Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

TANDEM-STRAIGHT, a research tool for L2 study enabling flexible manipulations of prosodic information

Hideki Kawahara

Faculty of Systems Engineering
Wakayama University, Japan
kawahara@wakayama-u.ac.jp

10.21437/SpeechProsody.2008-5

Abstract

A speech analysis, modification, and resynthesis system called STRAIGHT has been widely used in the speech research community. However, its foundation and implementation were not well established. This lecture introduces recent advances in STRAIGHT's foundation based on a new concept called TANDEM, a simple method for calculating temporally stable power spectra using two F0-adaptive time windows. A new interpretation of Shannon's sampling theory also provided a mathematical foundation and guidelines for designing STRAIGHT. A unified approach based on TANDEM and STRAIGHT in F0 extraction (at the same time, revisiting its conceptualization) and aperiodicity information representation provide flexibility in prosody-related feature manipulation. An overview on prospective applications of the new TANDEM-STRAIGHT in L2 study will also be presented.

1. Introduction

Understanding of human speech communication has been promoted by exploiting various capabilities of tools introduced to the speech research community. The first electrical speech processing system, VOCODER, established foundations of speech perception research primarily focusing on parsimonious representations of speech sounds [1]. Introduction of statistical time series analysis into speech signals [2] was the most successful successor of this parsimonious approach and formed the huge bodies of linear prediction analysis [3] technologies and applications. STRAIGHT, a speech analysis, modification, and synthesis system inherits VOCODER's basic architecture while laying this parsimony aside and focusing on conceptual isomorphism in representations with perception of periodic sounds [4, 5]. STRAIGHT decomposes input speech into three types of positive-valued parameters: an interference-free spectrogram, an aperiodicity map, and a fundamental frequency (F0) trajectory. These representations are easy to interpret in terms of source filter models and the positivity of parameters allows flexible manipulations. STRAIGHT (as well as speech morphing based on it) has been widely used in the speech research community [5, 6] because of the factors mentioned and relatively small degradations associated with manipulations. Despite the conceptual simplicity of the representations, the procedures for extracting them are complicated and the underlying principles of algorithms are not theoretically well established.

This paper revisits the underlying concepts of representations used in STRAIGHT and reformulates them [7] based on a simple new power spectrum estimation algorithm for periodic sounds that yields a temporally stable power spectral representation [8]. First, an interference-free power spectral representation is discussed and reformulated based on the newly introduced temporally stable representation and a new interpretation of Shannon's sampling theory [9, 10]. Second, discussions on excitation information that complements the proposed spectral representation are presented based on new formulations. Discussions on excitation information begin with questions on the dichotomy between periodicity and aperiodicity. Detection and representation of multiple periodicity is discussed based on a periodic component detector derived from the temporally stable power spectral representation and the interference-free power spectral representation. Following these discussions, the detector is extended to a generalized periodicity detector taking into account the interference-free spectrum and estimates of multiple periods. Finally, an acoustic event detector is introduced to complete the new formulation of STRAIGHT.

2. Temporally stable power spectrum

A simple new idea for extracting temporally stable power spectrum for periodic signals caused all procedures in STRAIGHT to be completely reformulated. The central idea behind the STRAIGHT VOCODER is to extract spectral information that does not consist of periodic structure in both the time and frequency domains. The current implementation of STRAIGHT solves this problem in two steps. The first step is to calculate temporally stable power spectrum using a set of complementary windows. The second step is to remove periodic variations in the frequency domain by using F0 adaptive smoothing and inverse filtering in the spatial frequency domain to preserve spectral levels at harmonic frequencies. Unfortunately, these conceptually straightforward ideas were implemented as Matlab procedures consisting of many tuning parameters and ad hoc functions and made the whole system difficult to understand and apply formal mathematical analysis to. The idea described below removed almost all these ad hoc and tunable components and made the new implementation of STRAIGHT transparent. F0 extraction and aperiodicity extraction procedures were also replaced by a new method based on the same spectral estimation methods.

The temporally stable power spectrum of a periodic signal is calculated as the sum of two power spectra using a pair of time windows temporally separated for half of the fundamental period [8]. Let $H(\omega)$ represent the Fourier transform of a time-windowing function. Assume that the width of the main lobe of $H(\omega)$ only covers two harmonic components of the fundamental period T_0. Therefore, it is sufficient to assume that the test signal $\delta(\omega) + \alpha e^{j\beta}\delta(\omega - \omega_0)$ represents the general periodic signals with fundamental period T_0, where $\omega_0 = 2\pi/T_0$. Since the Fourier transform of $H(\omega)$ yields $e^{-j\omega\tau}H(\omega)$ when the window is temporally displaced by the amount, τ, the power spectrum of test signal $|S(\omega,\tau)|^2$ is given by:

$$\begin{aligned} |S(\omega,\tau)|^2 &= H^2(\omega) + \alpha^2 H^2(\omega-\omega_0) \qquad (1)\\ &\quad + 2\alpha H(\omega)H(\omega-\omega_0)\cos(\omega_0\tau+\beta). \end{aligned}$$

The third term consists of window location τ and represents the temporal dependency of the power-spectrum estimation. The power spectrum of the same signal analyzed by a time window located at $\tau + T_0/2$ has a third term with an opposite sign because $\omega_0 T_0/2 = \pi$. Therefore, $|S(\omega,\tau)|^2 + |S(\omega,\tau+T_0/2)|^2$ has no time-dependent term (in this paper, the resultant spectrum is called the "TANDEM spectrum" below.)

This relation generalizes to multiple windows. By applying the identity relation $\sum_{k=0}^{N-1}\exp(jkT_0/N) = 0$, the following holds. Averaged power spectrum calculated using $N \in Z$ yields temporally stable spectrum $P_T(\omega)$ when the center locations of the time windows are separated by T_0/N.

$$\begin{aligned} P_T(\omega) &= \frac{1}{N}\sum_{k=0}^{N-1}\left|S\left(\omega,\tau+\frac{kT_0}{N}\right)\right|^2\\ &= H^2(\omega)+\alpha^2H^2(\omega-\omega_0). \qquad (2) \end{aligned}$$

In the following sections, $N = 2$ is assumed without losing generality.

Equation (1) suggests another trivial solution of the temporally stable spectrum. When a time window is long enough for $H(\omega)$ and $H(\omega-\omega_0)$ to have no overlap, the third term of Eq. (1) vanishes. However, this trivial solution is not useful for speech analysis because fine temporal resolution is necessary to track the dynamics of speech sounds due to articulatory movements. The effective duration of the TANDEM window can be made shorter than the fundamental period of the signal while retaining temporal stability.

2.1. Numerical example

Figure 1 shows the original power spectrum and corresponding TANDEM spectrum of a periodic pulse train with $T_0 = 5$ ms. The Blackmann Harris window is used in this example. The length of the window is $2.5T_0$. The FFT size is 8192, and the sampling frequency is 48 kHz. Their peak levels are normalized to 1. It is illustrated that the periodic temporal variation found in the power spectrum is effectively eliminated in the TANDEM spectrum.

3. Interference-free power spectrum

The periodic excitation of a set of resonators, such as the vocal tract, by a pulse train is also a sampling operation of the corresponding transfer function by a periodic pulse on the frequency axis. In other words, it is an analog-to-digital (discrete) conversion on the frequency axis. By this analogy, the problem becomes discrete-to-analog conversion on the frequency axis.

Because this process consists of both analog-to-discrete and discrete-to-analog conversions, and because the absolute value of the transfer function of the vocal tracts is not band-limited in terms of spatial frequency, adopting a formulation of consistent sampling [9] is better than adopting classic Shannon's sampling theory. A brief summary and excerpts of the main theorem [9, 10] are given below.

3.1. Consistent sampling (excerpts and summary)

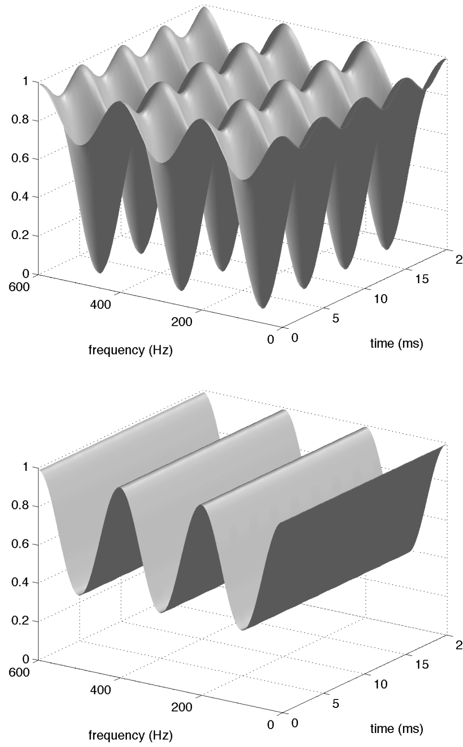


Figure 1: Original power spectrum (upper plot) and TANDEM spectrum (lower plot) of a pulse train with a 5 ms fundamental period. Blackman window with 48 kHz sampling and FFT size 8192.

Assume that a pre-filter, a sampler, a digital correcting filter, and a post-filter are connected in series. Let $\varphi_1(t)$ and $\varphi_2(t)$ represent the impulse responses of the pre- and post-filters, respectively. Then define the cross-correlation sequence $a_{12}(k)$ as an inner product of these functions:

$$a_{12}(k) = \langle \varphi_1(t-k), \varphi_2(t)\rangle. \qquad (3)$$

Theorem [9, 10] Let $f \in H$ be an unknown input function. Provided $m > 0$ exists such that $|A_{12}(e^{j\omega})| \geq m$ a.e., then there is unique signal approximation $\tilde{f}$ in $V(\varphi_2)$ that is consistent with f in the sense that

$$\forall f \in H,\ \ c_1(k) = \langle f, \varphi_1(x-k)\rangle = \langle \tilde{f}, \varphi_1(x-k)\rangle. \qquad (4)$$

This signal approximation is given by

$$\tilde{f} = \tilde{P}f(x) = \sum_{k\in Z}(c_1 * q)\varphi_2(x-k), \qquad (5)$$

where q is the impulse response of the digital correcting filter and is calculated by

$$Q(z) = \frac{1}{\sum_{k\in Z} a_{12}(k)z^{-k}}, \qquad (6)$$

and underlying operation $\tilde{P}$ is a projector from L_2 into $V(\varphi_2)$.

3.2. Envelope estimation based on consistent sampling

This theorem is applied to interference-free spectral estimation using the following interpretation of the underlying model of the theorem. Figure 2 summarizes this interpretation. In this

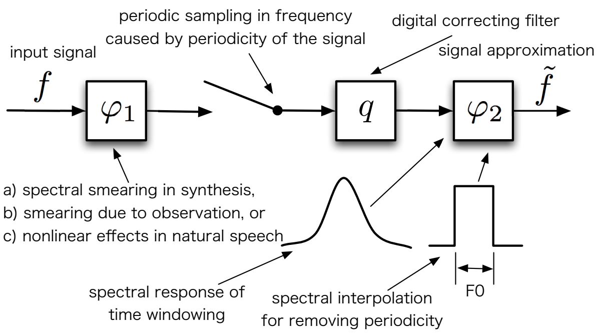

periodic sampling in frequency
caused by periodicity of the signal
digital correcting filter
signal approximation
input signal
f
φ_1
q
φ_2
$\tilde{f}$
a) spectral smearing in synthesis,
b) smearing due to observation, or
c) nonlinear effects in natural speech
spectral response of
time windowing
spectral interpolation
for removing periodicity
F0

Figure 2: Model for applying consistent sampling to spectral envelope estimation

interpretation, a TANDEM spectrum is an output approximation of this model, where the sampler is a periodic pulse on the frequency axis and the impulse response of the post-filter $\varphi_2(t)$ is $|H(\omega)|^2$ in Eq. (1). Pre- and digital correction filters and the spectral interpolation response are missing in this case. The problem is designing the missing correction filter and modifying the post-filter for consistency.

A simple illustrative case of the TANDEM method is to use the following Hanning window defined in $[-T_0, T_0]$:

$$h(t) = (1 + \cos(\pi t/T_0))/2. \tag{7}$$

The TANDEM spectrum of a periodic pulse train with a period of T_0 is also periodic on the frequency axis. This periodic fluctuation on the frequency axis represents interference caused by signal periodicity. This interference is completely eliminated by calculating the convolution with a rectangular smoothing function $r_{\omega 0}(\omega)$ when the width is set to ω_0. This smoothing function is the most localized anti-aliasing filter for discrete-to-analog conversion. Coefficients of the correction filter q_k are calculated using Eq. (6) with $(|H(\omega)|^2 * r_{\omega 0}(\omega))$ for $\varphi_2(t)$ and a delta function for $\varphi_1(t)$ to calculate a cross-correlation sequence $a_{12}(k)$. In this example, $a_{12}(k)$ consists of three non-zero elements: 0.0468, 1, and 0.0468 for $k = -1, 0, 1$. Coefficients $q_k = q_{-k}$ for $k = 0, 1, 2,$ and 3 are 1.0044, -0.0471, 0.0022, and -0.0001, respectively, and vanish rapidly for larger k.

The convolution of TANDEM spectrum $P_T(\omega)$ with $r_{\omega 0}(\omega)$ is calculated from the difference of the integrated TANDEM spectrum at two frequency points separated by ω_0. It is useful to truncate q_k in order to leave three dominant elements (for $k = -1, 0, 1$) because the large dynamic range usually found in speech spectra tends to introduce spectral smearing if q_k has long tails. Let $\tilde{q}_k$ represent the normalized and adjusted q_k to compensate for the effect of this truncation. The interference-free spectrum is assured to have no negative values when the correction filtering using $\tilde{q}_k$ is implemented in the cepstral domain. Taking into account these considerations, an interference-free spectrum, $P_{TST}(\omega)$ ("STRAIGHT spectrum" below) is calculated from the TANDEM spectrum $P_T(\omega)$ using the following set of equations:

$$\begin{aligned} C(\omega) &= \int_{\omega_L}^{\omega} P_T(\lambda) d\lambda && (8) \\ L_S(\omega) &= \ln\left[C(\omega + \omega_0/2) - C(\omega - \omega_0/2)\right] - \ln \omega_0 \\ P_{TST}(\omega) &= e^{[\tilde{q}_1(L_S(\omega-\omega_0)+L_S(\omega+\omega_0))+\tilde{q}_0 L_S(\omega)]}. && (9) \end{aligned}$$

3.3. Synthesis procedure and pre-filter

The pre-filter of the underlying model of consistent sampling corresponds to spectral smearing effects that are dependent on the specific implementation of the synthesis procedure. For example, when a window-based method for calculating the FIR response of given spectra is employed, the pre-filter corresponds to the power spectrum of the windowing function. When a sinusoidal model is employed and F0 is constant, the pre-filter yields a delta function.

4. Source periodicity and F0 extraction

The design objective of an F0 extractor for a speech analysis and synthesis system is to extract an F0 trajectory that is identical to the F0 trajectory generated by a re-synthesized version of the original signal. The fundamental period of the speech signal is updated on every glottal cycle. It is necessary for the F0 extractor to follow this cycle-by-cycle F0 change. To satisfy this condition, the F0 extractor has to operate pitch-synchronously or pitch-adaptively with temporal resolution comparable to the fundamental period. Both TANDEM and STRAIGHT spectra simultaneously satisfy finer temporal resolution and pitch synchronous analysis without need of precision in window positioning.

4.1. Notes on F0 extractors for previous STRAIGHT

It is better to briefly review F0 extractors [4, 11, 12] designed for the previous STRAIGHT and YIN [13]. Development of these extractors was motivated by observations that an extracted F0 trajectory has effects on reproduced speech quality, and failure in a voiced/unvoiced decision introduces severe degradation. The first two extractors are based on output behavior of log-linearly allocated bandpass filters having bandwidth able to separate fundamental components only. This design results in only the filter centering around the fundamental component having a stable output. The first F0 detector implementation [4] uses the total amount of AM and FM modulation as the measure to represent the "fundamentalness" of the filter outputs. The second implementation [11] extracts F0 candidates as fixed-points [11] of mapping from the filter center frequency to the output instantaneous frequency. The best candidate is selected based on the time-frequency stability of this mapping. Figure 3 illustrates stability-based candidate selection.

A common drawback of these methods is their sole reliance on the fundamental component. The harmonic structure of repetitive sounds is not directly taken into account in these methods. This sole reliance introduces susceptibility to environmental noise and missing fundamental problems. This problem was solved by introducing a repetition-based feature in generating candidates. YIN is a sophisticated extension of correlation-base methods and makes use of repetitive structure. By using an autocorrelation-based feature and an instantaneous frequency-based feature and introducing manually optimized post processing, a nearly defect-free F0 extractor was designed [12].

Introduction of these reliable F0 extractors revealed that it is not always relevant to assume one unique F0. Sometimes, vocal fold vibration has a hierarchical structure [15]. For example, a shorter cycle and a longer cycle are coupled to form a higher repetitive structure. Investigations on a framework that is capable of representing this phenomenon motivated introduction of a TANDEM based F0 extractor described in the following section.

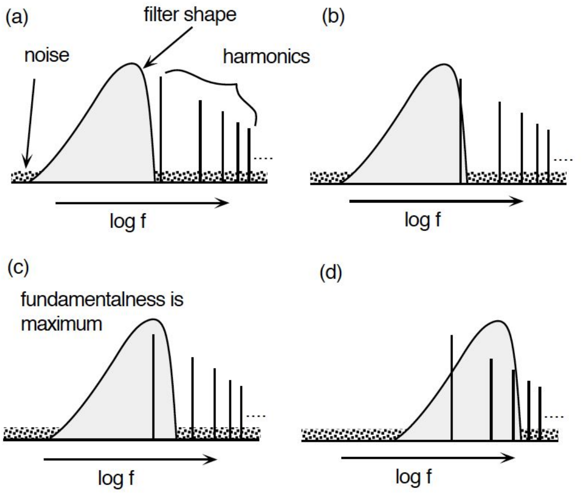


Figure 3: Bandpass filter design for extracting fundamental component [4]. Output modulation is minimum when only one harmonic component is isolated: case (c).

5. TANDEM-based F0 extraction

A repetitive structure introduces periodic modulation in a power spectrum. Normalizing power spectra with periodic modulation by corresponding smooth power spectra leaves only periodic modulation on the frequency axis. This is an underlying idea of the frequency domain method of F0 extraction [16, 17, 18]. TANDEM-based F0 extraction revisits this idea.

Assume that the F0 of a signal is temporally constant and known. Then define the fluctuation spectrum $P_C(\omega)$ using

$$P_C(\omega) = \frac{P_T(\omega)}{P_{TST}(\omega)} - 1. \qquad (10)$$

When the signal is a periodic pulse train and the analysis window for the TANDEM method is a Hanning window defined by Eq. (7), $P_C(\omega)$ yields a simple sinusoid $\cos(2\pi\omega/\omega_0)/4$. The Fourier transform of $P_C(\omega)$ has a unique peak at T_0 on the lag axis. Neither half nor double pitch peaks occur.

For more complex spectral shapes, $P_C(\omega)$ defined by Eq. 10 effectively represents the sinusoidal spectral variation component due to periodicity because the STRAIGHT spectrum closely approximates the spectral envelope. The sinusoidal modulation of the frequency axis reflecting signal periodicity found in the TANDEM spectrum is completely suppressed in the STRAIGHT spectrum. Normalizing the TANDEM spectrum by the STRAIGHT spectrum leaves constant bias plus sinusoidal spectral variations. Please note that it is not critical to use the correction filter derived in the previous section for Eq. 10 to work effectively. For F0 estimation and aperiodicity estimation, the following simplified definition of $P_{TST}(\omega)$ is used.

$$P_{TST}(\omega) = [C(\omega + \omega_0/2) - C(\omega - \omega_0/2)]/\omega_0, \qquad (11)$$

where $C(\omega)$ is defined by Eq. 8.

5.1. F0 detector without *a priori* information

When analyzing actual speech, F0 is not constant in time and is not known in advance. F0 changes in time introduce amplitude modulation of $P_C(\omega)$ on the frequency axis. This amplitude modulation is approximately modeled by $1 + \cos(c_m\omega)$. Modulation (spatial) frequency c_m is proportional to the speed of the F0 change. This modulation introduces spurious peaks in the Fourier transform of $P_C(\omega)$.

This artifact can be removed using the lower frequency portion of $P_C(\omega)$ with frequency weighting $w_{\omega 0,N}(\omega)$ defined in $[-N\omega_0, N\omega_0]$. N is set to satisfy $\pi/N\omega > c_m$. A practical implementation of $w_{\omega 0,N}(\omega)$ is given below:

$$w_{\omega 0,N}(\omega) = c_0\left(1 + \cos\left(\pi\omega/N\omega_0\right)\right), \qquad (12)$$

where c_0 is a constant so that $\int_{-\infty}^{\infty} w_{\omega 0,N}(\omega)d\omega = 1$.

Considering this, a weighted Fourier transform of the fluctuation spectrum is defined as

$$A(\tau; T_0) = \int_{-\infty}^{\infty} w_{\omega 0,N}(\omega)P_C(\omega; T_0)e^{-j\omega\tau}d\omega, \qquad (13)$$

where the assumed fundamental period T_0 is explicitly delineated. Note that $A(\tau; T_0)$ retains dominant peak salience.

Since no *a priori* information about the F0 is available, it is necessary to provide F0 candidates and to define a function to evaluate the possiblities. $A(\tau; T_0)$ does not have a peak at $\tau = T_0$ when the period of the input signal is $2T_0$ (by definition; see Eq. (7)). It also has a smaller peak at $\tau = T_0$ when the period of the input signal is $T_0/2$ because spectral smoothing does not attenuate spectral fluctuation effectively due to a mismatch in the F0 hypothesis. Therefore, it is necessary to reshape each $A(\tau; T_0)$ to have proper peak at T_0. This shaping is done by introducing a weighting function $w_{LAG}(\tau; T_C)$ in the lag domain. Taking these considerations into account, a weighted average of $A(\tau; T_0)$ is defined as follows to estimate T_0 by selecting the maximum peak:

$$\bar{A}(\tau) = \frac{1}{M}\sum_{k=1}^{M} w_{LAG}(\tau; T_L 2^{\frac{1-k}{L}})A\left(\tau; T_L 2^{\frac{1-k}{L}}\right), \quad (14)$$

where L represents the number of hypothesized F0 candidates in one octave. A constant T_L is the longest limit of the fundamental period, and M represents the total number of frequency bands.

5.2. Refinement of estimates

Parabolic interpolation around each peak is introduced to estimate F0 from quantized F0 candidates and their score $\bar{A}(\tau)$. Parabolic interpolation generally provides good estimates because the shaping weight is closely approximated by second order polynomials in the vicinity of peaks. Extracted F0 estimates by this interpolation are refined further using instantaneous frequencies of lower harmonic components. Details are given in Appendix B.

5.3. Response to random signals and weight design

In this section, response to random signals of this detector is discussed and a method to tune parameters is introduced. The upper plot of Figure 4 shows the distribution of peaks of $A(\tau; T_0)$ with a hypothesized $T_0 = 25$ Hz using a random signal as the input signal. Due to the spatial low-pass filtering effect of time windowing and the spatial high-pass filtering effect of the spectral normalization by the STRAIGHT spectrum, $A(\tau; T_0)$ has a band-pass characteristic. Because response to random noise represents interference to F0 estimation, it is necessary to make the ratio between response to the desired periodic signal and this random response be maximized. For example, when using

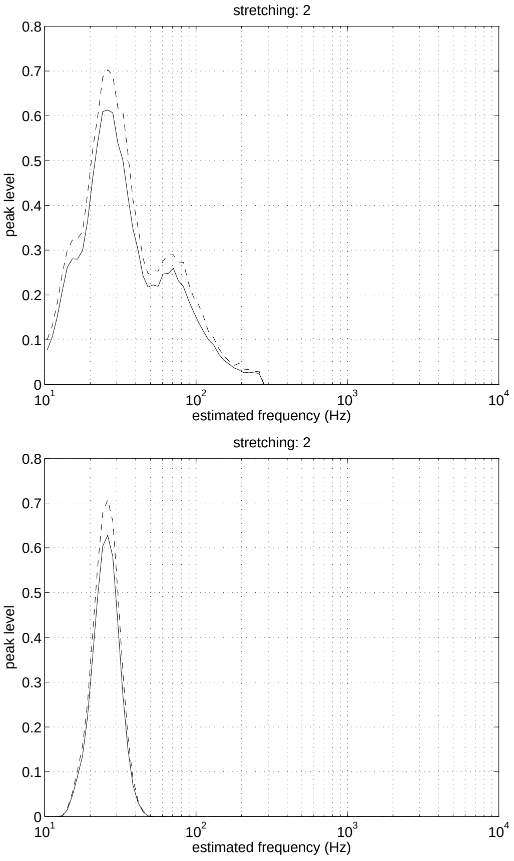


Figure 4: Single band upper bound for peak levels corresponding to 90% (solid line) and 95% (dashed line) cumulative probability when the size satisfies TANDEM condition and STRAIGHT conditions. (Blackman window with $4T_0$.)

a Blackmann window, the case illustrated in Figure 4, using $4T_0$ for window size yields the response of having maximum peak at F_0 and yields the best signal-to-noise ratio.

However, tuning the window length is not sufficient. As can be found in the upper plot of Figure 4, the response to noise has spurious peaks. The following shaping function is introduced to eliminate these additional peaks:

$$w_{LAG}(\tau;T_0) = 0.5 + 0.5\cos\left(\pi\log_2\left(\frac{\tau}{T_0}\right)\right) \quad . \tag{15}$$

The lower plot of Figure 4 shows the response after shaping. It does not have spurious peaks and is suitable for calculating the periodicity salience measure by allocating it for the desired F0 range using Eq. 14.

Figure 5 shows the distribution of peaks of $\bar{A}(\tau)$. F0 hypotheses in this case span from 40 Hz to 600 Hz with 2 hypotheses per one octave. The plot illustrates that the response is not dependent on F0 within the desired range. This implies that the periodicity score $\bar{A}(\tau)$ can be associated with the probability of peaks due to randomness.

Figure 6 illustrates the probability of random peaks as a function of the peak value. This relation can be used to determine the threshold values to detect voicing. For example, when acceptable false alarm for voicing detection is 1%, the threshold has to be set as 1.23. Thresholds have to be set as 1.43 and 1.53 for 0.1% and 0.01% false alarm rates respectively. Please note that this threshold setting does not assume that only one F0 exists at a specific moment, and that the maximum level of $\bar{A}(\tau)$ is uniquely determined as a function of the time window shape and F0 hypotheses spacing. In this simulation, the value is around 2.

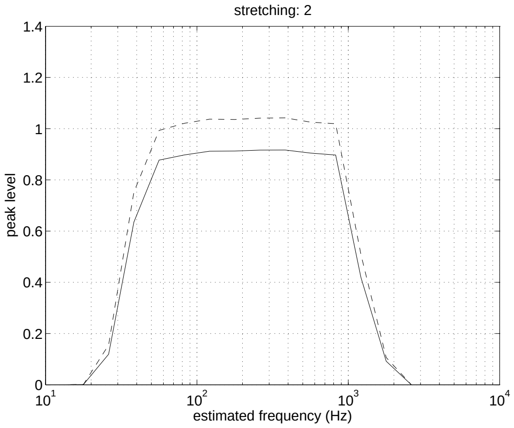


Figure 5: Multi band upper bound for peak levels corresponding to 90% (solid line) and 95% (dashed line) cumulative probability when the size satisfies TANDEM condition and STRAIGHT conditions. (Blackman window with $4T_0$. Two bands per octave allocation spans from 40 Hz to 600 Hz.)

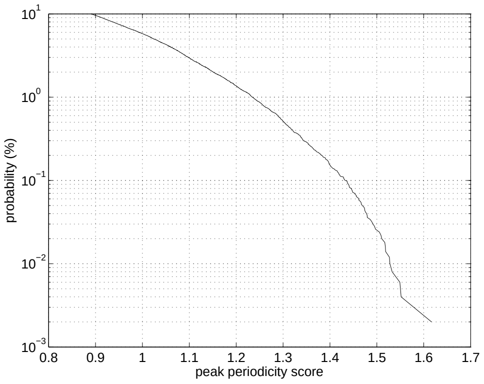


Figure 6: Probability of $\bar{A}(\tau)$ peaks for random inputs.

5.4. Response to pseudo-periodic signals

A test signal and a natural speech example were analyzed to test the proposed procedure. The upper plot of Figure 7 shows extracted F0 candidates for a pulse train with increasing repetition periods. A dot represents a candidate and a circle represents the best candidate of F0. The plot illustrates that the best candidates correspond to the correct trajectory. The lower plot of Figure 7 shows the periodicity score. It is clearly shown that the best score of each frame stands out. Please note that the other candidates do not have scores significantly higher than random peaks. Please also note that the waving behavior of some of the lower scores is due to the relatively sparse allocation of detectors. Increasing the number of detectors per octave makes wave length and depth small.

Figure 8 shows F0 candidates and their periodicity scores for a Japanese vowel sequence /auieo/ spoken by a male speaker. The best candidate at each frame is represented as

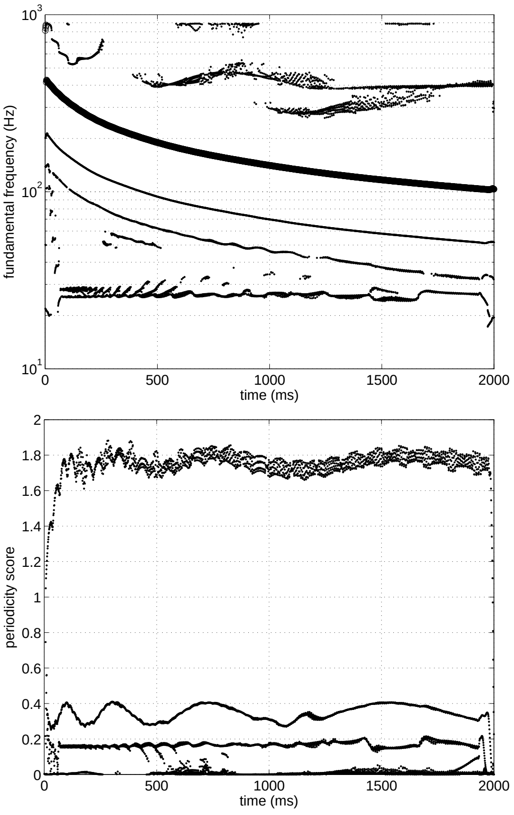


Figure 7: F0 candidates (upper plot) and their periodicity scores (lower plot). The most likely F0 candidate at each frame is represented using a circle. (Test signal: pulse train with increasing fundamental period)

an open circle in the upper plot. It is clearly shown that the best candidate also stands out. It is interesting to observe that the other candidates in the region from 300 ms to 450 ms also have relatively high peak values. They correspond to candidates around 300 Hz. It suggests that the fundamental period corresponding to the primary F0 ($\simeq$ 150 Hz) is subdivided into two periods.

6. Periodicity spectrogram

Speech sounds are not strictly periodic. F0 and amplitude fluctuations introduce FM and AM on each harmonic component. In addition, the excitation source signal fluctuates cycle by cycle, and the vocal-tract transfer function varies because of the movement of the articulators. These factors introduce deviations from the precise repetition of the waveform of each cycle. The deviations are also frequency-dependent and are represented as a periodicity spectrogram. Several approaches [11, 19] were tested to extract this periodicity spectrogram in the previous version of STRAIGHT. TANDEM and STRAIGHT spectra also introduce a unified approach for periodicity spectrogram extraction.

To define aperiodicity properly, these factors must be separated into two groups. The first group consists of factors dependent on F0 fluctuations and STRAIGHT spectral fluctuations. The second group consists of residual fluctuations. The aperiodicity that has to be defined for flexible speech manipulation is the second group. Effects caused by the first group have to be removed from the final results of aperiodicity analysis in order to prevent double counting, as both the F0 and the STRAIGHT spectrum are used in synthesizing the manipulated speech signals.

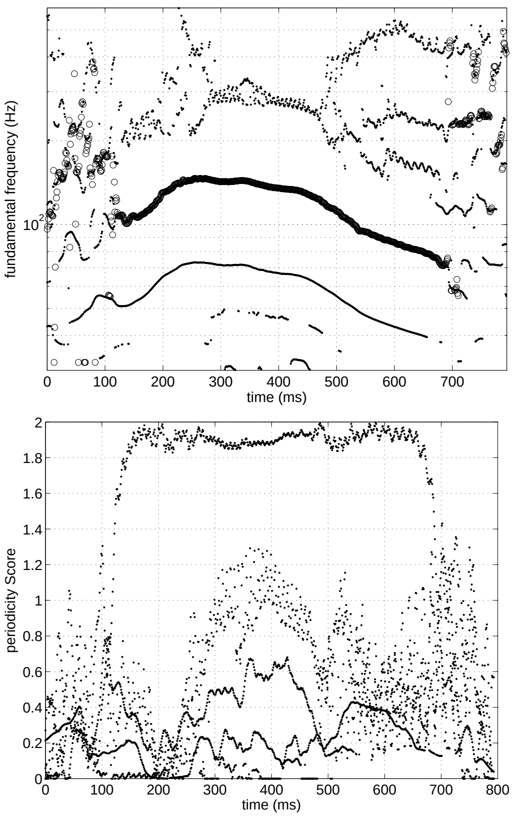


Figure 8: F0 candidates of natural speech. The material is a Japanese vowel sequence /aiueo/ spoken by a male speaker. (Upper plot) F0 candidates. The best candidate of each frame is represented as a circle. (lower plot) Periodicity scores of each candidate.

6.1. Normalization of F0 movement

Non-stationary F0 has to be stabilized prior to the following analysis because F0 movement is proportionally magnified by the harmonic numbers and introduces significant side band power due to frequency modulation in the higher frequency range. Converting the time axis t to $\tau(t)$ using the instantaneous frequency of the fundamental component $f_0(t)$ and target F0 f_{fix} in Equation $\tau(t) = \int_0^t f_{\text{fix}}/f_0(\lambda)d\lambda$, the F0 of the signal converted onto the new time axis has constant value f_{fix} [14, 11].

This F0 stabilization procedure eliminates the amplitude modulation of $P_C(\omega)$ on the frequency axis mentioned in section 5.1. Therefore, periodicity can be evaluated locally on the frequency axis irrespective of frequency position.

6.2. Periodicity extraction using quadrature signal

Since F0 is already known, the only interesting component of $A(\tau;T_0)$ is at $\tau = T_0$. Component $A(\tau;T_0)|_{\tau=T_0}$ is calculated using a quadrature signal $h_N(\omega)$ defined below.

$$h_N(\omega) = w_{\omega 0,N}(\omega)\exp\left(2\pi j\omega/\omega_0\right), \qquad (16)$$

where a signal envelope function $w_{\omega 0,N}(\omega)$ defines the spectral resolution of the aperiodicity calculation. In terms of TB (time and bandwidth) product, the wider the frequency span, the more reliable the estimation is.

In this implementation, the following raised cosine function is used as envelope $h_N(\omega)$ for simplicity.

$$w_{\omega_C,N}(\omega) = c_0\left(1+\cos\left(\pi\omega/N\omega_C\right)\right), \qquad (17)$$

where constant c_0 is used to normalize $\int w_{\omega_C,N}(\omega)d\omega = 1$. Using this quadrature signal, initial evaluation of periodicity is defined as follows:

$$Q_C^2(\omega;T_C) = \left|\int_{-\infty}^{\infty} h_N(\lambda;T_C)P_C(\omega-\lambda;T_C)d\lambda\right|^2 . \qquad (18)$$

The problem to be solved is estimation of the aperiodic component based on this periodicity measure.

6.3. Estimation of aperiodic component

Let $\circ$ represent convolution to make Eq. 18 simple. Using the definition of fluctuation spectrum Eq. 10, the following holds:

$$\begin{aligned} Q_C^2 &= \left|h_N \circ P_C(\omega;T_C)\right|^2 \\ &= \left|h_N \circ \frac{P_T(\omega;T_C)-P_{TST}(\omega;T_C)}{P_{TST}(\omega;T_C)}\right|^2 . \qquad (19) \end{aligned}$$

Periodic fluctuation in the TANDEM spectrum originates from two sources, signal periodicity and random fluctuation. Assume ΔP_P to represent the periodicity related fluctuation in the TANDEM spectrum and ΔP_R to represent the random component related fluctuation in the TANDEM spectrum. Also assume P_P to represent the STRAIGHT spectrum of the periodic component and P_R to represent the STRAIGHT spectrum of the random component. Then, as a first order approximation, assume that $P_P(\omega;T_C)$ and $P_R(\omega;T_C)$ are locally constant within the effective length of h_N. Then, by calculating the expectation, the following holds.

$$Q_C^2 = \frac{V[h_N \circ \Delta P_P]}{P_P^2+P_R^2} + \frac{V[h_N \circ \Delta P_R]}{P_P^2+P_R^2}, \qquad (20)$$

where $V[x]$ represents the variance of x.

For periodic signals, the ratio between $V[h_N \circ \Delta P_P]$ and P_P is uniquely determined as a constant C_P once a window is determined. For random signals, the ratio between $V[h_N \circ \Delta P_R]$ and P_R is also determined as a constant C_R depending on the effective TB product once the window function and h_N are determined. These yield the following:

$$Q_C^2 = \frac{C_P^2 P_P^2}{P_P^2+P_R^2} + \frac{C_R^2 P_R^2}{P_P^2+P_R^2}. \qquad (21)$$

Using these constants, the root mean squared value of the random component amplitude $a_{RND}(\omega)$ and periodic component amplitude $a_{PRD}(\omega)$ are represented by the following equation:

$$a_{RND}(\omega) = \sqrt{\frac{C_P^2-Q_C^2}{C_P^2-C_R^2}}\ ,\ a_{PRD}(\omega) = \sqrt{\frac{Q_C^2-C_R^2}{C_P^2-C_R^2}}. \qquad (22)$$

Note that these relations hold for expectations. Parameter extraction for speech resynthesis requires other criteria based on these relations. It is also necessary to evaluate effects due to spectral variation that were neglected in derivation up to this point.

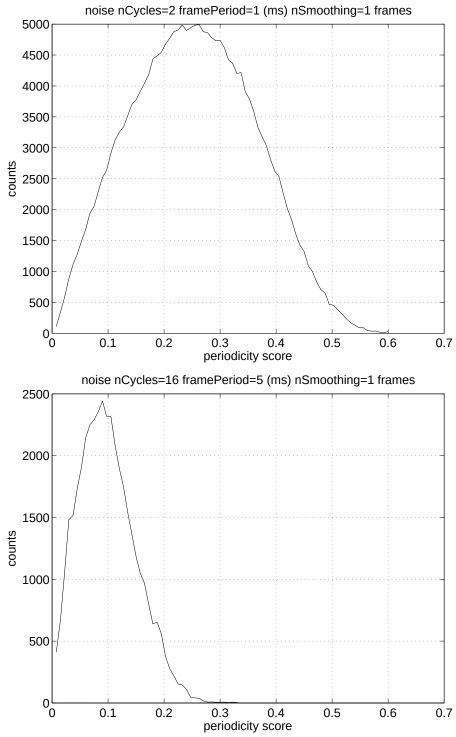


Figure 9: Histogram of periodicity rates. Frequency spreads are $N = 2$ for upper plot and $N = 16$ for lower plot.

6.4. Determination of constants C_P and C_R

The constant for periodic signals C_P represents the averaged amplitude of the spectral variation component having period ω_0 on the frequency axis. This value is directly calculated using the Fourier transform of the original time window $H(\omega)$ of Eq. 1, the frequency weighting window $w_{w0,N}(\omega)$ of Eq. 12, and the lag shaping function $w_{LAG}(\tau;T_0)$ of Eq. 15. Constant $C_P = 0.56$ for a Blackman window with the length $2.4T_0$ is used in the current implementation.

The constant for random signals C_R also depends on $H(\omega)$, $w_{w0,N}(\omega)$ and $w_{LAG}(\tau;T_0)$. One important difference of this parameter is that it is probabilistic. Distribution of C_R is highly affected by $h_N(\omega;T_C)$ in $w_{w0,N}(\omega)$. Figure 9 shows distribution of Q_C for $N = 2$ and $N = 16$ for random signal input. When N is small, small effective degrees of freedom (in other words, TB product) yield widely spread distribution and make reliable estimation of underlying aperiodicity difficult. It is necessary to increase the effective degrees of freedom for reliable estimation of aperiodicity. One practical solution is to calculate the average periodicity measure $\bar{Q}_c$ using M multiple frames.

An exhaustive simulation for all plausible combinations of N, M and quantized frame shift periods was conducted to calculate the average and variance of $\bar{Q}_C$. Simulation results were stored in a three-dimensional table to calculate $\bar{Q}_c$ for given analysis conditions using bilinear interpolation. Final amplitude estimates of the aperiodic component $a_{RND}(\omega)$ and the periodic component $a_{PRD}(\omega)$ are calculated using the average and variance of $\bar{Q}_C$. Please note that it is important to to set appropriate C_R and C_P.

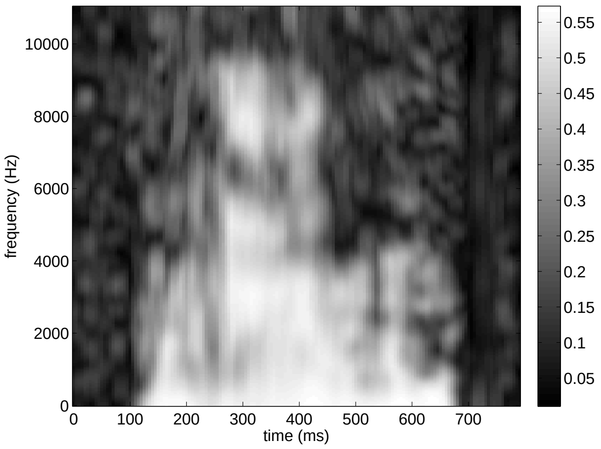


Figure 10: Observed periodic component $Q_C(\omega)$

6.5. Examples and comments

Since preliminary tests using simulated signals revealed that the proposed method performs as predicted, only the analysis for a natural speech example is presented here. A Japanese vowel sequence /aiueo/ spoken by a male speaker sampled at 22050 Hz was used. Figure 10 shows the $\bar{Q}_c$ map. It is observed that even in the silent period there are positive values in $\bar{Q}_c$ reflecting statistical fluctuation due to randomness. It is also observed that in the voiced part it does not completely reach a perfectly periodic value of 0.56.

Please note that an unsolved issue in representing aperiodicity remains. Due to highly nonlinear processing in the acoustic to neural conversion in human auditory system, sometimes the detection threshold of a brief burst noise varies more than 20 dB within one pitch period [20]. This effect has to be taken into account to represent aperiodicity for resynthesis.

7. TANDEM-STRAIGHT: Architecture

Figure 11 illustrates the schematic diagram of the revised STRAIGHT (TANDEM-STRAIGHT). Please note that a unified set of representations, TANDEM-spectrum and STRAIGHT-spectrum, is repeatedly used in all subsystems. This unified architecture and fewer tunable parameters are representative advantages of the new formulation over the previous version of STRAIGHT. However, in spite of this radical reformulation, the new TANDEM-STRAIGHT inherits all the functionality of the previous version of STRAIGHT. One important application of STRAIGHT is speech morphing [21], described in the next section.

8. Morphing of speech sounds

Morphing speech samples [21] introduces a complementary strategy, a deductive approach, for investigating the physical correlates of perceptual attributes. It enables us to generate a stimulus continuum between two or more exemplar stimuli by evenly interpolating STRAIGHT parameters even without knowing the physical correlates of a specific perceptual attribute in advance. This strategy was initially applied to investigate emotional expressions [22]. It was illustrated that using STRAIGHT-based morphing, the log-linear interpolation of a few speech parameters (namely, F0, STRAIGHT-spectrogram, and periodicity spectrogram) provides stimulus continuum in terms of emotional perception.

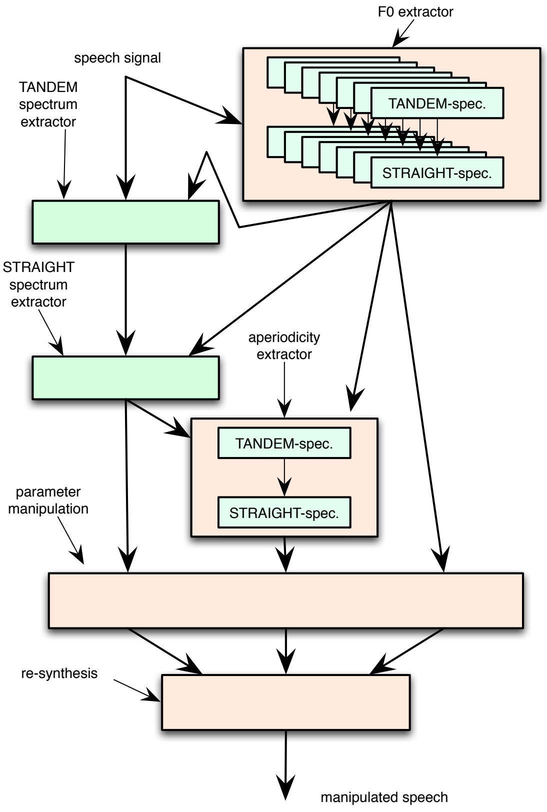
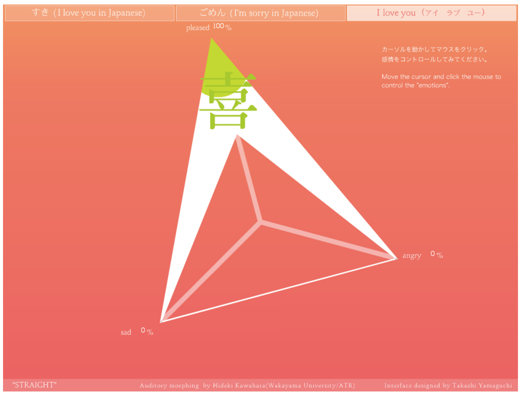


Figure 11: Schematic diagram of TANDEM-STRAIGHT.


すき（I love you in Japanese）
ごめん（I'm sorry in Japanese）
I love you（アイ　ラブ　ユー）
pleased 100 %
喜
angry 0 %
sad 0 %
"STRAIGHT"
Auditory morphing by Hideki Kawahara(Wakayama University/ATR)
Interface designed by Takashi Yamaguchi

カーソルを動かしてマウスをクリック。
感情をコントロールしてみてください。

Move the cursor and click the mouse to control the "emotions".

Figure 12: User interface for morphing demonstration (courtesy of the Mirainan, designed by Takashi Yamaguchi)

Emotional morphing demonstrations were displayed in the Miraikan, Japan's National Museum of Emerging Science and Innovation, from April 22 to August 15, 2005. Figure 12 shows a screenshot of the display. Three phrases were portrayed by one female and two male actors with three emotional styles (pleasure, sadness, and anger). Simple resynthesis of these original samples was placed at the vertices. Morphed sounds were located on the edges and the inside links of the triangle and reproduced by mouse clicks. Please note that only prosodic variation makes variations in perceived emotional expressions. This demonstration can be tested visiting the author's web page [6].

Singing is an extreme form of voice communication where stylized prosody plays a crucial role. STRAIGHT-based manipulations including speech morphing have been applied to investigate and control singing expressions [23, 24]. An extended morphing procedure that is capable of component-wise morphing was applied to singing voice manipulation [25]. In that study, individual control of the morphing rate using five extended parameters (in addition to the three parameters mentioned above, temporal axis mapping and frequency axis mapping functions were introduced) was implemented to test design reuse of singers' voice identity and their singing style. Subjective test results suggested that morphing of prosody-related physical parameters enables perceptual modifications of singing style. It also revealed that identification of the singer of the manipulated singing is mainly dependent on spectral information. Please note that these findings are dependent on specific singers participated in the experiment. Prosody-related physical parameters would play important roles when one or both of the singers have strong and unique singing style.

8.1. Application to L2 study

The exemplar-based approach that is based on morphing will serve as a powerful tool for investigating prosodic aspects of L2 perception and learning. A preliminary study was conducted to test if findings based on synthetic speech are replicated using morphing based procedures [26] and was encouraging. A real-time implementation of STRAIGHT [27] will be also useful for testing preliminary hypotheses on prosody-related aspects of speech perception.

9. Summary and conclusions

A unified framework was introduced based on a simple and novel power-spectrum estimation method called TANDEM that eliminates periodic temporal fluctuations. Based on this representation, extraction algorithms for interference-free spectrum (STRAIGHT spectrum), F0, and aperiodicity maps are formulated in a theoretically tractable manner. Preliminary tests indicated that the analysis results are compatible with the current version of STRAIGHT and yield re-synthesized speech that is indistinguishable from the current version. This reformulation of STRAIGHT makes it a more accessible and efficient tool for research communities. A STRAIGHT-based morphing procedure will serve as a powerful research tool for L2 study, especially for the prosodic aspect, because it enables a deductive approach and precise control of physical parameters at the same time.

10. Acknowledgments

The author deeply appreciates students, colleagues, collaborators and users of STRAIGHT for their contributions and support. Especially, the invention of TANDEM by Morise initiated the subsequent complete reformulation of STRAIGHT. Development of STRAIGHT and investigations based on its applications have been supported by CREST project by Japan Science and Technology Agency, "e-Society project" of Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research by Japan Society for the Promotion of Science (JSPS) and advanced research project of Wakayama University. This work is currently supported partly by Grant-in-Aid for Scientific Research (A)19200017 by JSPS and CrestMuse [28] project by JST.

11. References

[1] Dudley, H., "Remaking speech," J. Acoust. Soc. Amer., 11(2): 169–177, 1939.

[2] Itakura, F., and Saito, S., "A statistical method for estimation of speech spectral density and formant frequencies," Electro. Comm. Japan, 53-A(1): 36–43, 1970. [originally in Japanese]

[3] Atal, B. S., and Hanauer, S. L., "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Amer., 50: 637–655, 1971.

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

[5] Kawahara, H., "STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," Acoustic Science & Technology, 27(5): 349–353, 2006.

[6] Please check links from the following page. <http://www.wakayama-u.ac.jp/˜kawahara/>

[7] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., "TANDEM–STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," ICASSP 2008, [in press].

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

[9] Unser, M., and Aldroubi, A., "A general sampling theory for nonideal acquisition devices," IEEE Trans. Signal Processing, 42(11): 2915–2925, 1994.

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

[11] Kawahara, H., Katayose, H., de Cheveigné, A., and Patterson, R. D., "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," Proc. EUROSPEECH'99, 6: 2781–2784, 1999.

[12] Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., and Irino, T., "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT," Proc. Interspeech'2005, 537–540, 2005.

[13] de Cheveigné, A., and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Amer., 111(4): 1917–1930, 2002.

[14] Abe, T., Kobayashi, T., and Imai, S., "The IF Spectrogram: A New Spectral Representation," Proc. ASVA-97, Tokyo, 423–430, 1997.

[15] Titze, I., *Principles of voice production*, Prentice-Hall, 1994.

[16] Noll, A. M., "Cepstrum Pitch Detection," J. Acoust. Soc. Amer., 41(2): 293–309, 1967.

[17] Itakura, F., and Saito, S., "Analysis synthesis telephony based on the maximum likelihood method," Rep. 6th Int. Congr. Acoust., Tokyo, C-5-5, 1968.

[18] Fujisaki, H., and Tanabe, Y. A., "Time-Domain Technique for Pitch Extraction of Speech," J. Acoust. Soc. of Japan, 29(7): 418–419, 1973.

[19] Kawahara, H., Morise, M., Takahashi, T., Banno, H., Irino, T., and Fujimura, O., "Group delay for acoustic event representation and its application for speech aperiodicity analysis," Proc. EUSIPCO, 2007.

[20] Skoglund, J., and Kleijn, W. B., "On time-frequency masking in voiced speech," IEEE Trans. Speech and Audio Processing, 8(4): 361–369 2000.

[21] Kawahara, H., and Matsui, H., "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," ICASSP'03, Hong Kong, I: 256–259, 2003.

[22] Matsui, H., and Kawahara, H., "Investigation of Emotionally Morphed Speech Perception and its Structure Using a High Quality Speech Manipulation System," Eurospeech'03, 3157–3160, 2003.

[23] Yonezawa, T., Suzuki, N., Mase, K., and Kogure, K., "HandySinger: Expressive singing voice morphing using personified hand-puppet Interface," Proc. NIME2005, 121–126, 2005.

[24] Saitou, T., and Goto, M., and Unoki, M., and Akagi, M., "Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 21-24 Oct. 2007, 215–218, 2007.

[25] Kawahara, H., Ikoma, T., Morise, M., Takahashi, T., Toyoda, K., and Katayose, H., "Proposal on a morphing-based singing design manipulation interface and its preliminary study," Trans. IPSJ, 48(12): 3637–3648, 2007. [in Japanese].

[26] Kawahara, H., and Akahane-Yamada, R., "STRAIGHT as a research tool for L2 study: How to manipulate segmental and supra-segmental features," ASA and ASJ joint meeting, Hawaii, 2pSCb2, 2006.

[27] Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., and Kawahara, H., "Implementatioin of realtime STRAIGHT speech manipulation system: Report on its first implementation," Acoustic Science and Technology, 28(3): 140–146, 2007.

[28] <http://www.crestmuse.jp/index-e.html>

[29] Flanagan, J. L., and Golden, R. M., "Phase vocoder," Bell Syst. Tech. J., 45: 1493–1509, (1966).

[30] Charpentier, F. J., "Pitch detection using the short-term phase spectrum," Proc. ICASSP 1986, 113–116, 1986.

A. Zero frequency recovery

The power spectrum of speech has zero at zero frequency due to radiation impedance. This spectral zero has to be removed for two reasons in STRAIGHT spectrum calculation. One reason is for the re-synthesis procedure where a minimum phase response is calculated based on a complex cepstrum. Zero is incompatible with the cepstrum. The other reason is for F0 extraction of "pure tones." A single sinusoidal signal only has two components as its Fourier transform. The direct application of Eq. 10 only yields two peaks at $\pm\omega_0$ on the (angular) frequency axis. Distance between these peaks is $2\omega_0$ and yields $2F_0$ as the extracted fundamental frequency.

To avoid these problems, $P_{Tr}(\omega)$ is reshaped using the following equation in a region $(-\omega_0, \omega_0)$:

$$P_{Tr}(\omega) = w_{sh}(|\omega_0 - \omega|)P_T(\omega) + w_{sh}(\omega)P_T(|\omega_0 - \omega|), \quad (23)$$

where the weighting function has to satisfy the following relation.

$$\begin{aligned} 1 &= w_{sh}(|\omega_0 - \omega|) + w_{sh}(\omega) \\ w_{sh}(\omega) &= w_{sh}(-\omega) \\ \frac{dw_{sh}(\omega)}{d\omega} &= 0 \text{ at } \omega = \pm\omega_0,\ \omega = 0. \end{aligned} \quad (24)$$

The following raised cosine function is used in this implementation:

$$w_{sh}(\omega) = 0.5 + 0.5\cos\frac{\pi\omega}{\omega_0}. \quad (25)$$

B. Instantaneous frequency-based refinement

A procedure to refine the initial estimate of the fundamental frequency is designed based on instantaneous frequency and linear interpolation. As the initial estimate of F0 is available, it is possible to design a bandpass filter that extracts only the fundamental component. Such a filter has the following form:

$$h_N(t;\omega) = w_{T,N}(t)\exp\left(-2\pi j\omega t\right), \quad (26)$$

where ω represents the center frequency of the bandpass filter and $w_{T,N}(t)$ defines the envelope of the response. The Fourier transform of the envelope $w_{T,N}(t)$ has to have low-pass characteristics with suppression band from $(1-\alpha)\omega_0$, where α is a small positive number such as 0.1. In this implementation a Blackman Harris window function is used as $w_{T,N}(t)$ with the following form:

$$w_{T,N}(t) = 0.42 + 0.5\cos\frac{\pi t}{NT} + 0.08\cos\frac{2\pi t}{NT}, \quad (27)$$

where $-NT < t < NT$ and $w_{T,N}(t)$ is 0 outside.

To refine initial estimates of the fundamental frequency, instantaneous frequencies of filter outputs with center frequencies $\omega_a = (1-\alpha)\omega_0$ and $\omega_b = (1+\alpha)\omega_0$ have to be calculated. They can be calculated using Flanagan's method [29]. The refinement procedure is based on a finding that the output instantaneous frequency and the frequency of one dominant sinusoidal signal in the filter pass-band equals when the center frequency is set equal to that of the sinusoidal signal [30, 14, 11].

Assume that the output instantaneous frequencies of filters with center frequencies ω_a and ω_b are λ_a and λ_b respectively. Assume that the output instantaneous frequency is a linear function of ω in the vicinity of the initial estimate. Then, the following holds:

$$\begin{bmatrix} \lambda_a \\ \lambda_b \end{bmatrix} = \begin{bmatrix} \omega_a & 1 \\ \omega_b & 1 \end{bmatrix} \begin{bmatrix} u_a \\ u_b \end{bmatrix}. \quad (28)$$

Applying the condition that the center frequency and the instantaneous frequency have to be equal yields the following:

$$\omega_{r1} = \frac{u_b}{1-u_a}, \quad (29)$$

where ω_{r1} is the refined fundamental angular frequency.

This refinement procedure is applied again using the revised ω_{r1} as the initial estimate. Based on preliminary analysis, $N = 3$ is used in Eq. 27 and instantaneous frequencies of the lower three harmonic components are used to calculate the final estimate.