Sādhanā Vol. 36, Part 5, October 2011, pp. 713–727. © Indian Academy of Sciences

Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework

HIDEKI KAWAHARA[1] and MASANORI MORISE[2]

[1]Faculty of Systems Engineering, Wakayama University, Wakayama 640-8510, Japan
[2]College of Information Science and Engineering, Ritsumeikan University, Kusatsu 525-8577, Japan
e-mail: kawahara@sys.wakayama-u.ac.jp; morise@fc.ritsumei.ac.jp

Abstract. This article presents comprehensive technical information about STRAIGHT and TANDEM-STRAIGHT, a widely used speech modification tool and its successor. They share the same concept: the periodic excitation found in voiced sounds is an efficient mechanism for transmitting underlying smooth time–frequency representation. The tools are also based on the perceptual equivalence of two sets of independent Gaussian random signals. This equivalence makes it possible to discard input phase information intentionally and enables flexible manipulation of parameters.

Keywords. Speech analysis; fundamental frequency; speech synthesis; consistent sampling; periodic signals.

1. Introduction

STRAIGHT, a speech analysis, modification, and synthesis framework (Kawahara *et al* 1999a), was originally designed to promote speech perception research by providing a tool to manipulate naturally sounding speech material in terms of perceptually relevant and precisely controllable physical parameters (Kawahara 2006). The original STRAIGHT (legacy-STRAIGHT) was used for a decade and was superseded by TANDEM-STRAIGHT (Kawahara *et al* 2008), a complete reformulation and reengineering based on the same underlying concept. This article provides a comprehensive technical description of TANDEM-STRAIGHT. The following section describes the first step for solving this problem, spectral envelope estimation. Section 3 addresses its use of periodicity detection, which is crucial for implementing TANDEM-STRAIGHT. Section 4 discusses other issues of source information representations. The last section gives a summary and conclusions.

2. Power spectrum of periodic signals

Periodic signals, which are familiar auditory stimuli for humans, can convey rich and detailed information. They are usually perceived as smooth and comfortable. Voiced sounds are an example, but they are not strictly deterministic or stationary. This non-stationarity inevitably leads to

time–frequency analysis. The commonly used analysis tool in speech applications is a spectrogram, a power spectral sequence calculated by short-term Fourier analysis. However, periodic signals with many time-varying harmonic components are troublesome for short-term Fourier analysis. The output of a time invariant linear system excited by a periodic pulse train yields a spectrogram that has periodic interference both in the time and frequency domains, even if the system and the input are temporally stable and spectrally smooth. This is the major problem STRAIGHT and TANDEM-STRAIGHT were designed to address. The latest answer is called TANDEM (Morise *et al* 2007), a short-term power spectral representation of periodic signals that does not have a temporally varying component. Introduction of TANDEM completely reformulated STRAIGHT, as shown in the following section.

2.1 *Cancellation of temporal variation*

Assume a time-window function adaptively designed for a target signal with fundamental period T_0. The time window is designed to have equivalent transfer function $W(\omega)$, which covers up to two harmonic components of the target signal, and has negligible side lobes. To investigate the power spectra of the windowed periodic signals, it is general enough to assume signal $x(t)$ that consists of two sinusoidal components $\omega_0 = 2\pi/T_0$ apart:

$$x(t) = e^{jk\omega_0 t} + \alpha e^{j((k+1)\omega_0+\beta)}\,, \tag{1}$$

where α and β represent real numbers. Assuming $k = 0$ for simplicity, power spectrum $P(\omega, t)$ of the windowed signal yields the following:

$$P(\omega, t) = |W(\omega)|^2 + \alpha^2\,|W(\omega-\omega_0)|^2 + 2\alpha W(\omega)W(\omega-\omega_0)\cos(\omega_0 t+\beta)\,, \tag{2}$$

where the third term represents the temporal variation to be removed. The power spectrum calculated at $t + \frac{T_0}{2}$ has a third term with opposite polarity, suggesting that this third term is cancelled by adding $P(\omega, t)$ and $P\left(\omega, t+\frac{T_0}{2}\right)$:

$$\begin{aligned} P\left(\omega, t+\frac{T_0}{2}\right) &= |W(\omega)|^2 + \alpha^2\,|W(\omega-\omega_0)|^2 + 2\alpha W(\omega)W(\omega-\omega_0)\cos\left(\omega_0(t+\frac{T_0}{2})+\beta\right) \\ &= |W(\omega)|^2 + \alpha^2\,|W(\omega-\omega_0)|^2 - 2\alpha W(\omega)W(\omega-\omega_0)\cos\left(\omega_0 t+\beta\right). \end{aligned} \tag{3}$$

TANDEM spectrum $P_{\mathrm{T}}(\omega, t)$ is redefined based on this result with a modification to make it symmetric:

$$P_{\mathrm{T}}(\omega, t) = \frac{1}{2}\left[P\left(\omega, t-\frac{T_0}{4}\right) + P\left(\omega, t+\frac{T_0}{4}\right)\right]. \tag{4}$$

It should be noted that averaging N power spectra with temporal spacing $\frac{T_0}{N}$ also yields a temporally stable power spectrum. This is a specialized version of the Welch method (Welch 1967) for periodic signals.

2.1a *Selection of time-window function*: It is pragmatically important to decide which time-window function to use for calculating the TANDEM spectrum. The requirements for the window function described in the previous section cannot be fulfilled in a strict sense, because temporally bounded window function does not have compact support in the frequency domain. The leakage outside the effective pass band results in temporal variations of the TANDEM

spectrum made from the original window function. There is a trivial solution for suppressing temporal variations. The temporal variations of the power spectra obtained using the original window function are suppressed effectively by increasing the window length because it reduces their equivalent pass band to cover only one harmonic component. However, this trivial solution makes the logarithmic power spectra sensitive to background noise.

TANDEM is a procedure that shortens the window length while keeping the power spectra temporally constant and the logarithmic power spectra tolerant to background noise. To quantify these observations, we introduced measures for the window length, the temporal as well as the frequency variations of the power spectra, and the temporal variation of the logarithmic power spectra.

Let $w(t)$ represent a window function defined in region $-\frac{T_\mathrm{w}}{2} < t < \frac{T_\mathrm{w}}{2}$. Effective duration σ_t of window $w(t)$ is defined as the the square root of the second moment of squared window. As the window length is adaptively determined using the F0 information, the normalized version of the duration is used.

$$\sigma_t = \frac{1}{T_0}\sqrt{\frac{1}{T_\mathrm{w}}\int_{-\frac{T_\mathrm{w}}{2}}^{\frac{T_\mathrm{w}}{2}} t^2 w^2(t)dt}, \quad \text{where} \quad \frac{1}{T_\mathrm{w}}\int_{-\frac{T_\mathrm{w}}{2}}^{\frac{T_\mathrm{w}}{2}} w^2(t)dt = 1\ . \tag{5}$$

Let $P(\omega, t)$ represent the power spectrum calculated by an arbitrary window function. Normalized temporal variation η_t and frequency variation η_ω are defined as follows:

$$\eta_t = \frac{\sqrt{\int_{-\infty}^{\infty}\int_0^{T_0}\left|P(\omega,t)-\overline{P(\omega)}\right|^2 dt\, d\omega}}{\int_{-\infty}^{\infty}\int_0^{T_0} P(\omega,t)\, dt\, d\omega}, \quad \eta_\omega = \frac{\sqrt{\int_0^{T_0}\int_{-\infty}^{\infty}\left|P(\omega,t)-\overline{P(t)}\right|^2 d\omega\, dt}}{\int_{-\infty}^{\infty}\int_0^{T_0} P(\omega,t)\, dt\, d\omega}, \tag{6}$$

$$\text{where} \quad \overline{P(\omega)} = \frac{1}{T_0}\int_0^{T_0} P(\omega,t)\, dt\ , \quad \overline{P(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega,t)\, d\omega\ .$$

Figure 1 summarizes the exemplar test results using a discrete test signal. The figure shows normalized temporal and frequency variations of the original window functions and their TANDEM versions. The original window functions used are Hanning, Blackman, Nuttall, and Kaiser ($\beta = 9$) windows (Harris 1978; Nuttall 1981). The test signal is a periodic pulse train with a fundamental period of 400 samples. Fast Fourier Transform (FFT) buffer length L is set to $L = 2^{\lfloor \log_2(4L_W)\rfloor}$, where L_W represents the original window function length in samples. The horizontal axis shows the normalized effective window length that is calculated by Eq. 5.

It should be noted that the temporal variations of the TANDEM windows reach stable levels with shorter (about 60% of the original) effective window lengths. The frequency variations of the TANDEM windows at the beginning of the stable points are about 5 dB smaller than the original ones.

This difference significantly affects the logarithmic power spectra when background noise exists. The temporal variations of the logarithmic power spectra represented in terms of dB η_{dBt} are defined below:

$$\eta_{dBt} = \sqrt{\left\langle \frac{1}{2\pi T_0}\int_{-\infty}^{\infty}\int_0^{T_0}\left|L(\omega,t)-\overline{L(\omega)}\right|^2 dt\, d\omega\right\rangle} \tag{7}$$

$$\text{where} \quad \overline{L(\omega)} = \left\langle \frac{1}{T_0}\int_0^{T_0} L(\omega,t)\, dt\right\rangle\ , \quad L(\omega,t) = 10\log_{10} P(\omega,t),$$

where $\langle X\rangle$ represents the ensemble average of the random variable X.

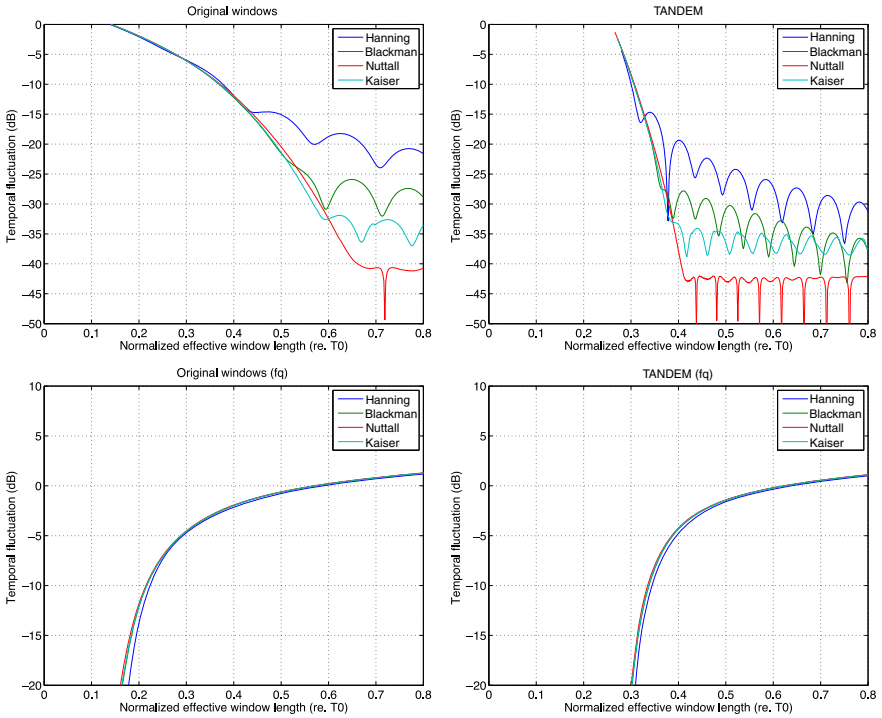

Figure 1. Normalized variations of power spectra for selected window functions: (top left) temporal variations of original time windows; (top right) temporal variations of TANDEM windows; (bottom left) frequency variations of original time windows; (bottom right) frequency variations of TANDEM windows.

Figure 2 shows the temporal variations of the logarithmic power spectra for the original and TANDEM windows under 60, 40, and 20 dB S/N conditions. The background noise is Gaussian white noise. It should be noted that the temporal variations of the TANDEM windows are around 0.5 dB even under 20 dB S/N, and those for the original windows are around 2 dB. The effective window length for the smallest temporal variations stay around $0.4T_0$ for the TANDEM windows under different S/N conditions. The temporal variations are virtually independent of the window functions in the 20 dB S/N condition because the side lobes are masked. These suggest that windows other than Hanning are relevant for applying TANDEM.

Figure 3 shows the actual window lengths of the original and their corresponding TANDEM windows. The vertical axis represents the normalized version of window length L_W (normalized by the fundamental period). A Blackman window with F0 (T_0) adaptive length $2.5T_0$ is used in the TANDEM-STRAIGHT implementation, because it is the shortest window with relevant behaviour when the effective window length is fixed. This $2.5T_0$ Blackman window is special. No power leakage occurs at the harmonic frequencies from other harmonic components, because zeros of the frequency representation of the $2.5T_0$ Blackman window coincide with other harmonic frequencies.

2.2 *Spectral envelope recovery*

The next step is to remove the spectral variations due to periodicity. It is worthwhile to revisit the signal periodicity role here and interpret it in terms of the analogue to discrete conversion

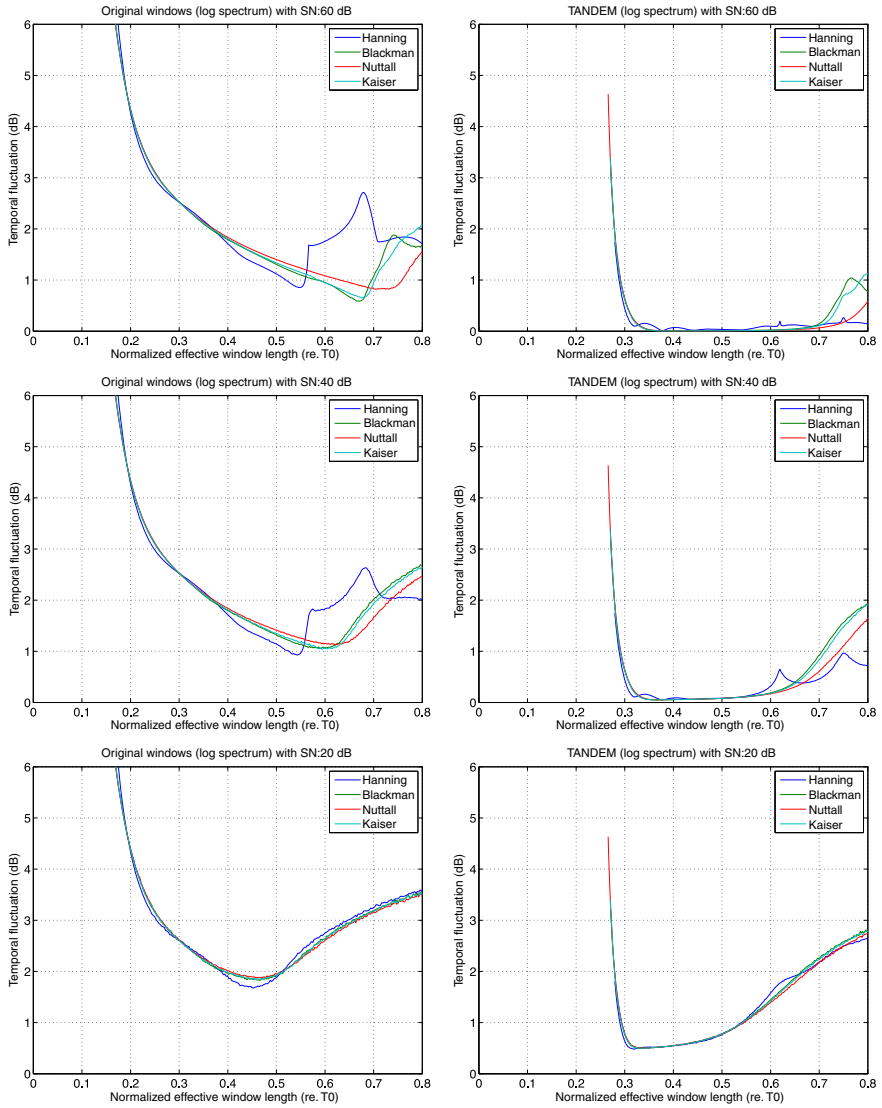

Figure 2. Temporal variation of logarithmic power spectra under different S/N: (left) original time windows; (right) TANDEM windows; S/Ns are 60, 40, and 20 dB SN from top to bottom.

problem. A new formulation of sampling theory, called consistent sampling (Unser 2000), provides the basis for this process.

The periodic excitation of a linear time invariant system in the time domain is periodic sampling of the corresponding transfer function in the frequency domain. This is a simplified description of voiced sounds. The TANDEM spectrum is a low-pass filtered (in the frequency domain) version of this sampled spectrum, where the impulse response of this low-pass filter is the frequency domain representation of the time-window function. In other words, spectral envelope recovery is an analogue to discrete conversion followed by a discrete to analogue conversion in the frequency domain. In this interpretation, this low-pass filter is found to be a poorly

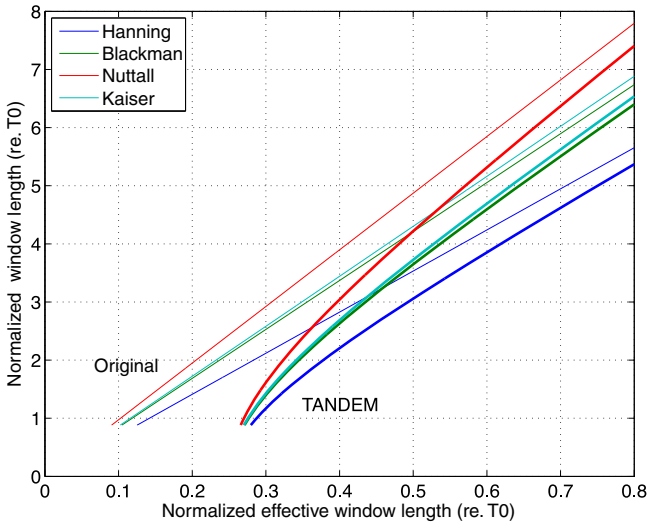

Figure 3. Normalized effective window length and actual window length.

designed anti-aliasing filter in the latter stage, because it does not have enough attenuation at the sampling frequency. Consequently, the filtered output (smoothed spectrum) still has frequency variations due to harmonic structure (in other words, periodic sampling in the frequency domain).

The former STRAIGHT uses F0 adaptive triangular smoothing function $h_1(\omega)$ as an additional anti-aliasing filter impulse response to eliminate this leakage. The base length is set to $2\omega_0$ in this case. TANDEM-STRAIGHT uses F0 adaptive rectangular function $h_2(\omega)$ instead. Its base length is set to ω_0. Smoothing function $h_1(\omega)$ is obtained by the convolution of $h_2(\omega)$ with itself. Functions $h_1(\omega)$ and $h_2(\omega)$ are also the basis functions of the cardinal B-spline family. Smoothing TANDEM spectra using this anti-aliasing smoother selectively removes spectral variations due to periodicity. However, at the same time, it smears the spectral levels at each harmonic frequency. Consistent sampling provides a solution that recovers each spectral level while suppressing the spectral variations due to periodicity.

2.2a *Spectral level recovery at harmonic frequencies*: Figure 4 shows a schematic diagram of the spectral envelope recovery process. The 'original spectral envelope' box corresponds to a hypothetical smooth spectral envelope behind the observed voiced speech, and 'recovered spectral envelope' box represents the desired goal of this process. The 'sampling by periodicity' box

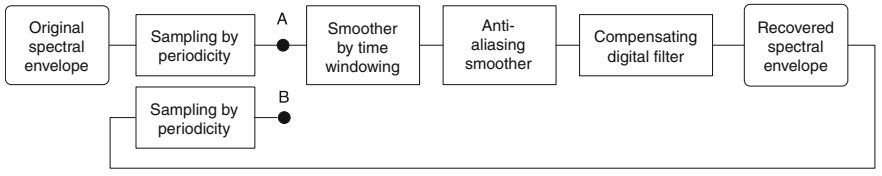

Figure 4. Spectral envelope recovery by consistent sampling.

represents the equivalent spectral sampling due to the periodic excitation of voiced sounds. The output of box 'smoother by time windowing' is the TANDEM spectrum. The 'anti-aliasing smoother' uses $h_2(\omega)$ in TANDEM-STRAIGHT. The output of this anti-aliasing smoother is the smeared version of the desired spectral envelope.

The spectral levels of this smeared spectrum are compensated at the harmonic frequencies to recover their original levels. The 'compensating digital filter' box is designed for this recovery. The sampling interval of this digital filter is ω_0. Consistent sampling provides a procedure to design this compensating digital filter. Instead of requiring a complete recovery of the original spectrum, consistent sampling only requires the resampled values to be recovered. The values at A and B in the figure must be identical. The procedure for designing the digital filter to fulfill this requirement is given below.

Recovered spectral envelope $P_{\mathrm{ST}}(\omega, t)$ is calculated using compensation digital filter coefficients q_k and the anti-aliasing smoother by the following equation:

$$P_{\mathrm{ST}}(\omega, t) = \sum_{k=-\infty}^{\infty} q_k P_{\mathrm{S}}(\omega - k\omega_0, t) \,, \tag{8}$$

$$\text{where} \quad P_{\mathrm{S}}(\omega, t) = \int_{-\infty}^{\infty} h(\lambda) P_{\mathrm{T}}(\omega - \lambda, t) d\lambda \,. \tag{9}$$

Using anti-aliasing smoother $h(\omega)$ and equivalent spectral smoother $W(\omega)$, which is the frequency domain representation of the time window function, the z-transform of compensating digital filter $Q(z)$ is calculated by the following equation.

$$Q(z) = \frac{1}{R(z)} = \frac{1}{\sum_{k=-\infty}^{\infty} r_k z^{-k}} = \sum_{k=-\infty}^{\infty} q_k z^{-k} \,, \tag{10}$$

$$\text{where} \quad r_k = \int_{-\infty}^{\infty} h(\omega - k\omega_0) \, |W(-\omega)|^2 \, d\omega \,.$$

It should be noted that the time-window function used for the TANDEM spectrum calculation is designed to cover only two harmonic components; only three of the coefficients r_k are different from zero ideally. In other words, $R(z)$ has three terms. Its reciprocal $Q(z)$ has an infinite number of terms. However, the absolute value of the k-th term vanishes very rapidly, and only some coefficients q_k are significantly different from zero. Figure 5 illustrates such behaviour.

Figure 5 shows the correlation coefficients (left plots) and digital compensation filter's coefficients (right plots) for a Blackman window with TANDEM. The horizontal axis represents the normalized effective window length in terms of T_0. The top two plots are meant for anti-aliasing smoother $h_1(\omega)$ and the bottom two plots are meant for anti-aliasing smoother $h_2(\omega)$.

2.2b *Implementation by cepstrum liftering*: Figure 6 suggests a problem in implementing this procedure: violation of positivity. The figure shows smoothed and recovered version of a line spectrum. The recovered spectra (for $h_1(\omega)$ and $h_2(\omega)$) have zeros at harmonic frequencies, indicating that compensation is successfully implemented. However, between the harmonic frequencies, the recovered spectra have negative values. The recovered spectra are not always positive semidefinite and cannot be proper power spectra.

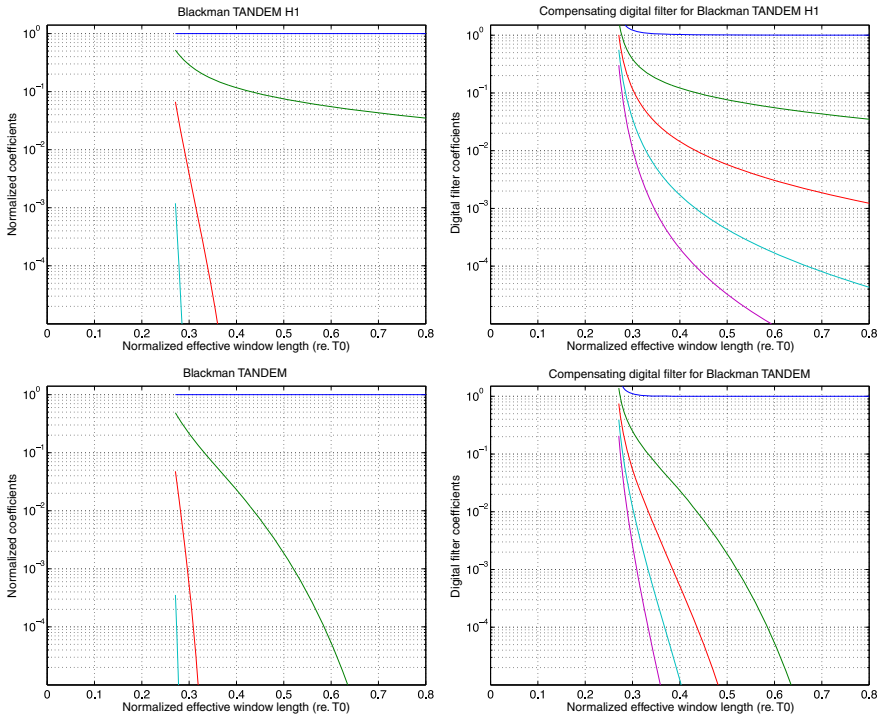

Figure 5. Correlation coefficients (left plots) for Blackman window and smoothers (top plots) $h_1(\omega)$ and (bottom plots) $h_2(\omega)$ and compensating digital filter coefficients (right plots). Absolute values are used to display the compensating coefficients. The normalized effective length of $2.5T_0$ Blackman window is 0.388.

The recovered spectra are made positive definite when digital compensation filtering is applied on the logarithmic power spectra and converted back to the power spectrum using an exponential function. Considering that logarithmic conversion has a unit bias, $\log(1 + x)$, is closely approximated by x when $|x| \ll 1$, the logarithmic conversion of the smoothed TANDEM spectra is

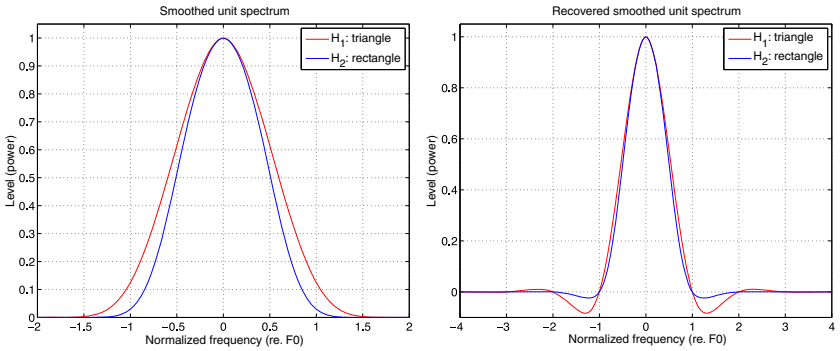

Figure 6. Smoothed line spectrum and recovered smoothed line spectrum using Blackman window and smoothers $h_1(\omega)$, $h_2(\omega)$.

closely approximated by the smoothed result of the logarithmic conversion of the TANDEM spectra. This approximation of Eq. 8 is given below.

$$P_{\mathrm{ST}}(\omega, t) \approx \exp\left(\sum_{k=-\infty}^{\infty} q_k \log(P_{\mathrm{S}}(\omega - k\omega_0, t))\right). \tag{11}$$

Figure 7 illustrates examples of this approximation. The green line in each plot represents the target model spectrum, and the blue and red lines show smoothed and recovered model spectra, respectively. It should be noted that the recovered spectra closely match the target at each harmonic frequency even though this is an approximation. This convolution in the frequency domain can also be implemented using cepstrum liftering.

$$P_{\mathrm{ST}}(\omega, t) \approx \exp\left(\mathcal{F}^{-1}\left[\left(q_0 + 2\sum_{k=1}^{\infty} q_k \cos\left(\frac{2\pi k\tau}{T_0}\right)\right) C_{\mathrm{S}}(\tau)\right]\right), \tag{12}$$

where $C_{\mathrm{S}}(\tau)$ represents the cepstrum of the smoothed power spectrum and τ represents the frequency. Symbolic notation $\mathcal{F}^{-1}[\ ]$ is also used to represent the inverse Fourier transform for simplicity.

Further approximation is introduced for calculating $C_{\mathrm{S}}(\tau)$. First, instead of calculating the convolution of a power spectrum, the convolution of a logarithmic spectrum is calculated and then Fourier transformed.

$$\begin{aligned} C_{\mathrm{S}}(\tau) &= \mathcal{F}\left[\log\left(\int_{-\infty}^{\infty} h(\lambda) P_{\mathrm{T}}(\omega - \lambda, t) d\lambda\right)\right] \\ &\approx \mathcal{F}\left[\int_{-\infty}^{\infty} h(\lambda) \log(P_{\mathrm{T}}(\omega - \lambda, t)) d\lambda\right] = g(\tau) C_{\mathrm{T}}(\tau), \end{aligned} \tag{13}$$

where $g(\tau)$ represents the Fourier transform of anti-aliasing smoothing function $h(\omega)$ and $C_{\mathrm{T}}(\tau)$ represents a cepstrum of TANDEM spectrum $P_{\mathrm{T}}(\omega, t)$. Finally, these yield the following.

$$P_{\mathrm{ST}}(\omega, t) \approx \exp\left(\mathcal{F}^{-1}\left[\left(q_0 + 2\sum_{k=1}^{\infty} q_k \cos\left(\frac{2\pi k\tau}{T_0}\right)\right) g(\tau) C_{\mathrm{T}}(\tau)\right]\right). \tag{14}$$

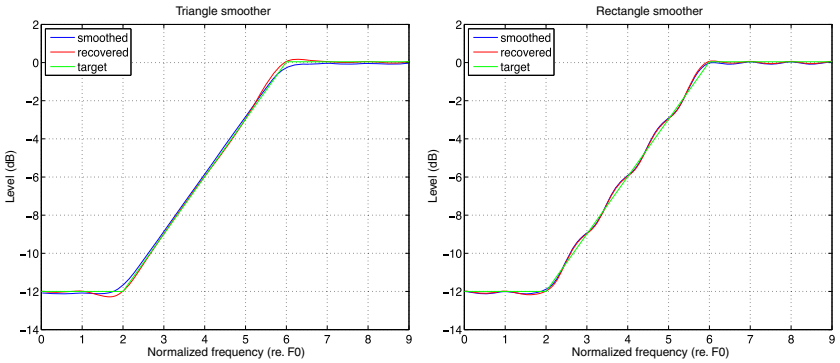

Figure 7. Smoothed model spectrum and approximately recovered smoothed model spectrum using Blackman window and smoothers (left plot) $h_1(\omega)$, (right plot) $h_2(\omega)$ with Eq. 11. Target model spectrum is shown using a green line.

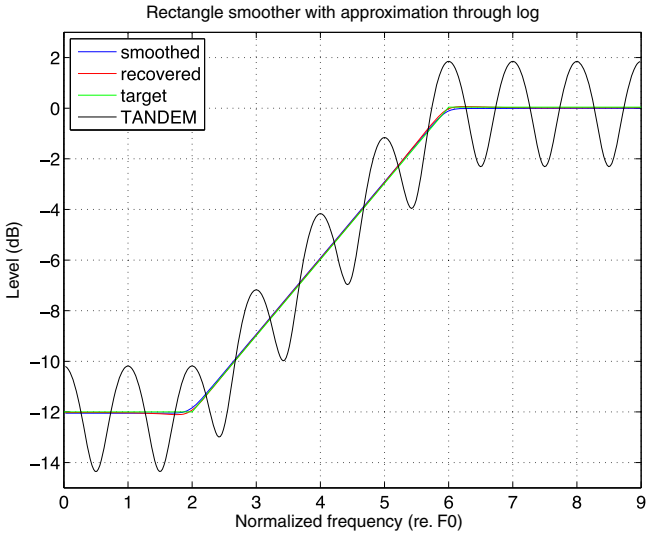

Figure 8. Smoothed model spectrum and approximately recovered smoothed model spectrum using Blackman window and smoother $h_2(\omega)$ with Eq. 14. Target model spectrum and TANDEM spectrum are shown using green and black lines.

Figure 8 shows the results using Eq. 14. Observe that the approximation yields better recovery than the original equation. However, this is not surprising because periodicity and windowing effects in the frequency domain are both periodic and multiplicative.

Table 1 shows the distances from the target spectrum in terms of the root mean squared (rms) error. The first column of the table shows the rms error result of the original TANDEM spectrum. The next four columns show the results using two smoother functions $h_1(\omega)$ and $h_2(\omega)$, and their recovered versions. The last two columns show the results using logarithmic TANDEM spectrum instead of directly using the power spectrum. It should be noted that the rectangular smoother on the logarithmic power spectra yields better approximation than all cases using direct power spectral smoothing.

Taking these into account, TANDEM-STRAIGHT spectrum $P_{\mathrm{TST}}(\omega)$ (STRAIGHT spectrum) is currently defined by the following equation.

$$P_{\mathrm{TST}}(\omega) = \exp\left(\mathcal{F}^{-1}\left[\left(\tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right)\right) g_2(\tau) C_{\mathrm{T}}(\tau)\right]\right), \tag{15}$$

$$\text{where} \quad g_2(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau} = \mathcal{F}[h_2(\omega)], \tag{16}$$

Table 1. Root mean squared dB distance from target spectrum. 'Logarithmic' indicates that smoothing and compensatory digital filtering are both applied to logarithmic TANDEM spectrum. (smth: smoothed spectrum, rcvr: recovered spectrum).

					logarithmic	
TANDEM	h_1 smth	h_2 smth	h_1 rcvr	h_2 rcvr	h_2 smth	h_2 rcvr
1.46	0.13	0.13	0.11	0.13	0.06	0.04

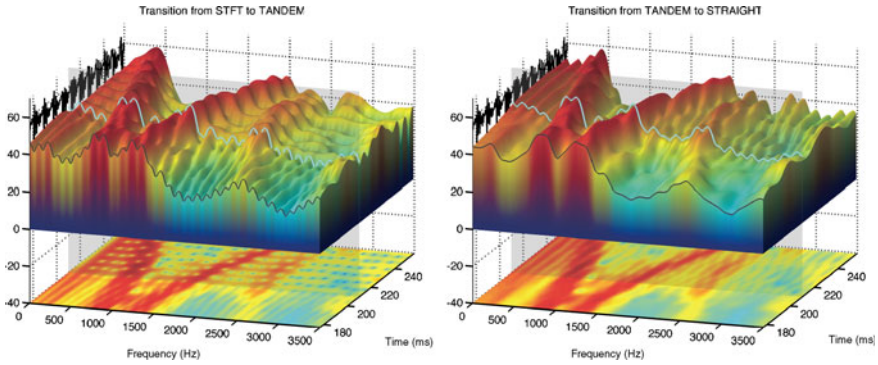

Figure 9. Real speech examples: (left) from short-term Fourier transform to TANDEM spectrogram transition and (right) from TANDEM spectrogram to STRAIGHT spectrogram transition. Transition from Japanese /a/ to /i/ spoken by a male speaker is analysed.

where $\tilde{q}_0$ and $\tilde{q}_1$ are truncated and adjusted versions of the compensating digital filter coefficients. Lifter $g_2(\tau)$ is the representation of rectangular smoother $h_2(\omega)$ in the quefrency domain. In the current implementation, $\tilde{q}_0 = 1.1$ and $\tilde{q}_1 = -0.05$ are used based on preliminary simulations.

2.2c *Natural speech examples*: Figure 9 shows three-dimensional displays of natural speech examples. The speech example is an excerpt of a Japanese vowel sequence /aiueo/ spoken by a male. The signal was sampled at 22025 Hz with 16-bit resolution. The analysis frame rate is set at 1 ms for illustration. The figure shows the transition from /a/ to /i/. Each figure shows two spectral representations transformed in the middle. The left figure shows the transition from a short-term Fourier transform (shown on the back side) to the TANDEM spectrogram (shown on the front side). The right figure shows the transition from the TANDEM spectrogram (shown on the back side) to the STRAIGHT spectrogram (shown on the front side). The grid-like structure in the short-term Fourier transform is systematically removed temporally and then spectrally.

3. Periodicity detection

The spectral envelope estimation procedure introduced in the previous section relies on F0 information, although the required precision is not very strict. Owing to this reliance, legacy-STRAIGHT and TANDEM-STRAIGHT inevitably embody F0 detection procedures. Four types of F0 extractors have been developed for this purpose. The first is based on the instantaneous frequency of the fundamental component (Kawahara *et al* 1999a). The second is based on a fixed-point calculation of the frequency to instantaneous frequency mapping using wavelet transform and an instantaneous frequency-based refinement procedure (Kawahara *et al* 1999b). The third integrates instantaneous frequency-based and autocorrelation-based methods with heavy post-processing (Kawahara *et al* 2005). The latest one, which was developed for TANDEM-STRAIGHT, is based on spectral division and an instantaneous frequency-based refinement procedure (Kawahara *et al* 2008). It also has a unique feature that enables excitation structure analysis. Owing to this feature, the latest is called the extraction structure extractor: XSX. The following sections introduces its details.

3.1 *Specialized periodicity detector*

STRAIGHT spectrum only consists of spectral envelope information. The TANDEM spectrum consists of both spectral envelope and periodicity information in a multiplicative manner. Dividing the TANDEM spectrum by the STRAIGHT spectrum yields the periodicity information and bias. By removing this bias, the following equation defines the spectral representation of periodicity information $P_{\mathrm{P}}(\omega)$.

$$P_{\mathrm{P}}(\omega) = \frac{P_{\mathrm{T}}(\omega)}{P_{\mathrm{TST}}(\omega)} - 1 \,, \tag{17}$$

where the constant 1 on the right hand side is the bias.

The next step represents the dominant periodic component as a salient peak. When the analysed signal is periodic and consists of all harmonic components, the inverse Fourier transform of $P_{\mathrm{P}}(\omega)$ has a unique peak at the fundamental frequency. The height of this peak represents the salience of the periodicity.

The actual voiced sounds are not strictly periodic. They consist of Frequency Modulation (FM) and Amplitude Modulation (AM) as well as random fluctuations. They yield modulation of the periodic variation of $P_{\mathrm{P}}(\omega)$ in the higher frequency region and result in split peaks of the inverse Fourier transform. Frequency weighting function $w_{\mathrm{B}}(\omega)$ is introduced to manage this problem by suppressing the higher harmonic components. Consequently, periodicity salience function $r_{\mathrm{A}}(\tau)$ is defined as a function of lag τ and is calculated using the following equation.

$$r_{\mathrm{A}}(\tau) = \int_{-\infty}^{\infty} w_{\mathrm{B}}(\omega) P_{\mathrm{P}}(\omega) e^{j\omega\tau} d\omega, \tag{18}$$

$$\text{where} \quad w_{\mathrm{B}}(\omega) = \begin{cases} c_{\mathrm{B}} \left(1 + \cos(\frac{\pi\omega}{N\omega_0})\right) & |\omega| \le N\omega_0 \\ 0 & |\omega| > N\omega_0 \end{cases},$$

where parameter N determines the range of harmonic components used to calculate the periodicity salience. Normalization constant c_{B} makes $\int_{-\infty}^{\infty} w_{\mathrm{B}}(\tau) d\tau = 1$. As this salience function is designed by assuming a specific fundamental frequency, it is better to explicitly represent the assumed frequency using f_c instead of f_0. Therefore, notation $r_{\mathrm{A}}(\tau;\, f_c)$ is used to represent the periodicity salience function designed using f_c.

Refined salience function $r(\tau;\, f_c)$ is defined by introducing a symmetric weighting function on the logarithmic frequency.

$$r(\tau;\, f_c) = w_{\mathrm{L}}(\tau;\, f_c) r_{\mathrm{A}}(\tau;\, f_c), \tag{19}$$

$$\text{where} \quad w_{\mathrm{L}}(\tau;\, f_c) = \begin{cases} 1 + \cos(\pi u(\tau)) & |u(\tau)| \le 1 \\ 0 & |u(\tau)| > 1 \end{cases},$$

$$u(\tau) = b_{\mathrm{w}} \log_2(\tau f_c), \tag{20}$$

where b_{w} represents a parameter that determines the sharpness of the salience function around assumed periodicity f_c.

3.2 *Integrated salience function*

The salience functions defined above have an identical shape on the logarithmic frequency (as well as the logarithmic lag) axes. By placing assumed fundamental frequency f_c evenly on the

logarithmic frequency axis, the overlap of each salience function with neighbouring functions is kept constant regardless of f_c. This makes summation of logarithmically allocated salience functions $r(\tau;\, f_c)$ yield integrated salience function $r_I(\tau)$ that covers a wide frequency range.

$$r_I(\tau) = c_0 \sum_{f_c \in F_c} r(\tau;\, f_c), \tag{21}$$

where F_c represents the set of assumed frequencies for specialized detectors. Normalization constant c_0 is defined so that the salience value for periodic pulse trains yield one. In our implementation, assumed frequency $f_c(k)$ of the k-th detector is defined below:

$$f_c(k) = f_{\mathrm{L}} 2^{\frac{k-1}{L}}, \tag{22}$$

where L represents the density of the specialized detectors in terms of the number of detectors in one octave and f_{L} represents the assumed frequency of the detector that covers the lowest end of the periodicity detection frequency range. The total number of detectors M is determined by the following equation:

$$M = \lceil L(\log_2(f_{\mathrm{U}}) - \log_2(f_{\mathrm{L}})) \rceil + 1, \tag{23}$$

where $\lceil x \rceil$ rounds x towards positive infinity and f_{U} represents the assumed frequency of the detector that covers the highest end of the periodicity detection frequency range.

3.3 *Implementation of periodicity detector*

Several factors must be considered for designing the proposed periodicity detector. The first design decision addresses the length of the time window, which was designed in two steps. The first step checks the shape of individual salience function $r_{\mathrm{A}}(\tau;\, f_c)$ to the periodic pulse input.

Figure 10 shows raw response $r_{\mathrm{A}}(\tau;\, f_c)$ of an individual detector with different base-bandwidth Ns. The time window is Blackman, and the length is set to $4T_0$. Setting the length to $4T_0$ gives $r_{\mathrm{A}}(\tau;\, f_c)$ a dominant peak at $\tau = 1/f_c$. The horizontal axis is normalized by assumed

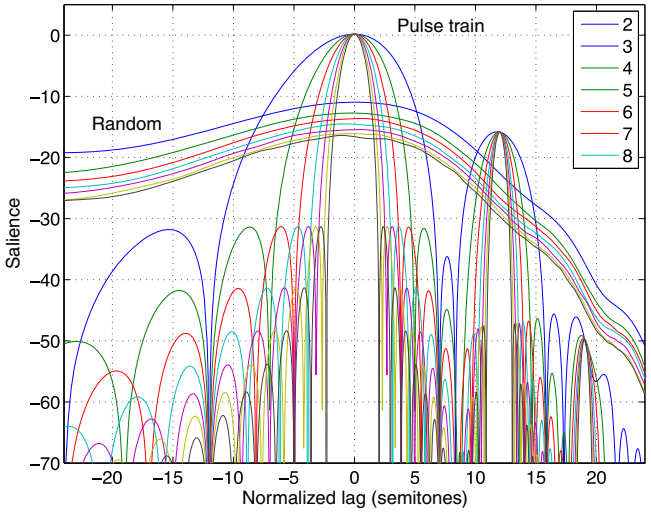

Figure 10. Response $r_{\mathrm{A}}(\tau;\, f_c)$ to a periodic pulse train and noise input.

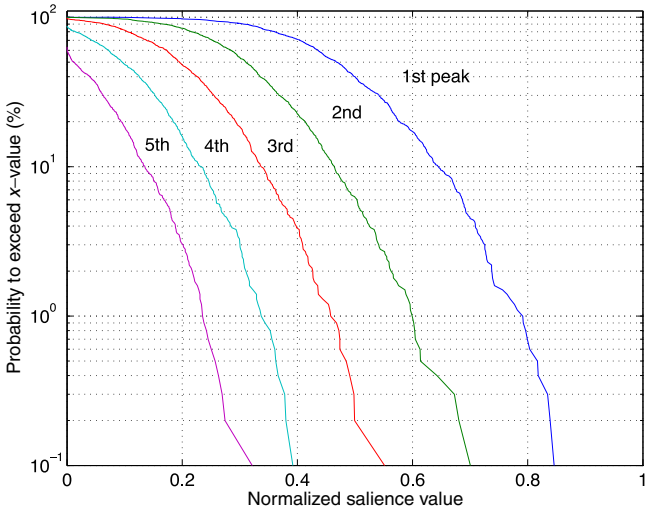

Figure 11. Probability of each salience peak to exceed values indicated by the horizontal axis.

fundamental period $T_c = 1/f_c$. Deviation is represented in terms of semitones. It should be noted that the period selectivity is sharper when a larger base-bandwidth N is used.

This response to a periodic signal should have a conspicuous distinctive value from the background peak levels due to random noise. Figure 10 also shows the averaged response to the random Gaussian noise. This averaged response is approximately flat in the vicinity of assumed period f_c. If this averaged response has a constant value, sharpness parameter b_w is uniquely determined as $b_w = 1/L$, where L represents the density of the individual detectors defined in Eq. 20. Owing to the slightly peaky shape of the averaged response, actual sharpness parameter b_w used in the implementation is slightly larger than $1/L$. For example, $b_w = 0.28$ is used instead of ideal case value $0.25 = 1/4$, $(L = 4)$ in the current implementation. Figure 11 shows the integrated salience values for a white Gaussian noise input and clearly indicates that periodic signals have a conspicuous distinctive value 1.

4. Other source-related information

Natural voiced sounds are not strictly periodic. Instantaneous frequency as well as instantaneous amplitude always fluctuate. They also consist of random components especially in the higher frequency range. Pathological voices such as diplophonia introduce irregularity in repetition and complex hierarchical vibration structures. These deviations from pure periodicity are represented in the aperiodicity information in the TANDEM-STRAIGHT implementation. The aperiodicity information is represented as a time–frequency map of the random to the deterministic power ratios. They are extracted by a pitch-range octave-band linear prediction on the original time axis and the time-warped time axis that gives the apparent fundamental frequency a predetermined constant value.

The current implementation of aperiodicity in TANDEM-STRAIGHT provides reasonably high-quality synthetic speech, but it sometimes produces artifacts that can be detected by experts in careful listening tests using headphones. Scope for improvement exists in terms of concept, algorithm and implementation. For example, the extraction and representation of the

temporal distribution of random components inside one pitch period is crucial for aperiodicity representation, especially for low-pitched male voices.

5. Conclusions

The technical details of a speech analysis, modification, and resynthesis framework called TANDEM-STRAIGHT, which is a completely reformulated version of STRAIGHT are presented here. Its conceptually simple decomposition, which separates the periodicity and response information almost perfectly, makes TANDEM-STRAIGHT a flexible tool for speech perception research. TANDEM itself, the temporally stable power spectral representation of periodic signals, is useful for other general speech processing applications, such as cepstrum-based methods, speech synthesis, enhancement and speech recognition. TANDEM-STRAIGHT keeps updating by introducing advances in signal processing theories and refinements on speech signal representations.

The authors appreciate the users who participated in the evolution of STRAIGHT and TANDEM-STRAIGHT. Without such participation, this evolution would not have been possible. The support of the following agencies was also indispensable: Advanced Telecommunication Research Institute International (ATRI), Japan Society for the Promotion of Science (JSPS), Japan Science and Technology agency (JST) and Wakayama University. Currently, this research is partly supported by Grants-in-Aid for Scientific Research (A) 19200017 and 22650042 by JSPS and the CrestMuse project by JST.

References

Harris F J 1978 On the use of windows for harmonic analysis with the discrete Fourier transform, *Proc. IEEE* 66(1): 51–83

Kawahara H, Masuda-Katsuse I, de Cheveigné A 1999a Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Commun.* 27(3–4): 187–207

Kawahara H, Katayose H, de Cheveigné A, Patterson R D 1999b Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity, In *Proc. EUROSPEECH'99, ESCA*. vol. 6, pp. 2781–2784

Kawahara H, de Cheveigné A, Banno H, Takahashi T, Irino T 2005 Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT, In *Proc. Interspeech 2005 ISCA*, pp. 537–540

Kawahara H 2006 STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds, *Acoust. Sci. Technol.* 27(5): 349–353

Kawahara H, Morise M, Takahashi T, Nisimura R, Irino T, Banno H 2008 A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, In *Proc. ICASSP 2008 IEEE*, pp. 3933–3936

Morise M, Takahashi T, Kawahara H, Irino T 2007 Power spectrum estimation method for periodic signals virtually irrespective to time window position, *Trans. IEICE* J90-D(12): 3265–3267 (in Japanese)

Nuttall A H 1981 Some windows with very good sidelobe behavior, *IEEE Trans. Audio Speech Signal Process.* 29(1): 84–91

Unser M 2000 Sampling – 50 years after Shannon, *Proc. IEEE* 88(4): 569–587

Welch P D 1967 The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.* AU-15(2): 70–73