TANDEM-STRAIGHT: A TEMPORALLY STABLE POWER SPECTRAL REPRESENTATION FOR PERIODIC SIGNALS AND APPLICATIONS TO INTERFERENCE-FREE SPECTRUM, F0, AND APERIODICITY ESTIMATION

*H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino**

Faculty of Systems Engineering
Wakayama University
930 Sakaedani, Wakayama, 640–8510 Japan

H. Banno

Meijo University
Nagoya, 468–8502 Japan

ABSTRACT

A simple new method for estimating temporally stable power spectra is introduced to provide a unified basis for computing an interference-free spectrum, the fundamental frequency (F0), as well as aperiodicity estimation. F0 adaptive spectral smoothing and cepstral liftering based on consistent sampling theory are employed for interference-free spectral estimation. A perturbation spectrum, calculated from temporally stable power and interference-free spectra, provides the basis for both F0 and aperiodicity estimation. The proposed approach eliminates ad-hoc parameter tuning and the heavy demand on computational power, from which STRAIGHT has suffered in the past.

***Index Terms*—** periodic signal, power spectrum, consistent sampling, periodicity, speech processing

1. INTRODUCTION

STRAIGHT [1], a speech analysis, modification, and synthesis system (as well as speech morphing based on it) is widely used in the speech research community [2, 3, 4, 5, 6]. STRAIGHT decomposes input speech into three types of positive valued parameters: an interference-free spectrogram, an aperiodicity map, and a fundamental frequency (F0) trajectory. Despite the conceptual simplicity of the parameters, the procedures for extracting them are complicated. They consist of nonlinear transformations and many coefficients for tuning performance, making theoretical analysis of the system intractable. This paper proposes a complete reformulation of STRAIGHT, based on a unified approach that uses a novel, simple method for calculating a temporally stable power spectrum.

2. TEMPORALLY STABLE SPECTRUM

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 [7]. 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 of τ, the power spectrum of test signal $|S(\omega,\tau)|^2$ is given by

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

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").

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 a finer temporal resolution is necessary to track the dynamics of speech sounds. The effective duration of the TANDEM window can be made shorter than the fundamental period of the signal while retaining temporal stability.

3. INTERFERENCE-FREE 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 transfer function of the vocal tracts is not band-limited, adopting a formulation of consistent sampling [8] is a better approach. A brief summary and excerpts of the main theorem [8, 9] are given below.

3.1. Consistent sampling (excerpts and summary)

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:

$$a_{12}(k) = \langle \varphi_1(t-k), \varphi_2(t) \rangle. \tag{2}$$

*Partially supported by Grants-in-Aid for Scientific Research (A)19200017 by JSPS.

Theorem 2 [8, 9] Let $f \in H$ be an unknown input function. Provided $m > 0$ exists such that $|A_{12}(e^{j\omega})| \geq m$ i.e., then there is unique signal approximation $\tilde{f}$ in $V(\varphi_2)$ that is consistent with f in the sense that

$$\forall f \in H,\ \ c_1(k) = \langle f, \varphi_1(x-k)\rangle = \langle \tilde{f}, \varphi_1(x-k)\rangle. \quad (3)$$

This signal approximation is given by

$$\tilde{f} = \tilde{P}f(x) = \sum_{k\in Z}(c_1 * q)\varphi_2(x-k), \quad (4)$$

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

$$Q(z) = \frac{1}{\sum_{k\in Z} a_{12}(k)z^{-k}} \quad (5)$$

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. In this interpretation, a TANDEM spectrum is an output 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 are missing in this case. The problem to be solved is how to design the missing correction filter and how to modify the post-filter to satisfy consistency.

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

$$h(t) = (1 + \cos(\pi t/T_0))/2. \quad (6)$$

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 the Harr function $r_{\omega 0}(\omega)$ when the width is set to ω_0.

The next step is to design the correction filter by introducing this Harr function into the post-filter. Coefficients of the correction filter q_k are calculated using Eq. (5) 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.

3.3. Practical implementation issues I

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:

$$C(\omega) = \int_{\omega_L}^{\omega} P_T(\lambda)d\lambda \quad (7)$$

$$L_S(\omega) = \ln\left[C(\omega + \omega_0/2) - C(\omega - \omega_0/2)\right] \quad (8)$$

$$P_{TST}(\omega) = e^{[\bar{q}_1(L_S(\omega-\omega_0)+L_S(\omega+\omega_0))+\bar{q}_0 L_S(\omega)]}. \quad (9)$$

3.4. 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. FUNDAMENTAL FREQUENCY ESTIMATION

The design objective of an F0 extractor for speech analysis and synthesis 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 at 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 that of the fundamental period. Both TANDEM and STRAIGHT spectra simultaneously satisfy a finer temporal resolution requirement and essentially yield pitch synchronous analysis without the need for precision in window positioning.

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

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

When the signal is a periodic pulse train and the analysis window for the TANDEM method is a Hanning window defined by Eq. (6), $P_C(\omega)$ 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, this relation still holds because the STRAIGHT spectrum closely approximates the spectral envelope. The sinusoidal modulation of the frequency axis reflecting signal periodicity is completely suppressed in the STRAIGHT spectrum. Therefore, $P_C(\omega)$ consists only of the effect of signal periodicity.

4.1. Practical implementation issues II

When analyzing actual speech, F0 is not constant in time and is not known in advance. F0 changes over 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

$$w_{\omega 0,N}(\omega) = c_0\left(1 + \cos\left(\pi\omega/N\omega_0\right)\right), \quad (11)$$

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

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

$$A(\tau; T_0) = \int_{-\infty}^{\infty} w_{\omega 0,N}(\omega) P_C(\omega; T_0) e^{-j\omega\tau} d\omega, \qquad (12)$$

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

Since no *a priori* information about the F0 is available, it is necessary to provide F0 candidates and to define a function to evaluate their possiblities. A weighting function $w_{LAG}(\tau; T_0)$, used to select the best response of each periodicity detector, is introduced to integrate each $A(\tau; T_0)$ to yield a F0 periodicity score $\bar{A}(\tau)$:

$$\bar{A}(\tau) = C_0 \sum_{k=1}^{M} w_{LAG}(\tau; T_L 2^{\frac{1-k}{L}}) A\left(\tau; T_L 2^{\frac{1-k}{L}}\right), \qquad (13)$$

where L represents the number of frequency bands in one octave. A constant T_L is the longest limit of the fundamental period, and M represents the total number of frequency bands. A coefficient C_0 is introduced to give the periodicity score a value of 1 when a purely periodic signal is analyzed. A preliminary test with $T_L = 32$(ms) suggested that $L = 2$ and $M = 9$ using frequency weighting $w_{\omega 0,N}(\omega)$ with $N = 4$ provides for reasonable F0 coverage and precision. A raised cosine centered at T_0 was used for $w_{LAG}(\tau; T_0)$.

5. APERIODICITY ESTIMATION

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 due to the movement of the articulators. These factors introduce deviations from the precise repetition of the waveform of each cycle.

To define aperiodicity properly, these factors must be separated into two groups. The first consists of factors dependent on F0 and the STRAIGHT spectrum. Effects caused by these factors have to be removed prior to the aperiodicity analysis in order to prevent double counting, as both the F0 and the STRAIGHT spectrum are used in synthesizing the speech signals. It is also reasonable to assume that F0 and the STRAIGHT spectrum are already available before aperiodicity analysis.

As a first-order approximation, assume that the effects of factors in the second group are random. Let σ_P^2 and σ_N^2 represent the power of the periodic and aperiodic (random) components, respectively. Let $\sigma_{P.obs}^2$ represent the power of the periodic component in the observed fluctuation spectrum $P_C(\omega; T_0)$. The power of the periodic component yields $\sigma_{P(\text{window})}^2$ when the signal is purely periodic. The value of $\sigma_{P(\text{window})}^2$ depends on the windowing function used for the TANDEM method. For example, $\sigma_{P(\text{Hanning})}^2$ is $1/16$, as found in the discussion pertaining to Eq. (10). The following equation defines aperiodicity implicitly using σ_P^2 and σ_N^2:

$$\frac{\sigma_P^2}{\sigma_P^2 + \sigma_N^2} = \frac{\sigma_{P.obs}^2}{\sigma_{P(\text{window})}^2}. \qquad (14)$$

5.1. Practical implementation issues III

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

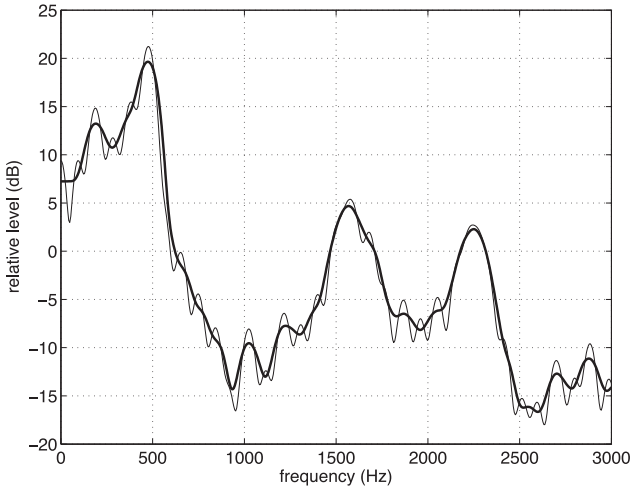

Fig. 1. TANDEM (light thin line) and STRAIGHT spectra (dark thick line) of the Japanese vowel /e/.

$\tau(t) = \int_0^t f_{\text{fix}}/f_0(\lambda)d\lambda$, the F0 of the signal converted onto the new time axis has a constant value f_{fix} [10, 11].

Because this transformation giving the input signals a constant F0 eliminates the amplitude modulation of $P_C(\omega)$ on the frequency axis, the transformation also eliminates the size limit N of frequency weighting in Eq. (11). 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 without relocation:

$$h_N(\omega) = w_{\omega 0,N}(\omega) \exp\left(2\pi j\omega/\omega_0\right). \qquad (15)$$

Aperiodicity is frequency-dependent. It is necessary to represent σ_P^2 and σ_N^2 as a function of ω. Since they are calculated from the power of the periodic component $\sigma_{P.obs}^2$ in the fluctuation spectrum $P_C(\omega; T_0)$, they can be estimated by calculating the convolution of $h_N(\omega)$ and $P_C(\omega; T_0)$. Let $\tilde{\sigma}_{P.obs}^2$ represent observed $\sigma_{P.obs}^2$. It follows that

$$\begin{aligned} \tilde{\sigma}_{P.obs}^2(\omega) &= \left| \int_{-\infty}^{\infty} h_N(\lambda) P_C(\omega - \lambda; T_0) d\lambda \right|^2 \qquad (16) \\ &= \sigma_{P.obs}^2(\omega) + \varepsilon_{w_N} \tilde{\sigma}_N^2(\omega), \end{aligned}$$

where ε_{w_N} represents a coefficient of a leaky, random component, since the component selectivity of $h_N(\omega)$ is not generally sharp. A preliminary test suggested that $N = 8$ provides a reasonable compromise between frequency selectivity and statistical fluctuation due to the degrees of freedom.

6. EXAMPLES

Since preliminary tests using simulated signals revealed that the proposed method performs as predicted, only the analysis for a natural speech example is presented. The Japanese vowel sequence /aiueo/ spoken by a male speaker sampled at 22050 Hz was used. Refer to the figure captions for discussion. Note that the STRAIGHT spectrum of this sample was calculated at 690 ms using Matlab on a Macinstosh Intel Core Duo 2.16 GHz computer (OS X) This calculation speed is faster than real time.

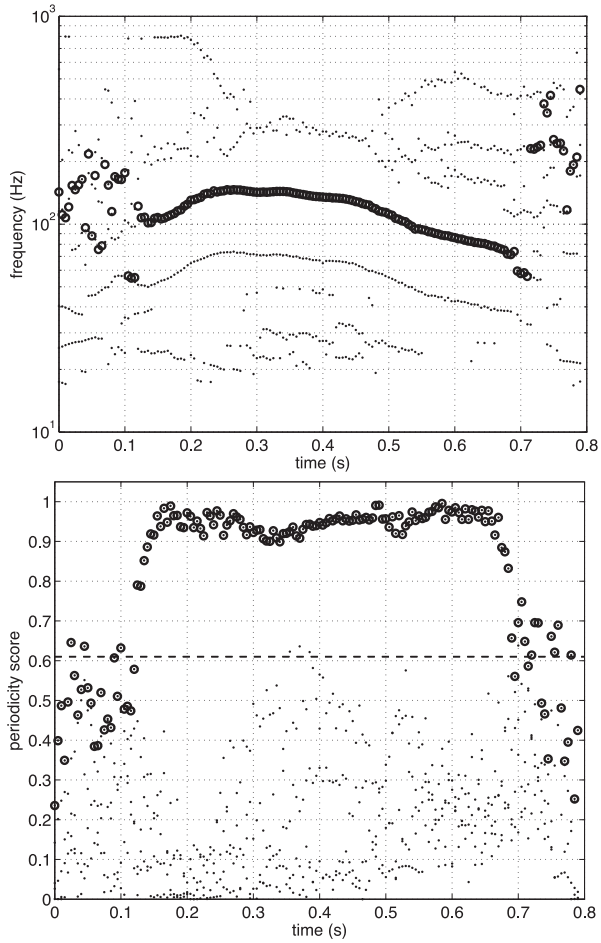

Fig. 2. F0 extraction. Upper plot shows F0 candidates. Five candidates are plotted for each frame. Thick open circles represent the best candidate for each frame. The bottom plot shows periodicity score $\bar{A}(\tau)$ of each candidate defined by Eq. 13. Only 1% of candidates due to random fluctuation a have higher periodicity score that exceeds the dashed line in the plot.

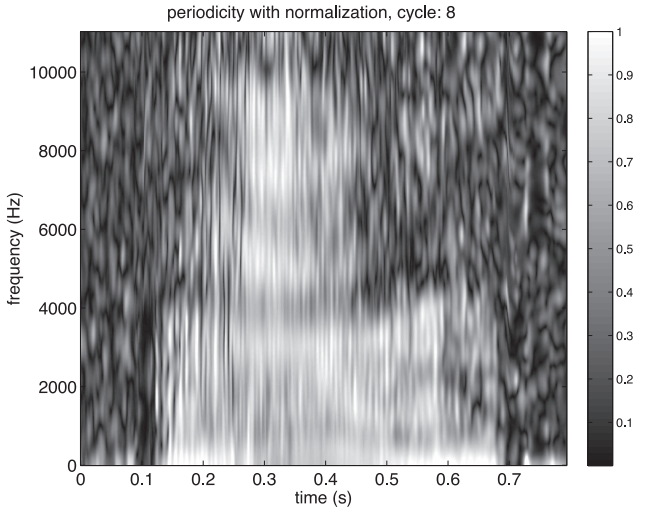

Fig. 3. Observed relative magnitude $\tilde{\sigma}^2_{P.obs}(\omega)$ of periodic component defined by Eq. 16. The magnitude $\tilde{\sigma}^2_{P.obs}(\omega)$ is normalized to have a value between 0 and 1.

7. CONCLUSION

A unified framework was introduced based on a simple and novel power-spectrum estimation method called TANDEM, which 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. Optimization and evaluation of the TANDEM-STRAIGHT approach are planned in the near future.

8. REFERENCES

[1] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

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

[3] K. von Kriegstein, J.D. Warren, D.T. Ives, R.D. Patterson, and T.D. Griffiths, "Processing the acoustic effect of size in speech sounds," *NeuroImage*, vol. 32, no. 1, pp. 368–375, 2006.

[4] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, no. 3–4, pp. 405–417, 2005.

[5] P. F. Assmann and W. F. Katz, "Synthesis fidelity and time-varying spectral change in vowels," *J. Acoust. Soc. Am.*, vol. 117, pp. 886–895, 2005.

[6] C. Liu and D. Kewley-Port, "Vowel formant discrimination for high-fidelity speech," *J. Acoust. Soc. Am.*, vol. 116, pp. 1224–1233, 2004.

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

[8] Michael Unser and Akram Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 2915–2925, 1994.

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

[10] T. Abe, T. Kobayashi, and S. Imai, "The if spectrogram: A new spectral representation," in *Proc. ASVA-97*, Tokyo, 1997, pp. 423–430.

[11] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. Eurospeech'99*, 1999, vol. 6, pp. 2781–2784.