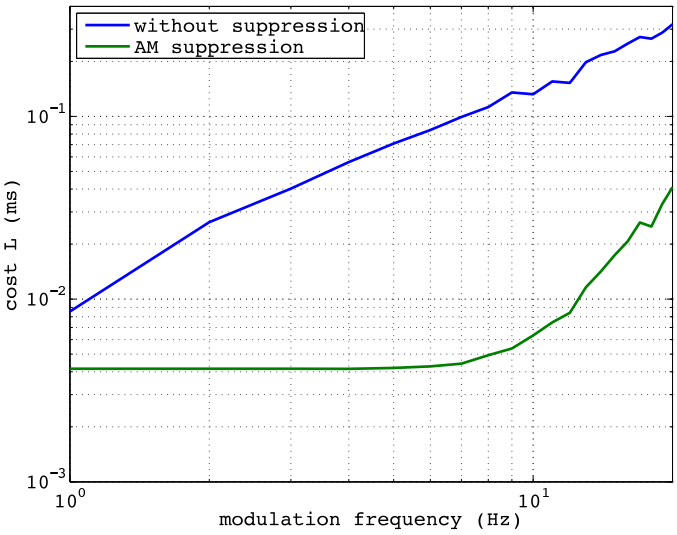

Fig. 12. Effect of AM and performance of AM supperssion.

where c_{FM} represents the frequency modulation depth and f_m represents the frequency of the fundamental frequency modulation.

4) Natural speech example: Fig. 15 shows an integrated display view of an analysis example of Japanese /a/ spoken by a male speaker. In this case, the static group delay represented by a thick blue line in the bottom right panel stays close to zero, even without AM and FM compensation, possibly because the signal is a sustained phonation.

VI. Discussion

The proposed group delay provides objective and quantitative means to represent deviation from periodicity in terms of HNR, since periodic signal yields constant output value zero. Effective insensitivity to phase and level of each harmonic

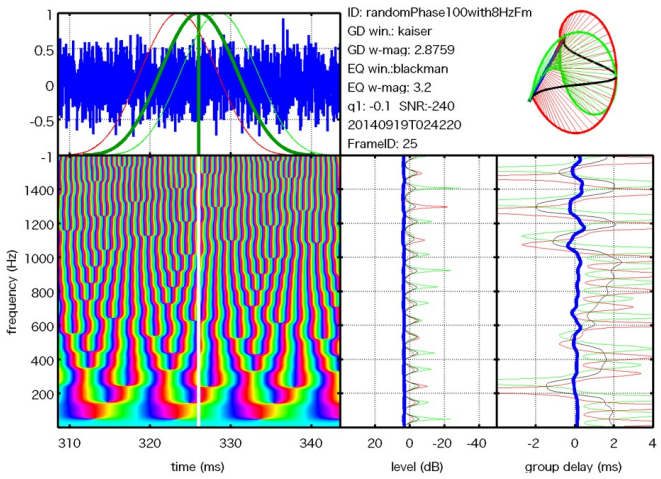

Fig. 13. Integrated display of the static group delay with additional intermediate information. The test signal is a periodic signal consisting of harmonically related sinusoids with random initial phase and applied FM with the following parameters ($f_m = 8$ Hz, $c_{FM} = \frac{1}{12}\log 2$).

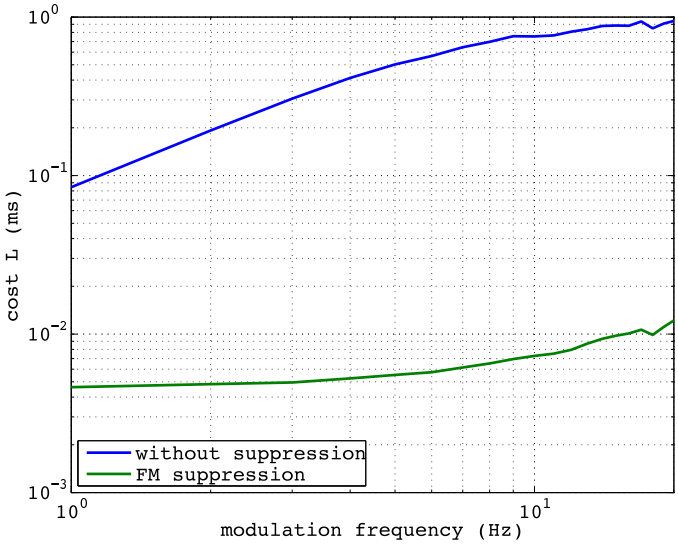

Fig. 14. Effect of FM and performance of FM supperssion.

component is a unique and valuable feature of the proposed representation. In addition of this feature, effects of known types of deviations such as AM and FM effects can be removed by introducing preprocessing procedures. These are useful for designing excitation source for resynthesis together with a group delay-based compensation, which is discussed in other articles [2], [4].

VII. Conclusions

A unified approach for designing aperiodic aspects of excitation source signals for high-quality speech analysis, modification and synthesis systems is introduced based on specially designed group delay representations. The temporally static group delay representation provides objective means

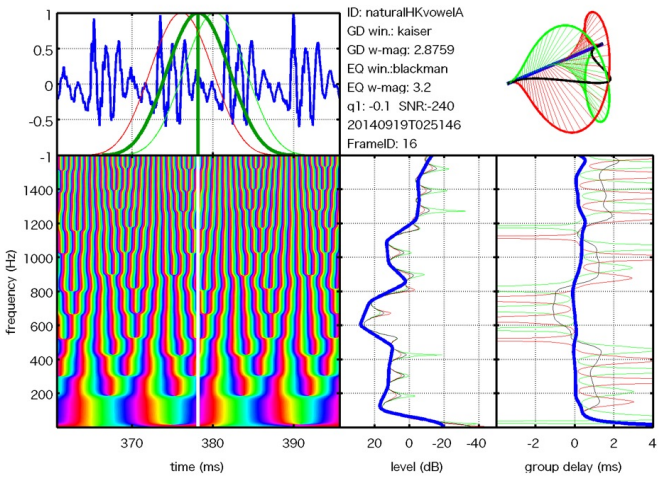

Fig. 15. Integrated display of an analysis example of sustained vowel /a/ spoken by a Japanese male speaker. Fundamental frequency of this example is 120 Hz.

for designing frequency distribution of aperiodicity and group delay-based compensation provides means to design temporal distribution of aperiodic energy. A series of systematic tests using subjective quality evaluation of synthesized speech sounds is currently undertaken.

Acknowledgment

This research is partly supported by Kakenhi (Aids for Scientific Research) of JSPS 24300073 and 24650085. The authors appreciate reviewers' constructive comments, which made the strength and impact of the proposed method clear and accessible. The authors also would like to thank Yegnanarayana for comments on the relation and role of the proposed method with his works on ZFF.

References

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

[2] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay," in *ICSLP 2000*, 2000, pp. 664–667.

[3] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J. C. Williams, "Noh voice quality," *Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[4] 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," *Proc. MAVEBA*, pp. 13–15, 2001.

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

[6] 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 *Proc. ICASSP 2008*, 2008, pp. 3933–3936.

[7] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," in *Proc. ICASSP 2011*, May 2011, pp. 5420–5423.

[8] H. Kawahara, "STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustic Science & Technology*, vol. 27, no. 5, pp. 349–353, 2006.

[9] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proc. ICASSP 2003*, vol. I, Hong Kong, 2003, pp. 256–259.

[10] S. R. Schweinberger, C. Casper, N. Hauthal, J. M. Kaufmann, H. Kawahara, N. Kloth, and D. M. Robertson, "Auditory adaptation in voice perception," *Current Biology*, vol. 18, pp. 684–688, 2008.

[11] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.

[12] 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.

[13] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, and R. Zäske, "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[14] 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 *Proc. Interspeech 2010*, 2010, pp. 38–41.

[15] R. McAulay and T. 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.

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

[17] J. Bonada, "High quality voice transformations based on modeling radiated voice pulses in frequency domain," in *Proc. Digital Audio Effects (DAFx)*, 2004.

[18] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085–2095, Oct 2013.

[19] D. P. Ellis, J. H. McDermott, and H. Kawahara, "Inharmonic speech: A tool for the study of speech perception and separation," in *Proc. SAPA-SCALE Conference 2012*, 2012, pp. 114–117.

[20] T. Nishi, R. Nisimura, T. Irino, and H. Kawahara, "Controlling linguistic information and filtered sound identity for a new cross-synthesis vocoder," *Acoustical Science and Technology*, vol. 34, no. 4, pp. 287–288, 2013.

[21] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *Signal Processing Magazine, IEEE*, vol. 24, no. 2, pp. 67–79, 2007.

[22] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1602–1613, 2008.

[23] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, July 2011. [Online]. Available: http://doi.org/10.1109/TASL.2010.2076806

[24] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[25] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, April 1967.

[26] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Am.*, vol. 82, no. 5, pp. 1560–1586, 1987.

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

[28] M. Morise, T. Takahashi, H. Kawahara, and T. Irino, "Power spectrum estimation method for periodic signals virtually irrespective to time window position," *Trans. IEICE*, vol. J90-D, no. 12, pp. 3265–3267, 2007, [in Japanese].

[29] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," in *Proc. ICASSP 2013*, 2013, pp. 6797–6801.

[30] ——, "An interference-free representation of group delay for periodic signals," in *Proc. APSIPA ASC 2012*, Dec 2012, pp. 1–4.

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

[32] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, November 1966.

[33] L. Cohen, *Time-frequency analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[34] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[35] 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.

[36] J. Kaiser and R. W. Schafer, "On the use of the i 0-sinh window for spectrum analysis," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 1, pp. 105–107, 1980.

[37] D. Slepian and H. O. Pollak, "Prolate spheroidal wave functions, fourier analysis and uncertainty–I," *Bell System Technical Journal*, vol. 40, no. 1, pp. 43–63, 1961.