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

Hideki Kawahara[1] and Masanori Morise[2]
[1]Faculty of Systems Engineering, Wakayama University
Wakayama, 640-8510 Japan
[2]Collage of Information Science and Engineering, Ritsumeikan University
Kusatsu, 525-8577 Japan

kawahara@sys.wakayama-u.ac.jp, morise@fc.ritsumei.ac.jp

Abstract

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

1 Introduction

A speech analysis, modification and synthesis framwork STRAIGHT (Kawahara et al. 1999a) was originally designed to promote speech perception research by providing a tool to manipulate naturally sounding speech materials in terms of perceptually relevant and precisely controllable physical parameters (Kawahara 2006). The original STRAIGHT (legacy-STRAIGHT) has been used for a decade and is superseded by TANDEM-STRAIGHT (Kawahara et al. 2008) a complete reformulation and reengineering based on the same underlying concept. This article provides a comprehensive set of technical descriptions of TANDEM-STRAIGHT framework to make them accessible. The following section introduces the first step for solving this problem, spectral envelope estimation.

2 Power spectrum of periodic signals

Periodic signals are familiar auditory stimuli for humans. They are able to convey rich and detailed information and perceived smooth and comfortable, usually. Voiced sounds are one of them but are not strictly deterministic nor stationary. This non-stationarity inevitably leads to time-frequency analysis. The commonly used one in speech applications is 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 stimulated by a periodic pulse train yields a spectrogram that has periodic interferences both in the time and in the frequency domain, 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 solve. The latest answer is called TANDEM (Morise et al. 2007), a short term power spectral representation of periodic signals that does not have temporally varying component. It enabled to reformulate STRAIGHT completely and is introduced in the following section.

2.1 Cancellation of temporal variation

Assume a time-windowing function adaptively designed for a target signal with the fundamental period T_0. The time window is designed to have its equivalent transfer function $W(\omega)$, which covers up to two harmonic components of the target signal and has negligible side lobes. Then, it is general enough to assume a 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 arbitrary real numbers. Assuming $k = 0$ for simplicity, power spectrum $P(\omega, t)$ of the windowed signal yields the following.

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

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

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

The TANDEM spectrum $P_T(\omega, t)$ is redefined based on this result with a modification to make it symmetric.

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

Note 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.

2.1.1 Selection of time windowing function

It is practically important to decide which time windowing function to be used for calculating TANDEM spectrum. Requirements for windowing functions described in the previous section cannot be fulfilled in a strict sense, because temporally bounded windowing functions does not have compact support in the frequency domain. The leakage outside of the effective pass band results into temporal variations of TANDEM spectrum that is made from the original windowing function. It also should be noted that there is a trivial solution for suppressing temporal variations. Temporal variations of the power spectra calculated by using the original windowing functions are suppressed effectively by increasing their window length for their equivalent pass band to cover only one harmonic component. However, this trivial solution makes logarithmic power spectra sensitive to background noise.

In other words, TANDEM is a procedure to shorten widow length while keeping power spectra temporally constant and logarithmic power spectra tolerant to background noise. To quantify these observations, measures for window length, temporal as well as frequency variations of power spectra and temporal variation of logarithmic power spectra are introduced.

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

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

Let $P(\omega, t)$ represent a power spectrum calculated by an arbitrary windowing function. The normalized temporal variation η_t and frequency variation η_ω are defined as follows.

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

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

Figure 1 summarizes exemplar test results using a discrete test signal. It shows normalized temporal and frequency variations of the original windowing functions and their TANDEM versions. The original windowing functions used are Hanning, Blackman,

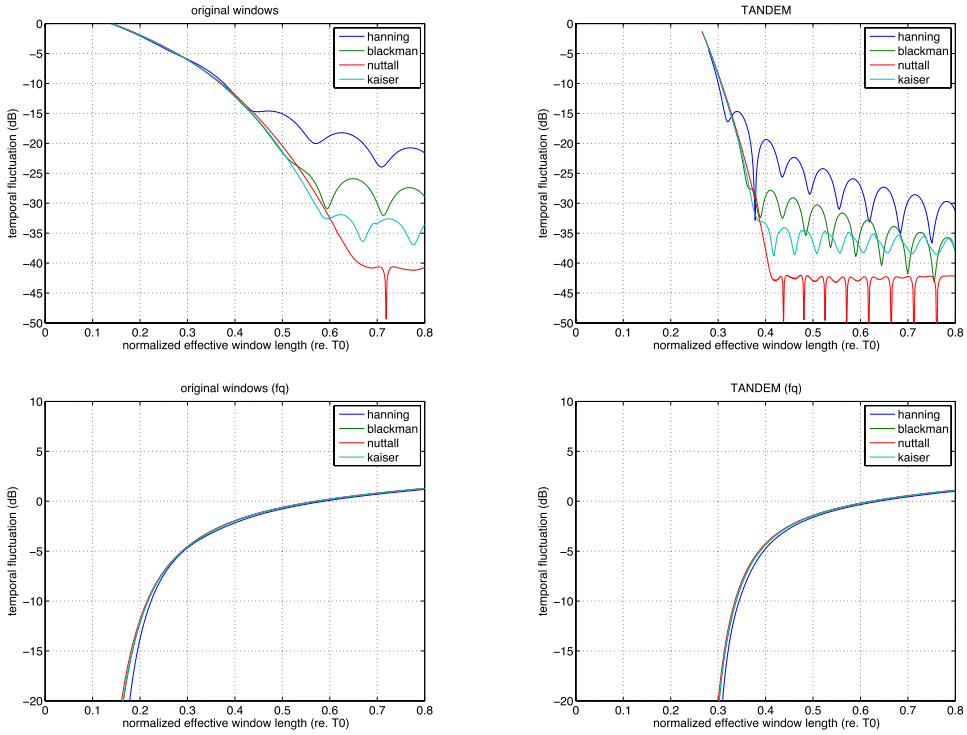

Figure 1: Normalized variations of power spectra for selected windowing functions. (top left) temporal variations of original time windows. (top right) temporal variations of TANDEM windows. (bottom left) frequency variations of original time windows. (bottom right) frequency variations of TANDEM windows.

Nuttall and Kaiser ($\beta = 9$) windows (Harris 1978, Nuttall 1981). The test signal is a periodic pulse train with the fundamental period of 400 samples. The FFT buffer length L is set to $L = 2^{\lfloor \log_2(4L_W) \rfloor}$, where L_W represents the original windowing function length in samples. The horizontal axis shows the normalized effective widow length that is calculated by Eq. 5

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

This difference has significant effects on logarithmic power spectra when background noise exists. Temporal variations of logarithmic power spectra represented in

terms of dB η_{dBt} is defined below.

$$\eta_{dBt} = \sqrt{\left\langle \frac{1}{2\pi T_0} \int_{-\infty}^{\infty} \int_{0}^{T_0} \left| L(\omega, t) - \overline{L(\omega)} \right|^2 dt\, d\omega \right\rangle} \tag{7}$$

$$\text{where } \overline{L(\omega)} = \left\langle \frac{1}{T_0} \int_{0}^{T_0} L(\omega, t)\, dt \right\rangle , \quad L(\omega, t) = 10 \log_{10} P(\omega, t),$$

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

Figure 2 shows temporal variations of logarithmic power spectra for the original and TANDEM windows under 60 dB, 40 dB and 20 dB S/N conditions. The background noise is Gaussian white noise. Note that temporal variations of TANDEM windows are around 0.5 dB even under 20 dB S/N while those for the original windows stay around 2 dB. Note also that the effective window length for the smallest temporal variations stay around 0.4 for TANDEM windows under different S/N consitions. Temporal variations are virtually independent of the windowing functions in 20 dB S/N condition because 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 normalized version of the window length L_W (normalized by the fundamental period). Blackman window with F0 (T_0) adaptive length $2.5T_0$ is used in TANDEM-STRAIGHT implementation, because it is the shortest window with relevant behavior when the effective window length is fixed. It is interesting to note that this $2.5T_0$ Blackman window is special. There exists no power leakage at 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

Next step is to remove spectral variations due to periodicity. It is worthwhile to revisit the role of signal periodicity here and interpret it in terms of analogue to discrete conversion problem. A new formulation of sampling theory, called consistent sampling (Unser 2000), provides a basis for this process.

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. 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 windowing function. In other words, spectral envelope recovery is an analogue to discrete conversion followed by a discrete to analog conversion in the frequency domain. In this interpretation, this low-pass filter found to be a poorly designed anti-aliasing filter in the latter stage, because it does not have enough attenuation at the sampling frequency. Consequently,

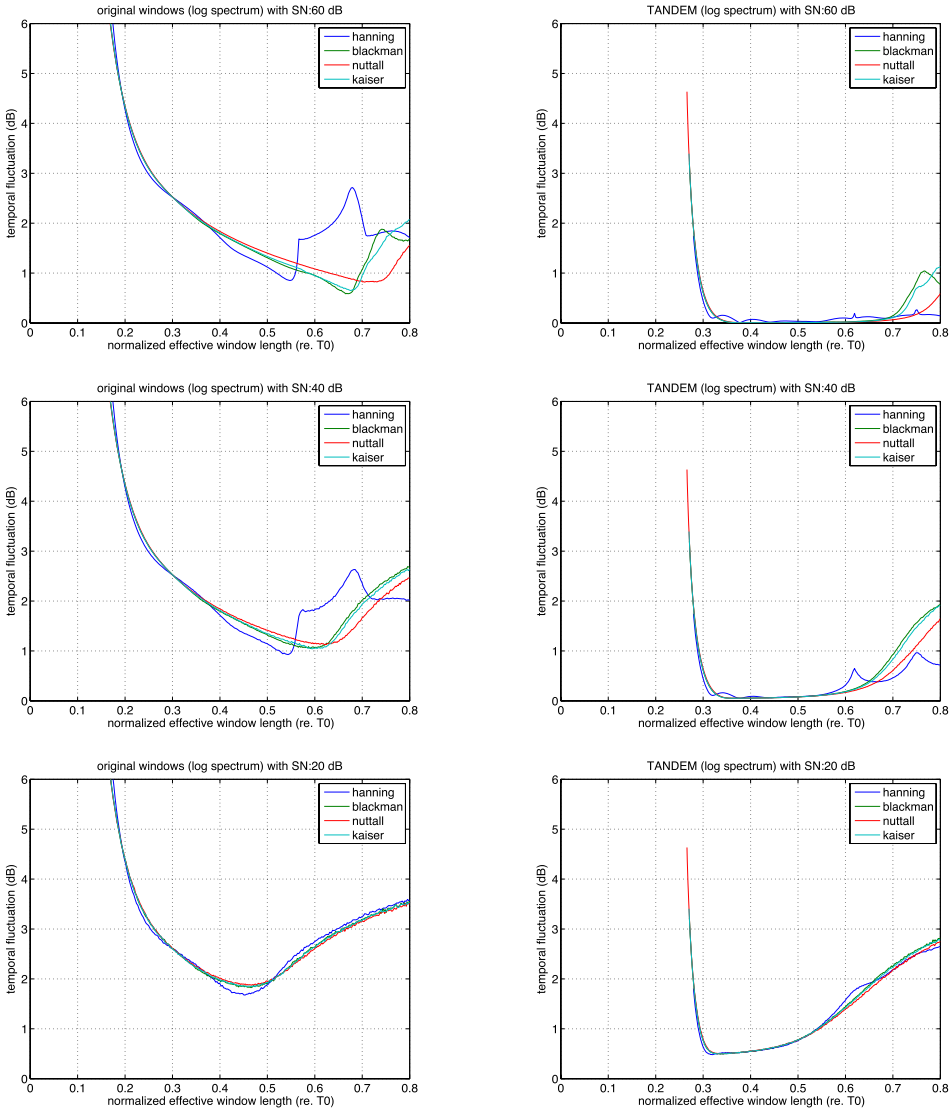

Figure 2: Temporal variation of logarithmic power spectra under different S/N. (left) original time windows. (right) TANDEM windows. The S/Ns are 60 dB, 40 dB and 20 dB SN from top to bottom

the filtered output (smoothed spectrum) still has frequency variations due to harmonic structure (in other word, sampling pulse).

Legacy-STRAIGHT uses an F0 adaptive triangular smoothing function $h_1(\omega)$ as an additional anti-aliasing filter impulse response to eliminate this leakage. The base length

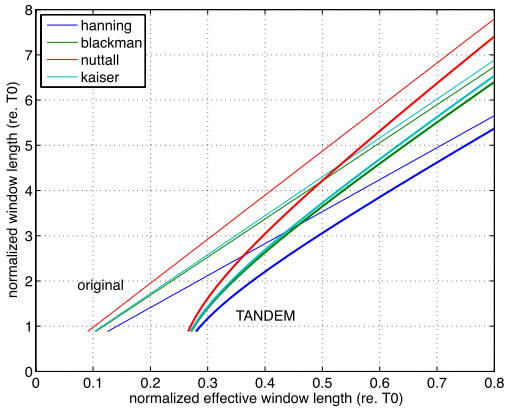

Figure 3: Normalized effective window length and actual window length

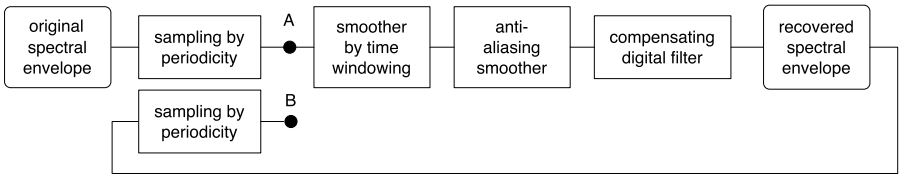

Figure 4: Spectral envelope recovery by consistent sampling

is set to $2\omega_0$ in this case. TANDEM-STRAIGHT uses an F0 adaptive rectangular function $h_2(\omega)$ instead. The base length is set to ω_0. The smoothing function $h_1(\omega)$ is yielded by convolution of $h_2(\omega)$ with itself. They are also basis functions of 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 spectral levels at each harmonic frequency. Consistent sampling provides a solution to recover each spectral level while suppressing spectral variations due to periodicity.

2.2.1 Spectral level recovery at harmonic frequencies

Figure 4 shows a schematic diagram of the spectral envelope recovery process. The box titled "original spectral envelope" corresponds to a hypothetical smooth spectral envelope behind the observed voiced speech and the box titled "recovered spectral envelope" represents the desired goal of this process. The box "sampling by periodicity" represents the equivalent spectral sampling due to periodic excitation of voiced sounds. The output of the box "smoother by time windowing" is 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.

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

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

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

$$\text{where} \quad P_S(\omega, t) = \int_{-\infty}^{\infty} h(\lambda) P_T(\omega - \lambda, t) d\lambda \tag{9}$$

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

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

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

Please note that the time windowing function used for TANDEM spectrum calculation is designed to cover only two harmonic components, only three of coefficients r_k are different from zero ideally. In other words, $R(z)$ has three terms. Consequently, its reciprocal $Q(z)$ has infinite number of terms. However, absolute value of k-th term vanishes very rapidly and also a few coefficients q_k are significantly different from zero. Figure 5 illustrates such behavior.

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

2.2.2 Implementation by cepstrum liftering

Figure 6 suggests a problem in implementing this procedure. It 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 compensation is successfully implemented.

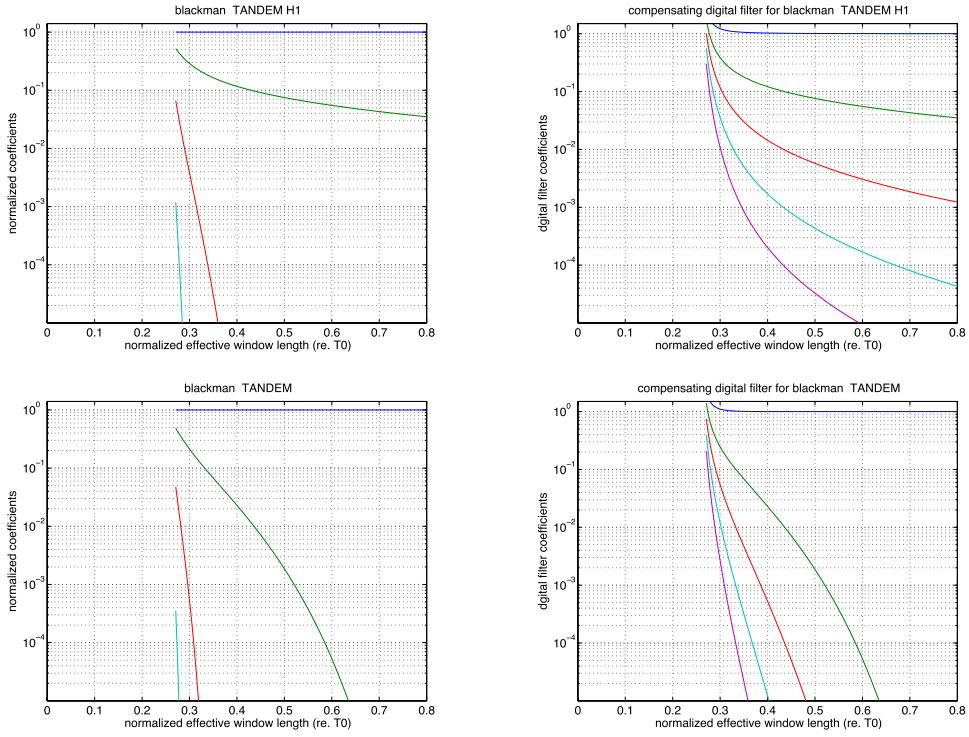

Figure 5: Correlation coefficients (left plots) for Blackman window and smoothers (top plots) $h_1(\omega)$ and (bottom plots) $h_2(\omega)$ and compensating digital filter coefficients (right plots). Absolute values are used to display the compensating coefficients. The normalized effective length of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

However, between harmonic frequencies, recovered spectra have negative values. The recovered spectra are not positive semidefinite and cannot be proper power spectra.

The recovered spectra are made positive definite when the digital compensation filtering is applied on the logarithmic power spectra and converted back to power spectrum using exponential function. Taking into account of the fact that the logarithmic conversion with a unit bias, $\log(1 + x)$, is closely approximated by x when $|x| \ll 1$, this approximation is relevant for smoothed TANDEM spectra in usual situations. This approximation of Eq. 8 is given below.

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

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 the smoothed model spectrum and the recovered model spectrum, respectively. Note that recovered spectra

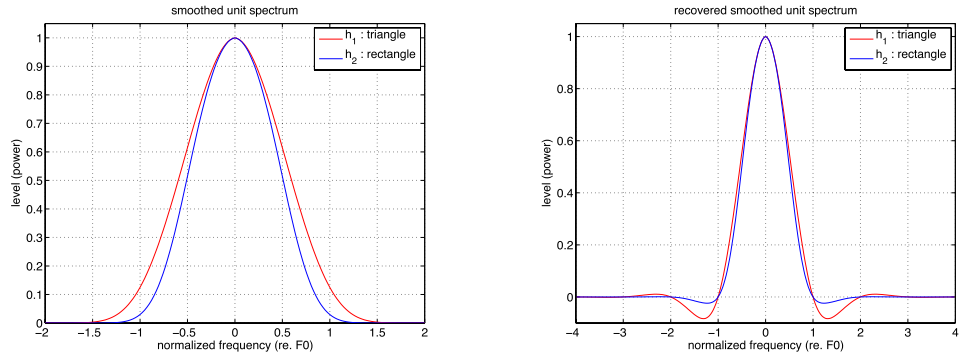

Figure 6: Smoothed line spectrum and recovered smoothed line spectrum using Blackman window and smoothers $h_1(\omega)$, $h_2(\omega)$

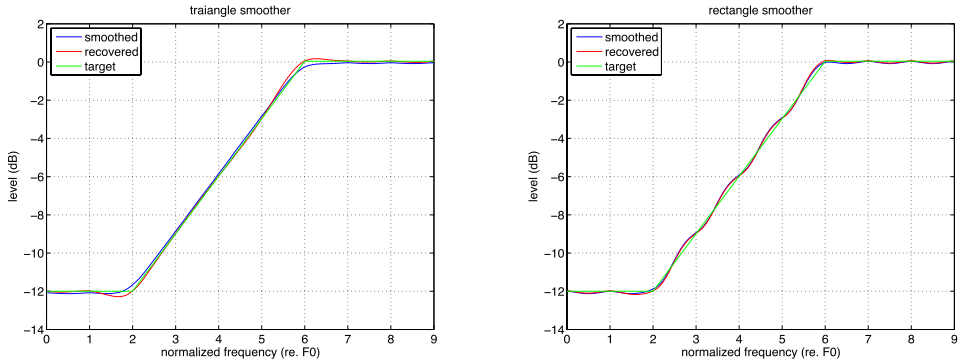

Figure 7: Smoothed model spectrum and approximately recovered smoothed model spectrum using Blackman window and smoothers (left plot) $h_1(\omega)$, (right plot) $h_2(\omega)$ with Eq. 11. The target model spectrum is shown using a green line.

closely match the target at each harmonic frequency even though this is an approximation. It also should be noted that the convolution in the frequency domain can be implemented using cepstrum liftering.

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

where $C_S(\tau)$ represents cepstrum of the smoothed power spectrum and τ represents quefrency. Also symbolic notation $\mathcal{F}^{-1}[\ ]$ is used to represent inverse Fourier transform for simplicity.

Further approximation is introduced for calculating $C_S(\tau)$. First, instead of calculat-

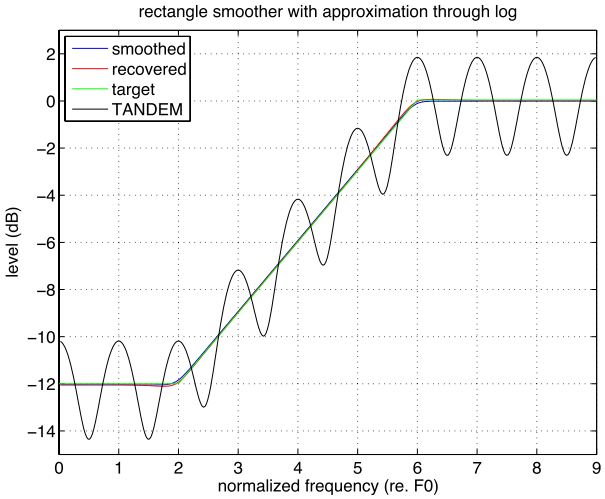

Figure 8: Smoothed model spectrum and approximately recovered smoothed model spectrum using Blackman window and smoother $h_2(\omega)$ with Eq. 14. The target model spectrum and TANDEM spectrum are shown using a green line and a black line, respectively

ing convolution of a power spectrum, convolution of a logarithmic spectrum is calculated and then Fourier transformed.

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

where $g(\tau)$ represents Fourier transform of the anti-aliasing smoothing function $h(\omega)$ and $C_T(\tau)$ represents cepstrum of the TANDEM spectrum $P_T(\omega,t)$. Finally, these yields the following.

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

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

Table 1 shows dB distances from the target spectrum in terms of root mean squared error. The original TANDEM spectrum, two smoothed spectra using $h_1(\omega)$ and $h_2(\omega)$,

					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

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

their recovered spectra are listed. The last two columns show results using logarithmic TANDEM spectrum instead of directly using power spectrum. It should be noted that rectangular smoother on logarithmic power spectra yields better approximation than all cases using direct power spectral smoothing.

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

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

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

where $\tilde{q}_0$ and $\tilde{q}_1$ are truncated and adjusted version of the compensating digital filter coefficients. The 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 a preliminary simulations.

3 Periodicity detection

The spectral envelope estimation procedure introduced in the previous section relies on F0 information, although required precision is not very strict. Due 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 extractor is based on instantaneous frequency of the fundamental component (Kawahara et al. 1999a). The second extractor is based on a fixed-point calculation of frequency to instantaneous frequency mapping using wavelet transform and instantaneous frequency-based refinement procedure (Kawahara et al. 1999b). The third extractor integrates instantaneous frequency based method and autocorrelation based method with heavy post processing (Kawahara et al. 2005). The latest one, which is developed for TANDEM-STRAIGHT and is based on spectral division and instantaneous frequency-based refinement procedure (Kawahara et al. 2008). The latest one also has a unique feature which enables excitation structure analysis.

3.1 Specialized periodicity detector

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

$$P_P(\omega) = \frac{P_T(\omega)}{P_{TST}(\omega)} - 1 \ , \tag{17}$$

where the second term 1 is the approximated bias.

The next step is to represent the dominant periodic component as a salient peak. When the analyzed signal is periodic and consists of all harmonic components, the inverse Fourier transform of $P_P(\omega)$ has a unique peak at the fundamental frequency. Hight of this peak represents salience of the periodicity.

However, actual voiced sounds are not strictly periodic. They consists of FM and AM as well as random fluctuations. These yield into modulation of periodic variation of $P_P(\omega)$ in the higher frequency region and result into split peaks of the inverse Fourier transform. A frequency weighting function $w_B(\omega)$ is introduced to manage this problem by suppressing higher harmonic components. Consequently, the periodicity salience function $r_A(\tau)$ is defined as a function of lag τ and calculated using the following equation.

$$\begin{aligned} r_A(\tau) &= \int_{-\infty}^{\infty} w_B(\omega) P_P(\omega) e^{j\omega\tau} d\omega, \\ \text{where} \quad w_B(\omega) &= \begin{cases} c_B\left(1 + \cos(\frac{\pi\omega}{N\omega_0})\right) & |\omega| \le N\omega_0 \\ 0 & |\omega| > N\omega_0 \end{cases} \ , \end{aligned} \tag{18}$$

where parameter N determines range of harmonic components used to calculate periodicity salience. The normalization constant c_B makes $\int_{-\infty}^{\infty} w_B(\tau) d\tau = 1$. Since, 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 $r_A(\tau; f_c)$ is used to represent the periodicity salience function designed using f_c.

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

$$\begin{aligned} r(\tau; f_c) &= w_L(\tau; f_c) r_A(\tau; f_c), \\ \text{where} \quad w_L(\tau; f_c) &= \begin{cases} 1 + \cos(\pi u(\tau)) & |u(\tau)| \le 1 \\ 0 & |u(\tau)| > 1 \end{cases} \ , \\ u(\tau) &= b_w \log_2(\tau f_c), \end{aligned} \tag{19}$$

where b_w represents a parameter that determines sharpness of the salience function around the 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, overlap of each salience function with neighboring functions is kept constant irrespective to f_c. This makes simple summation of logarithmically allocated salience functions $r(\tau; f_c)$ yield an integrated salience function $r_I(\tau)$ that covers wide frequency range.

$$r_I(\tau) = c_0 \sum_{f_c \in F_c} r(\tau; f_c), \tag{20}$$

where F_c represents the set of assumed frequencies for specialized detectors. The normalization constant c_0 is defined to make the salience value for periodic pulse train yield one. In our implementation, the assumed frequency $f_c(k)$ of the k-th detector is defined below.

$$f_c(k) = f_L 2^{\frac{k-1}{L}}, \tag{21}$$

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

$$M = \lceil L(\log_2(f_U) - \log_2(f_L)) \rceil + 1, \tag{22}$$

where $\lceil x \rceil$ rounds x toward positive infinity and f_U represents the assumed frequency of the detector which covers the highest end of the periodicity detection frequency range.

3.3 Implementation of periodicity detector

There are several factors to be taken into account for designing the proposed periodicity detector. The first designing decision is on the length of the time window. This is designed in two steps. The first step is to check the shape of the salience function $r_A(\tau; f_c)$ to periodic pulse input.

Figure 9 shows response of each individual detector. The horizontal axis is normalized by the assumed fundamental period $T_c = 1/f_c$. Deviation is represented in terms of semitone. Note that the period selectivity is sharper when larger base-band width N is used. This response to a periodic signal has to have a distinctive value that stands out from background peak levels due to random noise. Figure

4 Conclusions

Acknowledgement: The authors appreciate for users who participated in evolution of STRAIGHT and TANDEM-STRAIGHT. Without such participation, this evolution was

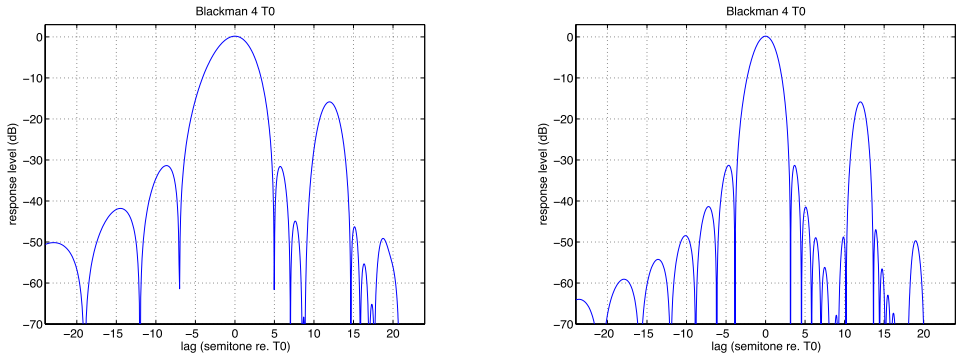

Figure 9: Response to a periodic pulse train with period T_c. The horizontal axis represents the normalized period. (Left plot: $N = 3$, Right plot: $N = 5$)

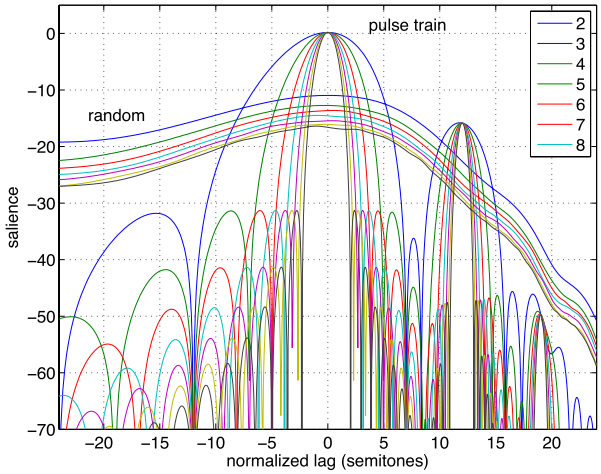

Figure 10: Response to a periodic pulse train and noise input

not possible. Various agencies' support were also indispensable. They are ATRI (Advanced Telecommunication Research Institute International), JSPS (Japan Society for the Promotion of Science), JST (Japan Science and Technology agency) and Wakayama University. Currently, this research is partly supported by Grants-in-Aid for Scientific Research (A) 19200017 by JSPS and the CrestMuse project by JST.

References

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

H. Kawahara. STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustic Science & Technology*, 27(5): 349–353, 2006.

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*, 27(3-4):187–207, 1999a.

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

Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proc. EUROSPEECH'99*, volume 6, pages 2781–2784. ESCA, 1999b.

Hideki Kawahara, Alain de Cheveigné, Hideki Banno, Toru Takahashi, and Toshio Irino. Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Proc. Interspeech2005*, pages 537–540. ISCA, 2005.

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

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

Michael Unser. Sampling–50 years after Shannon. *Proceedings of the IEEE*, 88(4): 569–587, 2000.

P. D. Welch. 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 and Electroacoustics*, AU-15(2):70–73, 1967.