Speech Communication 84 (2016) 57–65

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

D4C, a band-aperiodicity estimator for high-quality speech synthesis

Masanori Morise

Interdisciplinary Graduate School, University of Yamanashi, 4-3-11, Takeda, Kofu, Yamanashi, 400-8511, Japan

ARTICLE INFO

Article history:
Available online 14 September 2016

Keywords:
Speech analysis
Speech synthesis
Aperiodicity
Periodic signal
Group delay

ABSTRACT

An algorithm is proposed for estimating the band aperiodicity of speech signals, where "aperiodicity" is defined as the power ratio between the speech signal and the aperiodic component of the signal. Since this power ratio depends on the frequency band, the aperiodicity should be given for several frequency bands. The proposed D4C (Definitive Decomposition Derived Dirt-Cheap) estimator is based on an extension of a temporally static group delay representation of periodic signals. In this paper, the principle and algorithm of D4C are explained, and its effectiveness is discussed with reference to objective and subjective evaluations. Evaluation results indicate that a speech synthesis system using D4C can synthesize natural speech better than ones using other algorithms.

1. Introduction

There have been many studies in speech analysis based on Vocoder (Dudley, 1939). It basically consists of the fundamental frequency (F0) and spectral envelope estimators while the modern framework uses the aperiodic parameter to improve the sound quality of synthesized speech. In spectral envelope estimation, linear predictive coding (LPC) (Atal and Hanauer, 1971) and Cepstrum representation (Oppenheim, 1969) are typical estimators. Although the speech synthesis system of these conventional estimators was unable to achieve speech synthesis with quality as high as that of a waveform-based synthesizer (Black and Campbell, 1995), the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) (Kawahara et al., 1999) algorithm was able to achieve equivalent sound quality. STRAIGHT has been used as the fundamental framework for voice conversion such as voice morphing (Kawahara et al., 2009), and TANDEM-STRAIGHT (Kawahara and Morise, 2011; Kawahara et al., 2008) was proposed as the next version of STRAIGHT. For achieving higher-quality speech synthesis, the F0-Adaptive Multi-Frame Integration Analysis method (Nakano and Goto, 2012) and the CheapTrick spectral envelope estimator (Morise, 2015a,b) are still being studied as next-generation spectral envelope estimators.

Aperiodicity is a spectral parameter associated with mixed excitation (McCree and Barnwell, 1995) and is used as a parameter for voice conversion (Ohtani et al., 2006). Several aperiodicity estimators have been proposed (Deshmukh and Wilson, 2003; Griffin and Lim, 1985, 1988) for speech analysis and speech coding. For high-quality speech synthesis, STRAIGHT (Kawahara et al., 2001) (which it is called Legacy-STRAIGHT to distinguish it from TANDEM-STRAIGHT) and TANDEM-STRAIGHT (Kawahara and Morise, 2012) use original algorithms. Improving sound quality requires not only the F0 and spectral envelope, but also an aperiodicity estimator.

The goal in this study is to develop a Vocoder-based high-quality speech synthesis system. Here we propose an algorithm for estimating the band aperiodicity of speech signals for use in high-quality speech synthesis. We call it the D4C (Definitive Decomposition Derived Dirt-Cheap) estimator.

The rest of this paper is organized as follows. In Section 2, we define the problem to be solved, the concept of the algorithm used, and the equations used in the paper. In Section 3, we describe the algorithm in detail. In Section 4, we evaluate the estimator's performance objectively and subjectively. We conclude in Section 5 with a brief summary and a mention of future work.

2. Problem to be solved and definition of parameters used for algorithm

In this section, we define voiced sound and describe our derivation of the algorithm used. After that, we explain the concept of D4C and the parameters used for it. D4C uses a group-delay-based parameter. This parameter forms a sine wave of F0 Hz from arbitrary periodic signals with a fundamental period of T_0. Therefore, the power ratio between the sine wave and the other frequency components corresponds to the aperiodicity. We can obtain the band aperiodicity by limiting the frequency band used for the calculation. Furthermore, we can obtain the same result independently of the temporal position used for windowing. One problem in periodic signal analysis is that the windowed signal

E-mail address: mmorise@yamanashi.ac.jp

http://dx.doi.org/10.1016/j.specom.2016.09.001

depends on the temporal position. This means that the estimated speech parameter also depends on the temporal positions even if the signal does not change F0 and the spectral envelope. D4C overcomes this problem by using temporally static representation.

2.1. Definition of periodic signal and problem to be solved

In the Vocoder-based approach, voiced sound $y(t)$ is defined as the convolution of an impulse response $h(t)$ and a pulse train with a fundamental period of T_0. Signal waveform $y(t)$ and its spectrum $Y(\omega)$ are given by

$$y(t) = h(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT_0), \tag{1}$$

$$Y(\omega) = \frac{2\pi}{T_0} H(\omega) \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0), \tag{2}$$

$$H(\omega) = A(\omega) e^{j\phi(\omega)}, \tag{3}$$

where symbol $*$ represents convolution, and ω_0 represents the fundamental angular frequency ($=2\pi/T_0$). $A(\omega)$ and $\phi(\omega)$ represent the amplitude and phase spectrum, respectively. This equation shows that spectrum $Y(\omega)$ consists of fundamental and harmonic components. The arbitrary sequences associated with amplitude α_n and phase β_n are used to normalize $Y(\omega)$:

$$Y(\omega) = \sum_{n=-\infty}^{\infty} \alpha_n e^{j\beta_n} \delta(\omega - n\omega_0). \tag{4}$$

One aperiodic component of actual speech is noise, and aperiodicity is defined as the power ratio between the speech signal and the aperiodic component of the signal. Since this power ratio depends on the frequency band, the aperiodicity should be given for several frequency bands.

2.2. Concept of algorithm used

There are several algorithms for estimating aperiodicity (Deshmukh and Wilson, 2003; Kawahara et al., 2001; Kawahara and Morise, 2012). Legacy-STRAIGHT and TANDEM-STRAIGHT are mainly used for high-quality speech synthesis. Since Legacy-STRAIGHT requires post-processing for temporal smoothing after frame-by-frame processing with a 1 ms frame shift, it is not suitable for real-time applications (Banno et al., 2007; Morise et al., 2009). TANDEM-STRAIGHT is based on a waveform-based approach and uses F0 information as input. Specifically, its estimation performance is greatly degraded when the estimated F0 contour includes an error. D4C overcomes these problems by utilizing temporally static parameters (Kawahara et al., 2012, 2014).

2.3. Definitions of fundamental equations

The equations used in the algorithm are defined in this section. A fundamental discussion can be found elsewhere (Cohen, 1994). Group delay $\tau_g(\omega)$ is a parameter defined on the basis of the frequency derivation of phase $\phi(\omega)$.

$$\tau_g(\omega) = -\phi'(\omega), \tag{5}$$

$$\phi'(\omega) \equiv \frac{d\phi(\omega)}{d\omega},$$

where symbol $'$ represents the derivation of the frequency domain. The other approach is defined as follows:

$$\tau_g(\omega) = \frac{\Re\left(S'(\omega)\right)\Im(S(\omega)) - \Re(S(\omega))\Im\left(S'(\omega)\right)}{|S(\omega)|^2}, \tag{6}$$

where $S(\omega)$ represents the spectrum of input signal $s(t)$. Derivative of the spectrum $S(\omega)$, $S'(\omega)$ is given by the spectrum of signal $-jts(t)$.

$$S'(\omega) = \mathcal{F}[-jts(t)], \tag{7}$$

where symbol $\mathcal{F}[]$ represents the Fourier transform. The D4C algorithm uses Eq. (6) as a fundamental parameter. Temporally static representations (Kawahara et al., 2012) are used in both the numerator and denominator. A temporally static parameter based on the group delay is used for the aperiodicity estimation.

3. Algorithm details

In periodic signal analysis, the windowed waveform and its spectrum depend on the temporal position used for windowing. Even if the target spectral envelope is temporally static, the estimation result depends on the temporal position. Temporally static representation of the spectral envelope is important for high-quality speech synthesis. It is also important for aperiodicity estimation.

The D4C algorithm uses pitch synchronous analysis (Mathews et al., 1961) for designing the window function and a new parameter based on the temporally static group delay (Kawahara et al., 2012). D4C consists of three steps.

3.1. First step: calculation of temporally static parameter on basis of group delay

First, the numerator of Eq. (6) is defined as $E_{cs}(\omega)$.

$$E_{cs}(\omega) = \Re\left(S'(\omega)\right)\Im(S(\omega)) - \Re(S(\omega))\Im\left(S'(\omega)\right). \tag{8}$$

We assume the calculation of E_{cs} from periodic signal $y(t)$ (Eq. (1)). Eq. (2) shows that $Y(\omega)$ has harmonic components at $n\omega_0$ (n: integer value). Each component has a different amplitude and phase. We first design a window function that fulfills the following two requirements.

- The main-lobe bandwidth is ω_0.
- The amplitude of the side lobes is negligibly low compared with that of the main lobe.

Signal waveform $y(t)$ is windowed using the window function centered at τ. The windowed waveform is defined as $y(t, \tau)$.

The spectrum of the windowed waveform $Y(\omega, \tau)$ is the convolution of the spectrum of the window function and $Y(\omega)$.

$$\begin{aligned} Y(\omega, \tau) &= Y(\omega) * W(\omega) e^{-j\omega\tau}, \\ &= W(\omega) e^{-j\omega\tau} * \sum_{n=-\infty}^{\infty} \alpha_n e^{j\beta_n} \delta(\omega - n\omega_0). \end{aligned} \tag{9}$$

Fig. 1 illustrates examples of $Y(\omega)$ and $Y(\omega, \tau)$. $Y(\omega, \tau)$ has harmonic components at $n\omega_0$, but one harmonic $k\omega_0$ is spread across the frequency range from $(k-1)\omega_0$ to $(k+1)\omega_0$ by convolving the window function. Normally, interference among harmonic components must be calculated, but the window function used for windowing can simplify this calculation. We can calculate the effect of interference between neighboring components $k\omega_0$ and $(k+1)\omega_0$. To simplify the explanation, we explain the case for $k = 0$.

$$\begin{aligned} Y(\omega, \tau) &= \left(\delta(\omega) + \alpha e^{j\beta} \delta(\omega - \omega_0)\right) * W(\omega) e^{-j\omega\tau}, \\ &= W(\omega) e^{-j\omega\tau} + \alpha W(\omega - \omega_0) e^{-j(\omega\tau - \omega_0\tau - \beta)}, \end{aligned} \tag{10}$$

where $W(\omega)$ represents the spectrum of the window function, and α and β correspond to amplitude α_1 and phase β_1. Since the relative difference is important, α_0 and β_0 are normalized to 1 and 0, respectively. The component including τ is the time-varying component, and we remove it using the following derivation.

To remove the time-varying component, we insert parameter C into Eq. (7).

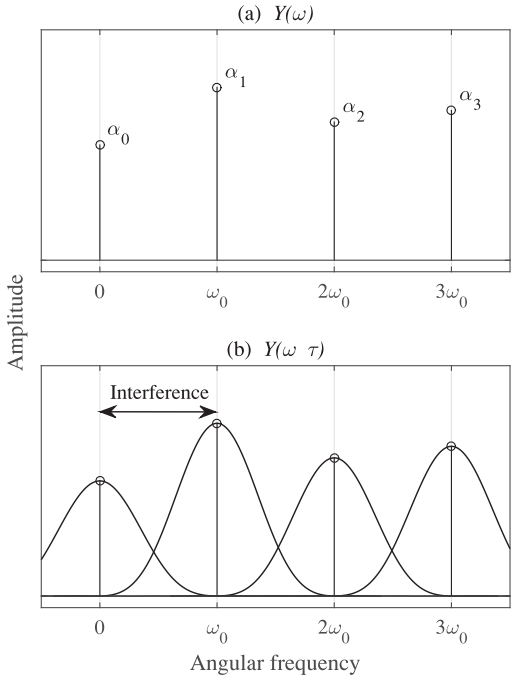

Fig. 1. Examples of $Y(\omega)$ and $Y(\omega, \tau)$. In spectrum $Y(\omega, \tau)$, the components between neighboring harmonics interfere due to convolving of window function. Value between them therefore depends on temporal position τ. We can discuss only interference between $k\omega_0$ and $(k+1)\omega_0$ due to features of main and side lobes of window function.

$$Y_0(\omega, \tau) = \mathcal{F}[-j(t+C)y(t, \tau)] = Y'(\omega, \tau) - jCY(\omega, \tau), \tag{11}$$

$$\begin{aligned} Y'(\omega, \tau) = & -j\tau e^{j\omega\tau}W(\omega) + e^{-j\omega\tau}W'(\omega) \\ & - j\alpha\tau e^{j(\omega\tau-\omega_0\tau-\beta)}W(\omega-\omega_0) \\ & + \alpha e^{-j(\omega\tau-\omega_0\tau-\beta)}W'(\omega-\omega_0), \end{aligned} \tag{12}$$

where $W'(\omega)$ represents the frequency derivation of $W(\omega)$. Parameter C corresponds to the temporal shift of the signal. It is thus essential to obtain the temporally static parameter on the basis of the group delay. Here we use $Y_0(\omega, \tau)$ as the frequency derivation of $Y(\omega, \tau)$. We assume that $W(\omega)$ has no imaginary part. A window function that is temporally symmetrical and ranges from $-N/2$ to $N/2$ fulfills this assumption. N represents the width of the window function.

Both the real and imaginary parts of $Y(\omega, \tau)$ and $Y_0(\omega, \tau)$ are used to calculate $E_{cs}(\omega, \tau)$.

$$\begin{aligned} \Re(Y(\omega, \tau)) = & W(\omega)\cos(\omega\tau) \\ & + \alpha W(\omega-\omega_0)\cos(\omega\tau-\omega_0\tau-\beta), \end{aligned} \tag{13}$$

$$\begin{aligned} \Im(Y(\omega, \tau)) = & -W(\omega)\sin(\omega\tau) \\ & - \alpha W(\omega-\omega_0)\sin(\omega\tau-\omega_0\tau-\beta), \end{aligned} \tag{14}$$

$$\begin{aligned} \Re(Y_0(\omega, \tau)) = & W'(\omega)\cos(\omega\tau) - \tau W(\omega)\sin(\omega\tau) \\ & + \alpha W'(\omega-\omega_0)\cos(\omega\tau-\omega_0\tau-\beta) \\ & - \tau\alpha W(\omega-\omega_0)\sin(\omega\tau-\omega_0\tau-\beta) \\ & - C\alpha W(\omega-\omega_0)\sin(\omega\tau-\omega_0\tau-\beta) \\ & - CW(\omega)\sin(\omega\tau), \end{aligned} \tag{15}$$

$$\begin{aligned} \Im(Y_0(\omega, \tau)) = & -W'(\omega)\sin(\omega\tau) - \tau W(\omega)\cos(\omega\tau) \\ & - \alpha W'(\omega-\omega_0)\sin(\omega\tau-\omega_0\tau-\beta) \\ & - \tau\alpha W(\omega-\omega_0)\cos(\omega\tau-\omega_0\tau-\beta) \\ & - C\alpha W(\omega-\omega_0)\cos(\omega\tau-\omega_0\tau-\beta) \\ & - CW(\omega)\cos(\omega\tau). \end{aligned} \tag{16}$$

We can obtain the following equation by plugging these four terms into Eq. (8).

$$\begin{aligned} E_{cs}(\omega, \tau) = & (C+\tau)W^2(\omega) + \alpha^2(C+\tau)W^2(\omega-\omega_0) \\ & + 2W(\omega)W(\omega-\omega_0)\alpha(C+\tau)\cos(\omega_0\tau-\beta) \\ & + \alpha(W'(\omega)W(\omega-\omega_0) - W(\omega)W'(\omega-\omega_0)) \\ & \times \sin(\omega_0\tau-\beta). \end{aligned} \tag{17}$$

Eq. (11) includes the Fourier transform, and its integration range is defined as $(-\infty, \infty)$. This range is limited to $(\tau - N/2,\ \tau + N/2)$ because the signal is windowed by the window function shifted by τ. Since parameter C corresponds to the temporal shift, the integration range in Eq. (11) is equivalent to $(\tau + C - N/2,\ \tau + C + N/2)$. Once the integration range is set to $(\tau_0 - N/2,\ \tau_0 + N/2)$, parameter C is automatically set to $-\tau + \tau_0$. The following equation is obtained by using the integration range $(\tau_0 - N/2,\ \tau_0 + N/2)$.

$$\begin{aligned} E_{cs}(\omega, \tau) = & \tau_0 W^2(\omega) + \tau_0\alpha^2 W^2(\omega-\omega_0) \\ & + 2W(\omega)W(\omega-\omega_0)\alpha\tau_0\cos(\omega_0\tau-\beta) \\ & + \alpha(W'(\omega)W(\omega-\omega_0) - W(\omega)W'(\omega-\omega_0)) \\ & \times \sin(\omega_0\tau-\beta). \end{aligned} \tag{18}$$

The important thing is that τ_0 can be set to an arbitrary value not related to temporal position τ. Parameter C is automatically set to the value for canceling temporal position τ. Therefore, the first and second terms in the equation are temporally static.

In this equation, the signs of the sin and cos terms are reversed by calculating $E_{cs}(\omega, \tau + T_0/2)$. The third and fourth terms are canceled by calculating $E_{cs}(\omega, \tau) + E_{cs}(\omega, \tau + T_0/2)$. In the following equation, the temporal positions are shifted so that the centroid of the analysis positions is τ.

$$\begin{aligned} E_D(\omega, \tau) = & E_{cs}\left(\omega, \tau - \frac{T_0}{4}\right) + E_{cs}\left(\omega, \tau + \frac{T_0}{4}\right) \\ = & 2\tau_0 W^2(\omega) + 2\tau_0\alpha^2 W^2(\omega-\omega_0). \end{aligned} \tag{19}$$

$E_D(\omega, \tau)$ is the parameter used as the numerator of Eq. (6). $E_D(0, \tau)$ and $E_D(\omega_0, \tau)$ are $2\tau_0 W^2(0)$ and $2\tau_0\alpha^2 W^2(0)$, respectively. This suggests that the values of each harmonic component and the powers of $Y(\omega)$ are virtually equal. The equation shows the result for $k = 0$. This equation is generalized for an arbitrary k.

$$E_D(\omega, \tau) = 2\tau_0 W^2(\omega - k\omega_0) + 2\tau_0\alpha^2 W^2(\omega - (k+1)\omega_0). \tag{20}$$

This equation is effective provided that the main lobe of the window is below ω_0 and the amplitude of the side lobes is negligible. In practical use, a window function is used in which the side lobes are low enough to achieve the required performance. Although a long window function can better approximate the requirements for both the main and side lobes, it is not appropriate because the parameters of speech temporally change. Since there are many window functions that fulfill the requirements, we carried out an exploratory experiment including an unofficial listening test with limited speech and few subjects. On the basis of the results, we decided to use the Blackman window with a length of $4T_0$. The other parameters used for the D4C algorithm were also based on the results.

The temporally static parameter on the basis of the group delay is obtained by calculating the temporally static power spectrum and using it as the denominator of Eq. (6). We next explain how to calculate the temporally static power spectrum.

Since this power spectrum is used as the denominator of Eq. (6), the full frequency band should not include any zeros. For D4C estimation, we focus on limiting the window function and thus use a Hanning window with a length of $4T_0$. Since the main lobe of

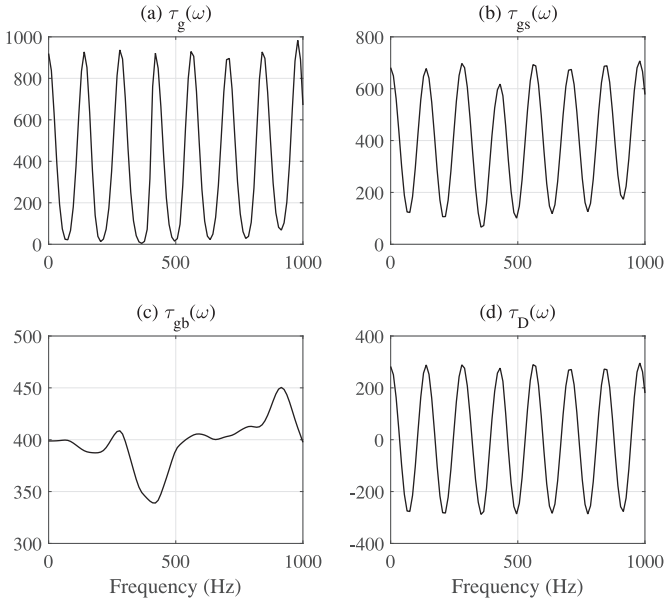

Fig. 2. Examples of four parameters.

this window function is $\omega_0/2$, the main lobe does not cause interference between neighboring harmonic components. The effect of interference between the side lobes is therefore below 30 dB, which is the side lobe amplitude. This effect is ignored here. Zeros in the spectrum are removed by a simple spectral smoothing with a rectangular window that has a width of ω_0.

$$P_s(\omega,\tau)=\frac{1}{\omega_0}\int_{-\frac{\omega_0}{2}}^{\frac{\omega_0}{2}}P(\omega+\lambda,\tau)d\lambda, \tag{21}$$

where $P(\omega,\ \tau)$ represents the power spectrum of the waveform windowed at τ. We can ignore the interference between neighboring harmonic components because of the main lobe of the window function and the width of the smoothing.

A temporally static parameter based on group delay is given by

$$\tau_g(\omega,\tau)=\frac{E_D(\omega,\tau)}{P_s(\omega,\tau)}. \tag{22}$$

In both the numerator and denominator, the values at $n\omega_0$ are the same. Eq. (22) therefore shows that the values of all harmonic components are the same. This means that Eq. (22) outputs the same result for arbitrary signals with the same F0.

With Eq. (22), the value at $(k+0.5)\omega_0$ is smaller than that at $k\omega_0$, provided that window functions shown in the paper are used. This means that $\tau_g(\omega,\ \tau)$ is shaped as a periodic signal with a period of ω_0. Here we neglect temporal position τ and use $\tau_g(\omega)$ because $\tau_g(\omega,\ \tau)$ does not depend on the temporal position.

3.2. Second step: calculation of parameter shaping

In the first step, we calculate the temporally static parameter on the basis of group delay $\tau_g(\omega)$. In the second step, $\tau_g(\omega)$ is converted into the parameter used for the calculation in the third step. We explain how this is done by using the examples in Fig. 2, which shows the parameters for frames calculated from actual speech.

Since $\tau_g(\omega)$ (panel (a) in Fig. 2) is a periodic signal with a period of ω_0, its waveform is obtained by inverse Fourier transformation and has peaks at nT_0. Eqs. (20) and (21) show that the frequency fluctuation depends on the spectrum of the window function. Here Blackman and Hanning windows are used for calculating $E_D(\omega)$ and $P_s(\omega)$, respectively. These spectra therefore consist of superimposed sinc functions.

In the second step, the component of nT_0 ($n>1$) is reduced so that $\tau_g(\omega)$ is shaped into a sine wave. The component is reduced by the spectral smoothing given by

$$\tau_{gs}(\omega)=\frac{2}{\omega_0}\int_{-\frac{\omega_0}{4}}^{\frac{\omega_0}{4}}\tau_g(\omega+\lambda)d\lambda. \tag{23}$$

The waveform of the smoothing function has zeros at $2nT_0$ (n: natural number). Although this smoothing cannot reduce the component at $(2n+1)T_0$, we neglect this effect because powers of $\tau_g(\omega)$ above $3T_0$ are negligibly low.

Panel (b) in Fig. 2 shows the low frequency component. The component is removed by using the following equations.

$$\tau_D(\omega)=\tau_{gs}(\omega)-\tau_{gb}(\omega), \tag{24}$$

$$\tau_{gb}(\omega)=\frac{1}{\omega_0}\int_{-\frac{\omega_0}{2}}^{\frac{\omega_0}{2}}\tau_{gs}(\omega+\lambda)d\lambda, \tag{25}$$

where $\tau_D(\omega)$ represents the parameter used for the third step. Panels (c) and (d) in Fig. 2 show examples of $\tau_{gb}(\omega)$ and $\tau_D(\omega)$, respectively. We can obtain a sine wave with a frequency of ω_0 from any periodic signal. The amount of power for the noise excluding the sine wave is used to estimate the aperiodicity.

3.3. Third step: estimation of band-aperiodicity

Parameter $\tau_D(\omega)$ fits a sine wave with a frequency of ω_0 provided that the input signal does not contain any aperiodic noise. Band aperiodicity is calculated as the power ratio between total power and the power of the sine wave for each frequency band. The D4C algorithm can therefore estimate the band aperiodicity for a center frequency with a certain bandwidth. To calculate the power of the sine wave, the waveform is windowed using a window function with low side lobes. The power at around T_0 calculated on the basis of the main-lobe bandwidth represents the target power.

In the D4C algorithm, a Nuttall window (Nuttall, 1981) is used as the window function with low side lobes. The side lobes are 90 dB lower than the main lobe, and the window function works better than a Blackman window. Parameter $\tau_D(\omega)$ is windowed using window function $w(\omega)$ shifted to an arbitrary frequency ω_c; ω_c is an angular frequency and equals $2\pi f_c$ Hz, where f_c is the center frequency. Waveform $p(t,\ \omega_c)$ is calculated using the inverse Fourier transform.

$$p(t,\omega_c)=\mathcal{F}^{-1}\left[w(\omega)\tau_D\left(\omega-\left(\omega_c-\frac{w_l}{2}\right)\right)\right], \tag{26}$$

where w_l represents the length of the window function, and $\mathcal{F}^{-1}$ represents the inverse Fourier transform. Fig. 3 shows examples of two parameters calculated from the actual speech corresponding to the signals in Fig. 2. To simplify the discussion, the total energy of $|p(t,\ \omega_c)|^2$ was normalized to 1. Power waveform $|p(t,\ \omega_c)|^2$ (panel (a)) was calculated from $p(t,\ \omega_c)$, and parameter $p_c(t,\ \omega_c)$ (panel (b)) was calculated using

$$p_c(t,\omega_c)=1-\int_0^t p_s(\lambda,\omega_c)d\lambda, \tag{27}$$

where $p_s(t,\ \omega_c)$ represents the parameter calculated by sorting $|p(t,\ \omega_c)|^2$ in descending order on the time axis. Band aperiodicity $ap(\omega_c)$ is given by

$$ap(\omega_c)=-10\log_{10}\left(p_c(2w_{bw},\omega_c)\right), \tag{28}$$

where w_{bw} represents the main-lobe bandwidth of window function $w(\omega)$. The dimension of w_{bw} is time. Since the main-lobe bandwidth is defined as the shortest frequency range from 0 Hz to the frequency at which the amplitude indicates 0, $2w_{bw}$ is used for Eq. (28).

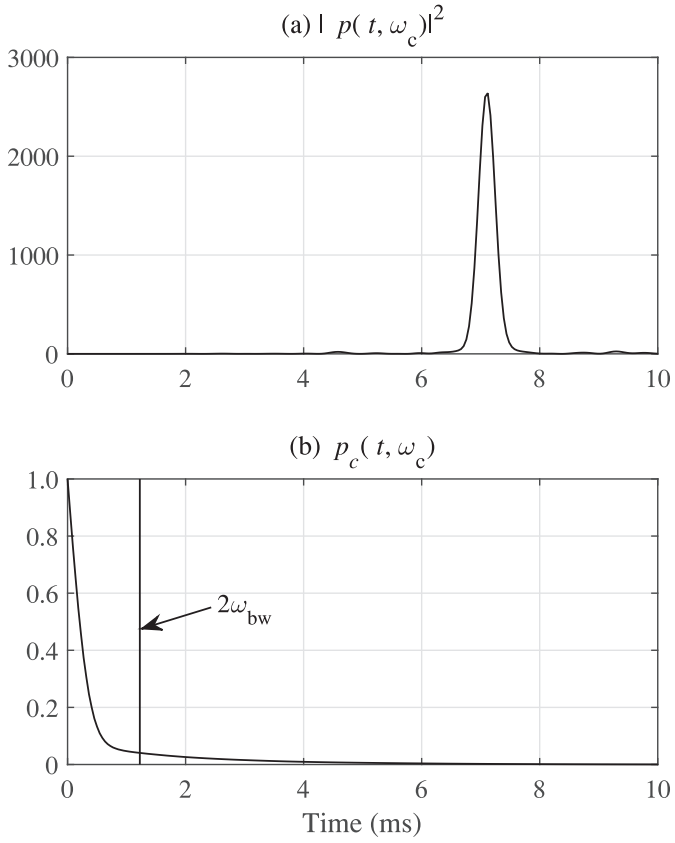

Fig. 3. Examples of $|p(t, \omega_c)|^2$ and $p_c(t, \omega_c)$. Total energy of waveforms was normalized to 1.

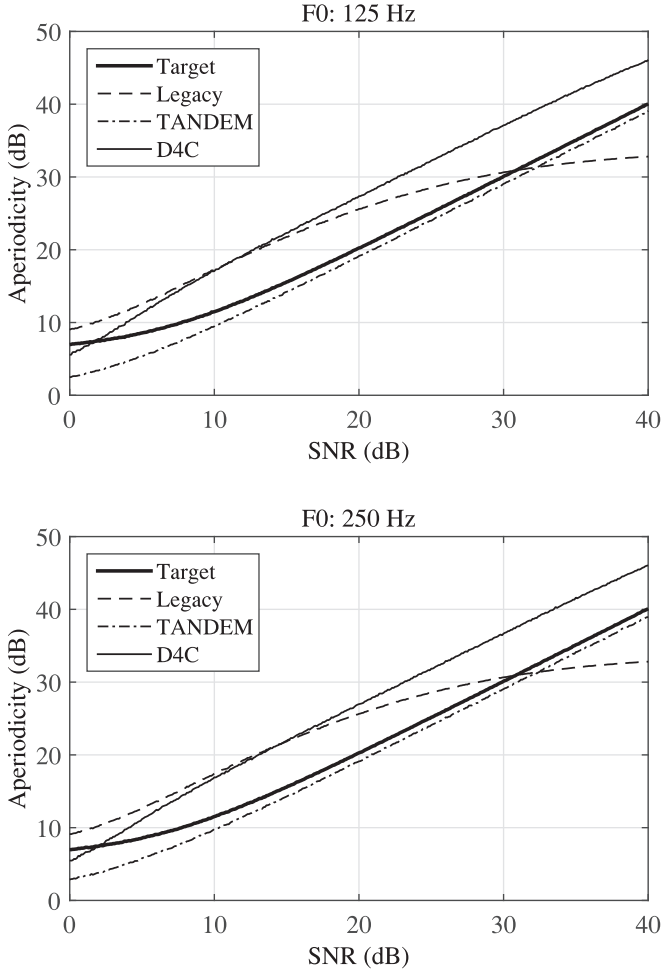

Fig. 4. Relationship between the SNR and estimated aperiodicity.

4. Evaluation

The effectiveness of D4C was investigated both objectively and subjectively. Since the goal of this study is to develop a Vocoder-based high-quality speech synthesis system, subjective evaluation using synthesized speech is important. On the other hand, objective evaluation is also important to quantify the performance. In actual speech, the F0 is time-varying, and the F0 contour estimated from the speech waveform usually contains an error. The purpose of the objective evaluation was to quantify both the basic performance and robustness against these factors.

4.1. Objective evaluation

4.1.1. Common conditions

In the evaluation, Legacy-STRAIGHT and TANDEM-STRAIGHT were compared. The sampling frequency of the signal used for the evaluation was 48 kHz. The signal length was set to 1 s, and the frame shift was set to 1 ms. The number of estimation results for each center frequency was 1000. The fast Fourier transform length used for D4C was set to 4096.

The length of the window function, w_l, was set to 6 kHz. Five center frequencies (3, 6, 9, 12, and 15 kHz) were used. In the objective evaluation, since the other algorithms output the aperiodicity of the spectral representation, they were discretized for each center frequency. In the subjective evaluation, linear interpolation was used in D4C to obtain the aperiodicity of the spectral representation. For the interpolation, we added the aperiodicity of a value (60 dB) at 0 Hz on the basis of our past research (Kawahara and Morise, 2012). The validity of these parameters will be discussed on the basis of the results.

4.1.2. Experiment 1: relationship between SNR and estimated results

The first experiment was carried out to determine the relationship between the SNR and the estimated results. The signal used is given by

$$y(t) = n(t) + \sum_{k=0}^{K} \cos(k\omega_0 t + \theta_k), \tag{29}$$

where $n(t)$ represents white noise, and θ_k represents the phase characteristic in each component. The θ_k were set to random values, and K was set to the maximum value at which $K\omega_0$ does not exceed the Nyquist frequency (24 kHz). Since the aperiodicities had the same value for all center frequencies, conclusive results were defined as the averages for all center frequencies and frames. The SNR was set from 0 to 40 dB, and F0s of 125 and 250 Hz were used.

The results are plotted in Fig. 4. The thick line shows the target aperiodicity, calculated using $n(t)$ and $y(t)$ in Eq. (29). Their power spectra, $|N(\omega)|^2$ and $|Y(\omega)|^2$, were used to calculate the target aperiodicity for each frequency band. The center frequency and bandwidth were equivalent to the parameters of the aperiodicity estimators.

Since similar results were observed for both F0s, we discuss only the result for 125 Hz. Legacy-STRAIGHT estimated the aperiodicity from 0 to around 20 dB and could not estimate it above 30 dB. TANDEM-STRAIGHT estimated the aperiodicity most accurately, and the estimated aperiodicity around 0 dB was below the target. D4C estimated the aperiodicity from 0 to 40 dB.

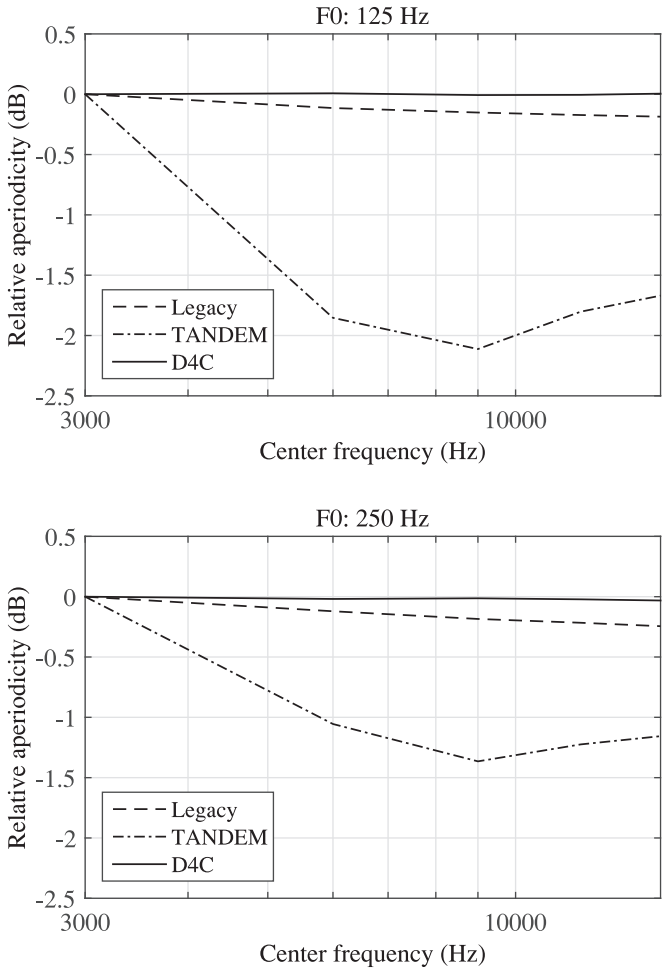

Fig. 5. Relative aperiodicities for center frequencies (white noise).

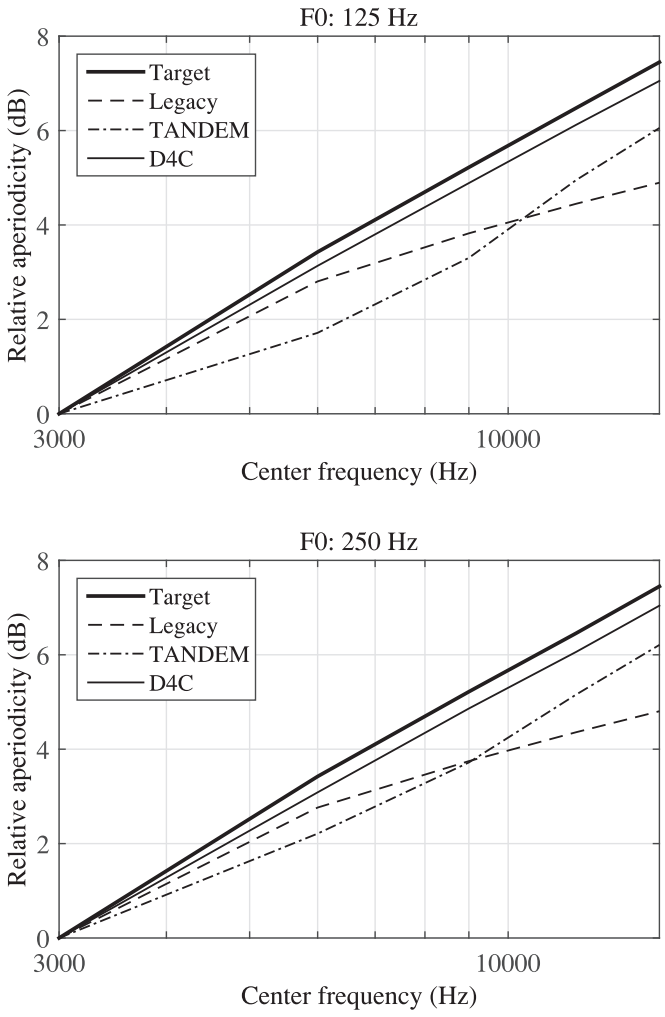

Fig. 6. Relative aperiodicities for center frequencies (pink noise).

4.1.3. Experiment 2: estimation performance for each center frequency using different noise signals

The second experiment was carried out to evaluate the estimation results for each center frequency. The signal was designed using Eq. (29), and either white or pink noise was used as $n(t)$. The relative aperiodicity was calculated for each center frequency under the same F0 conditions as those used in the first experiment. Since similar tendencies were observed, the SNR was set to 20 dB in the three subsequent experiments.

The results when white noise was used are plotted in Fig. 5. The results for all algorithms at 3 kHz were set to 0 dB and used as a baseline for comparison. Since the SNRs for all center frequencies were the same, the ideal result is 0 dB for all frequencies. TANDEM-STRAIGHT had the largest error while D4C had an error smaller than Legacy-STRAIGHT, but the difference was only about 0.2 dB.

The results when pink noise was used are plotted in Fig. 6. The target aperiodicity is represented by the thick line. D4C estimated the aperiodicity for all center frequencies more accurately than the other algorithms. While both sets of results show that D4C had the best performance, the differences among them were small.

4.1.4. Experiment 3: effect of F0 estimation error

Since the three algorithms use F0 for estimation, we evaluated them for robustness against F0 estimation error. Experiment 3 was done the same as experiment 1 except that the input F0 had an error (from −10 to 10%).

The results are plotted in Fig. 7. The relative error (shown on the y-axis) was defined as the difference between the estimated aperiodicity and the aperiodicity for an F0 error of 0%. TANDEM-STRAIGHT tended to have the highest error. Legacy-STRAIGHT had an error of around ± 4 dB, and that of D4C was within 3 dB. TANDEM-STRAIGHT is thus less robust against F0 estimation error than the other two algorithms.

4.1.5. Experiment 4: effect of F0 contour frequency modulation (FM)

Since the F0 contour of actual speech temporally varies, this effect should be measured to evaluate the effectiveness of D4C. In this experiment, the F0 contours of test signals were designed using the following equation, which includes FM parameter α for controlling the modulation gradient.

$$f_0(t) = f + \sqrt{\alpha f}\cos\left(\sqrt{\alpha f}t + \theta\right), \tag{30}$$

where f represents the standard F0, and θ represents the phase. The F0 contour had a maximum gradient of αf. The value of FM parameter α ranged from 0.0 to 25.0, and a θ from 0.0 to 2π was used to calculate the F0 contour. The values calculated for all θ were averaged and used as the result.

The results are plotted in Fig. 8. In the figure, F0 corresponds to f. All the results were subtracted from what so that the result for α of 0 corresponds to 0 dB. The results should therefore indicate 0 dB regardless of α.

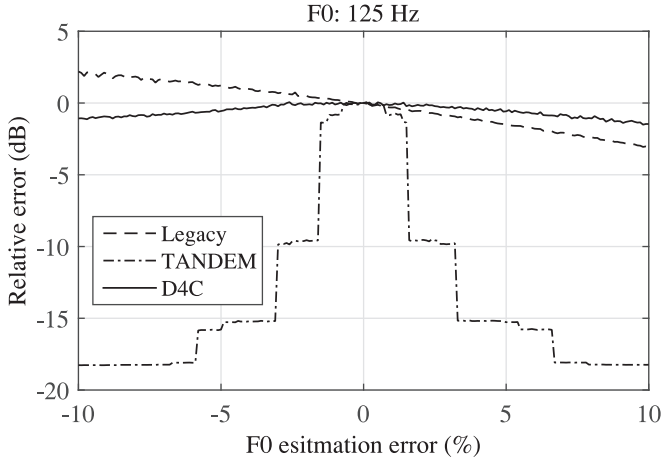

Fig. 7. Effect of F0 estimation error on estimation performance.

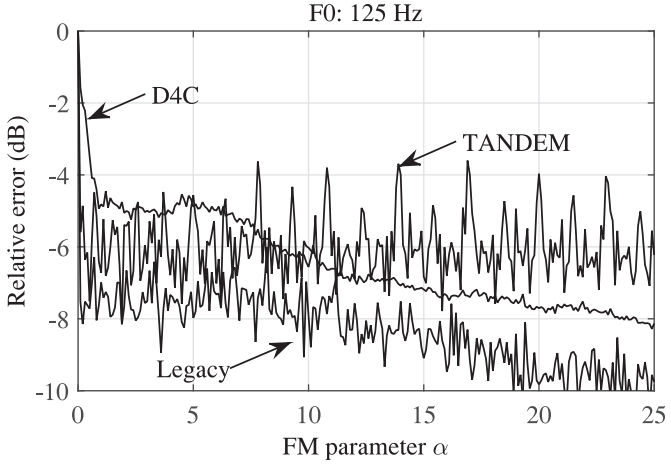

Fig. 8. Relationship between FM parameter and estimation result.

The results show that D4C was the most effective of the three algorithms provided that the FM parameter was under 5. When the parameter was above 10, TANDEM-STRAIGHT was the most effective. However, TANDEM-STRAIGHT was more sensitive to α than the other two algorithms.

4.1.6. Experiment 5: effect of vocal cord vibration amplitude modulation (AM)

In the final experiment, a periodic signal in which the amplitude of each pulse differed was used to evaluate robustness against AM of vocal cord vibration. The signal was designed using

$$y(t) = \sum_{n=-\infty}^{\infty} (\delta(t - 2nT_0) + \beta\delta(t - (2n+1)T_0)), \tag{31}$$

where β represents the AM parameter. The value of β ranged from 1.0 to 1.2.

The results are plotted in Fig. 9. The results were subtracted from what so that the result for β of 1 corresponds to 0 dB. They clearly show that D4C was the most robust against AM regardless of the F0.

4.2. Discussion for objective evaluation

The results of experiment 1 show that the aperiodicity estimated by D4C had a bias when the SNR was above 5 dB. Since the bias was around 6 dB, we can compensate for it by subtracting 6 dB from the estimated aperiodicity. When the SNR was below 5 dB, the trend in the bias differed, but an algorithm used for evaluation cannot accurately estimate F0 from low-SNR speech. A frame with this SNR is identified as an unvoiced part, and its aperiodicity is not used in speech synthesis. We covered this range in our experiments because the purpose of our objective evaluation was to evaluate the performance of D4C. The results for SNR below 5 dB do not affect our subjective evaluation.

The F0 estimation error and AM results showed that D4C had the best performance. The other results showed that its performance was similar to those of the other two algorithms. Since actual speech does not have perfect periodicity, the temporal positions of vocal cord vibration include temporal fluctuation. Subjective evaluation with re-synthesized speech is therefore important to evaluate the effectiveness of the proposed algorithm.

4.3. Subjective evaluation

The speech used for the subjective evaluation was 40 words spoken by two men and two women, and the sampling rate was 48 kHz/16 bits. The words consisted of Japanese four-mora words including consonants. A block diagram of the speech analysis/synthesis procedure is shown in Fig. 10. To enable pure evaluation of the difference in aperiodicity, Legacy-STRAIGHT was used for F0 and spectral envelope estimation. Sixteen people with normal hearing participated in the evaluation. They sat in a room with an A-weighted sound pressure level of 26 dB. We did not use a sound proof room as we wanted to evaluate the differences in a natural environment. The speech was reproduced using a laptop

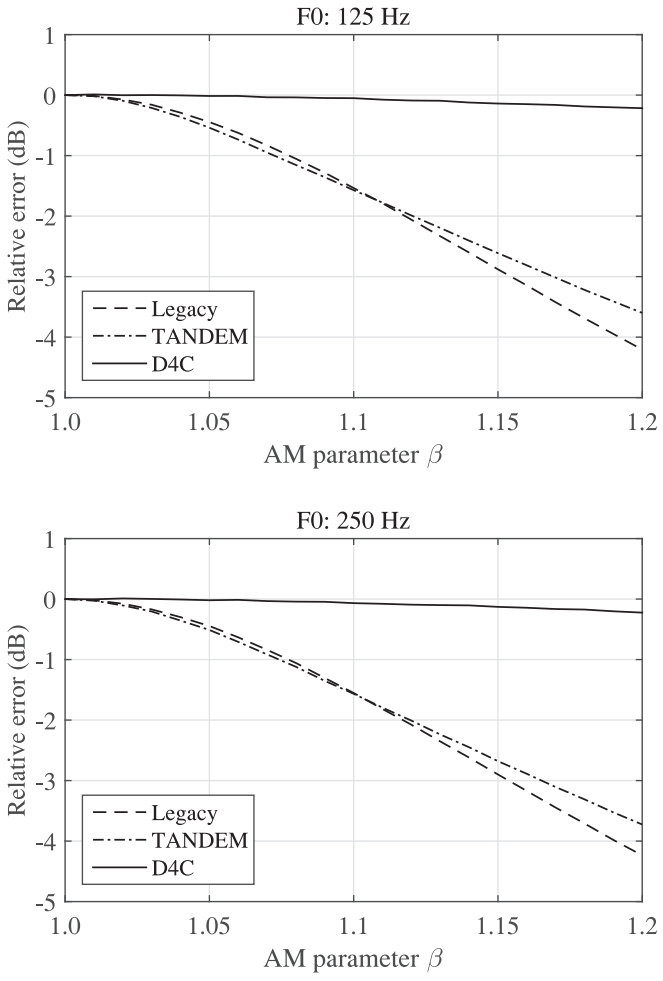

Fig. 9. Relationship between AM parameter and estimation result.

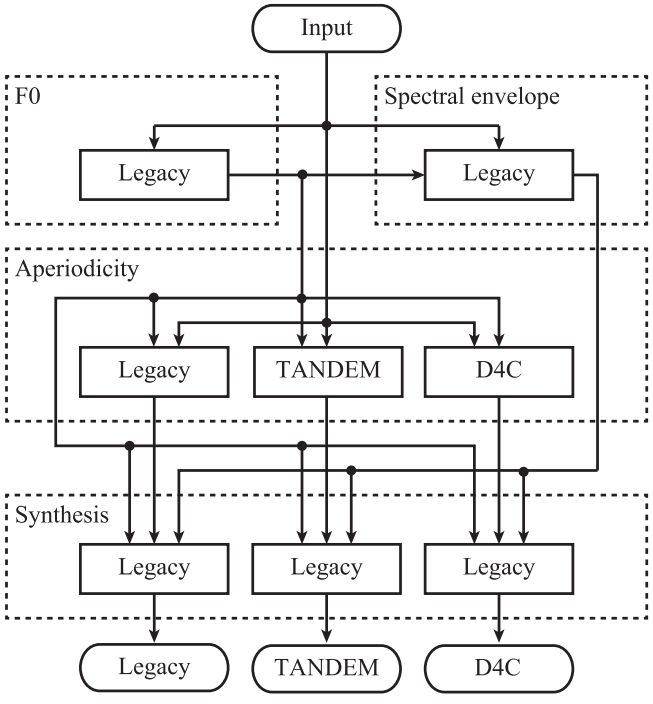

Fig. 10. Block diagram of speech analysis/synthesis procedure. F0 and spectral envelope were from results estimated by Legacy-STRAIGHT. Algorithm based on Legacy-STRAIGHT was used for speech synthesis.

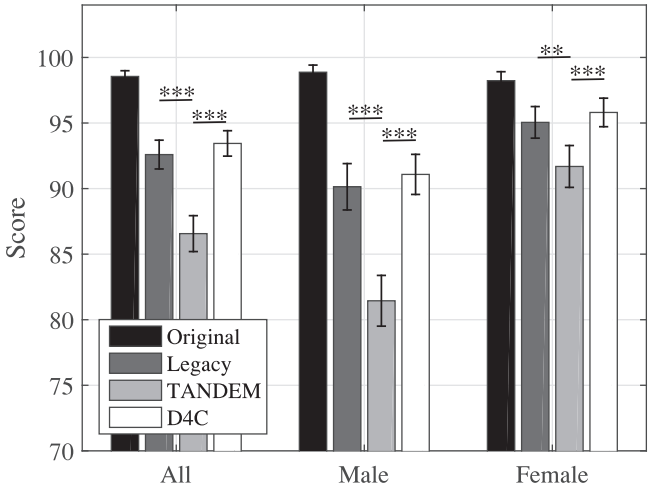

Fig. 11. Results of MUSHRA subjective evaluation. Symbols ** and *** represent significant differences ($p < 0.01$) and ($p < 0.0001$), respectively. Error bars represent 95% confidence intervals. There were significant differences between original and synthesized speech for all algorithms ($p < 0.0001$).

PC with an audio interface (EDIROL Quad-Capture), and the participants listened to the speech through headphones (SENNHEISER HD650).

A MUSHRA evaluation based on ITU-R recommendation BS.1534-1 was carried out to compare the sound quality of the original and re-synthesized speech. The participants rated the speech on a scale of 0 to 100 using an interface that simultaneously displayed four kinds of stimuli (the original speech and speech synthesized using Legacy-STRAIGHT, TANDEM-STRAIGHT, and D4C). The four sets of speech were defined as one set of stimuli, so the number of sets was 40.

The results (Fig. 11) show that not all of the algorithms could synthesize speech as natural as the input speech. TANDEM-STRAIGHT was significantly inferior to the others for both male and female speech. Under all conditions, Legacy-STRAIGHT and D4C did not significantly differ. The results for female speech were higher than those for male speech for all the algorithms.

4.4. Discussion for subjective evaluation

The results of the subjective evaluation showed that D4C can synthesize natural speech better than TANDEM-STRAIGHT. Legacy-STRAIGHT requires post-processing for temporal smoothing whereas D4C can estimate aperiodicity without post-processing. TANDEM-STRAIGHT has the same advantage, but it had sound quality inferior to that of D4C.

Band aperiodicity based on an auditory scale (e.g., mel and Bark scales) is generally used, and a previous study (Lin et al., 2000) used different center frequencies. It is interesting that D4C had the highest sound quality even though a linear scale was used. This suggests that only one estimated aperiodicity (3 kHz) is enough to synthesize natural speech. In conclusion, all the results indicate that D4C is suitable for high-quality speech synthesis.

5. Conclusion

Our proposed D4C band-aperiodicity estimator for high-quality speech synthesis uses a temporally static parameter calculated on the basis of group delay and does not require post-processing. The results of objective evaluation show that D4C can effectively estimate band aperiodicity. In particular, it is highly robust against F0 estimation error and amplitude modulation of vocal cord vibration. The results of subjective evaluation show that D4C is superior

to two conventional algorithms. Although the difference between D4C and Legacy-STRAIGHT was small, D4C can estimate aperiodicity without post-processing. While this evaluation demonstrated the effectiveness of D4C, its parameters were not optimized. Optimizing them remains for future work. Future work also includes applying D4C to voice morphing, voice conversion (Ohtani et al., 2006), and statistical parametric speech synthesis (Koriyama et al., 2014; Zen et al., 2009).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP15H02726, JP16H05899, and JP16H01734.

References

Atal, B.S., Hanauer, S.L., 1971. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. 50 (2B), 637–655.
Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., Kawahara, H., 2007. Implementation of realtime STRAIGHT speech manipulation system. Acoust. Sci. & Tech. 28 (3), 140–146.
Black, A.W., Campbell, N., 1995. Optimising selection of units from speech databases for concatenative synthesis. In: Proceedings of EUROSPEECH95, 1, pp. 581–584.
Cohen, L., 1994. Time Frequency Analysis. Prentice Hall.
Deshmukh, O., Wilson, C.E., 2003. A measure of aperiodicity and periodicity in speech. In: Proceedings of ICASSP2003, pp. 448–451.
Dudley, H., 1939. Remaking speech. J. Acoust. Soc. Am. 11 (2), 169–177.
Griffin, D.W., Lim, J.S., 1985. A new model-based speech analysis/synthesis system. In: Proceedings of ICASSP1985, 10, pp. 513–516.
Griffin, D.W., Lim, J.S., 1988. Multiband excitation vocoder. IEEE Trans. Acoust., Speech, Signal Process. 36 (8), 1223–1235.
Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: Proceedings of MAVEBA2001, pp. 59–64.
Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction. Speech Commun. 27 (3–4), 187–207.
Kawahara, H., Morise, M., 2011. Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. SADHANA - Acad. Proc. Eng.Sci. 36 (5), 713–728.
Kawahara, H., Morise, M., 2012. Simplified aperiodicity representation for high-quality speech manipulation systems. In: Proceedings of ICSP2012, pp. 579–584.
Kawahara, H., Morise, M., Nisimura, R., Irino, T., 2012. An interference-free representation of group delay for periodic signals. In: Proceedings of APSIPA ASC 2012, pp. 1–4.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In: Proceedings of ICASSP2008, pp. 3933–3936.
Kawahara, H., Morise, M., Toda, T., Banno, H., Nisimura, R., Irino, T., 2014. Excitation source analysis for high-quality speech manipulation systems based on an interference-free representation of group delay with minimum phase response compensation. In: Proceedings of INTERSPEECH2014, pp. 2243–2247.
Kawahara, H., Nisimura, R., Irino, T., Morise, M., Takahashi, T., Banno, H., 2009. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In: Proceedings of ICASSP2009, pp. 3905–3908.
Koriyama, T., Nose, T., Kobayashi, T., 2014. Statistical parametric speech synthesis based on gaussian process regression. IEEE J. Select. Topics Signal Process. 8 (2), 173–183.
Lin, W., Koh, S.N., Lin, X., 2000. Mixed excitation linear prediction coding of wideband speech at 8 kbps. In: Proceedings of ICASSP'00, 2, pp. 1137–1140.
Mathews, M.V., Miller, J.E., David, E.E., 1961. Pitch synchronous analysis of voiced sounds. J. Acoust. Soc. Am. 33 (2), 179–186.
McCree, A.V., Barnwell, T.P., 1995. A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Trans. Speech Audio Process. 3 (4), 242–250.
Morise, M., 2015a. CheapTrick, a spectral envelope estimator for high-quality speech synthesis. Speech Commun. 67, 1–7.
Morise, M., 2015b. Error evaluation of an F0-adaptive spectral envelope estimator in robustness against the additive noise and F0 error. IEICE Trans. Inf. Syst. E98-D (7), 1405–1408.
Morise, M., Onishi, M., Kawahara, H., Katayose, H., 2009. v.morish'09: A morphing-based singing design interface for vocal melodies. Lecture Notes Comput. Sci. LNCS 5709, 185–190.
Nakano, T., Goto, M., 2012. A spectral envelope estimation method based on F0-adaptive multi-frame integration analysis. In: Proceedings of SAPA-SCALE 2012, pp. 11–16.
Nuttall, A.H., 1981. Some windows with very good sidelobe behavior. IEEE Trans. Acoust., Speech, Signal Process. 29 (1), 84–91.
Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2006. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In: Proceedings of ICSLP, pp. 2266–2269.
Oppenheim, A.V., 1969. Speech analysis-synthesis system based on homomorphic filtering. J. Acoust. Soc. Am. 45 (2), 458–465.
Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.