A unified approach for F0 extraction and aperiodicity estimation based on a temporally stable power spectral representation

Hideki Kawahara¹, Masanori Morise¹, Toru Takahashi², Ryuichi Nisimura¹, Hideki Banno³, Toshio Irino¹

¹Faculty of Systems Engineering, Wakayama University, Wakayama, Japan
²Graduate School of Informatics, Kyoto University, Kyoto, Japan
³Faculty of Science and Technology, Meijo University, Nagoya, Japan

kawahara@sys.wakayama-u.ac.jp

Abstract

A power spectrum estimation method for periodic signals was proposed to provide temporally stable representation and has been applied to reformulate STRAIGHT, a system for speech analysis modification and synthesis based on stable spectral envelope estimation. This article proposes a specialized F0 detector based on a ratio between this stable spectrum and corresponding spectral envelope. By allocating multiple specialized F0 detectors and integrating individual clues, the proposed method selectively detects only fundamental components and yields a probability measure for each estimate. It also provides a method to estimate aperiodicity in each frequency band by making use of estimated fundamental frequency information to design a quadrature signal on the frequency axis for filtering periodic spectral component due to the signal periodicity. The proposed method shed new lights on source filter representation/decomposition of speech signals.

Index Terms: periodic signal, STRAIGHT, fundamental period, aperiodic component, probability

1. Introduction

It is still amazing to listen to the voice of VODER that was generated by human operation using pre-computer age technologies. It effectively demonstrated that speech can be transmitted using a far narrower frequency bandwidth, which was an important motivation of telecommunication research in the 1930s. This aim was recapitulated in the original paper on VOCODER [1] and led to the development of speech coding technologies. The demonstration also provided a foundation for the conceptualization of a source filter model of speech sounds, the other aspect of VOCODER.

It is not a trivial concept that our auditory system decomposes input sounds in terms of excitation (source) and resonant (filter) characteristics [2]. Retrospectively, this decomposition can be considered an ecologically relevant strategy that evolved through selection pressure. This article tries to focus on this conceptualization, especially on excitation “source,” based on a recent finding of a temporally stable power spectral representation of periodic signals.

2. STRAIGHT and its reformulation

3. 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 [3]. 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 \( \tau \), the power spectrum of test signal \( |S(\omega, \tau)|^2 \) is given by

\[
|S(\omega, \tau)|^2 = H^2(\omega) + \alpha^2 H^2(\omega - \omega_0) + 2\alpha H(\omega)H(\omega - \omega_0)\cos(\omega_0\tau + \beta).
\]

The third term consists of window location \( \tau \) 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”).

3.1. Envelope estimation

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 envelope estimation problem becomes discrete-to-analog conversion on the frequency axis.

A speech analysis, modification and resynthesis framework STRAIGHT [4, 5] is based on this concept and recently reformulated using TANDEM. It is worthwhile to note that the transfer function of the vocal tracts is not band-limited on the frequency axis. This suggests that it is more relevant to adopt a formulation of consistent sampling [6] than to adopt classical sampling theory. Please refer to our ICASSP’08 article [7] for details.

In short, it is virtually always possible to design a digital compensation filter to retain spectral information at harmonic frequencies while eliminating periodic variation on the frequency axis. This procedure is approximately implemented using the following equations. A stable envelope spectrum, \( P_{\text{EST}}(\omega) \) (“STRAIGHT spectrum” below), is calculated from the TANDEM spectrum \( P_{\text{T}}(\omega) \) using the following set of equa-

June 4-6, 2008, Aalborg University, Denmark
tions:
\[ C(\omega) = \int_{-\infty}^{\infty} P_T(\lambda) d\lambda \]  
\[ L_S(\omega) = \ln \left[ C(\omega + \omega_0/2) - C(\omega - \omega_0/2) \right] \]  
\[ P_{TST}(\omega) = \frac{\| q \| L_S(\omega + \omega_0 + \omega) - L_S(\omega + \omega_0 - \omega) \|}{\| q \| L_S(\omega)} \]
where \( q \) and \( \| q \| \) represent the coefficients of the compensation filter. (They are numerically optimized because the coefficients were truncated from a rapidly decaying infinite sequence.)

4. F0 extraction

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
\[ P_C(\omega) = \frac{P_T(\omega)}{P_{TST}(\omega)} - 1. \]  
When the signal is a periodic pulse train and the analysis window for the TANDEM method is a Hanning window with two pitch period in length, \( P_C(\omega) \) yields a simple sinusoidal \( \cos(2\pi\omega/\omega_c) / A \). The Fourier transform of \( P_C(\omega) \) has a prominent peak at \( T_0 \) on the log axis. 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. Specialized periodicity detector

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 \( 1 + \cos(\pi\omega/\omega_c) \). Modulation (spatial) frequency \( \omega_c \) 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 \( w_{\omega_c,N}(\omega) \) defined in \([-N\omega_0, N\omega_0]\). \( N \) is set to satisfy \( \pi/\omega_c > N \). A practical implementation of \( w_{\omega_c,N}(\omega) \) is
\[ w_{\omega_c,N}(\omega) = c_0 \left( 1 + \cos \left( \frac{\pi\omega}{N\omega_0} \right) \right), \]  
where \( c_0 \) is a constant so that \( \int_{-\infty}^{\infty} w_{\omega_c,N}(\omega) d\omega = 1. \)

Considering these factors, a weighted Fourier transform of the fluctuation spectrum is defined as
\[ A(\tau; T_0) = \int_{-\infty}^{\infty} w_{\omega_c,N}(\omega) P_C(\omega; T_0) e^{-j\omega\tau} d\omega, \]  
where the assumed fundamental period \( T_0 \) is explicitly delineated. The measure \( A(\tau; T_0) \) is considered to be a periodicity detector.

Magnitude of this measure \( A(\tau; T_0) \) has band-pass behavior, because time windowing introduces low-pass filtering on the frequency axis and spectral normalization using STRAIGHT spectrum introduces high-pass filtering on the frequency axis. In other words, it is possible to design its best response to match \( T_0 \) by adjusting the length of the time window.

Figure 1 shows equal probability contour of peak frequencies of \( A(\tau; T_0) \) for random signal input. Please note that the dominant peak is located at 25 Hz and that corresponds to \( 1/T_0 \) by using a Blackman window with \( 4T_0 \) in length. \( T_0 = 40 \text{ms} \).

4.2. Integration of specialized F0 detectors

Since no \textit{a priori} information about the F0 is available, it is necessary to provide F0 candidates and to define a function to evaluate their possibilities. A weighting function \( w_{LAG}(\tau; T_0) \), used to suppress spurious peaks and to select the best response of each periodicity detector, is introduced to integrate each \( A(\tau; T_0) \) to yield a F0 periodicity score \( A(\tau) \):
\[ A(\tau) = \frac{1}{M} \sum_{k=1}^{M} w_{LAG}(\tau; T_2) e^{j\frac{2\pi k\tau}{T_2}}, \]  
where \( L \) represents the number of frequency bands in one octave. A constant \( T_2 \) is the longest limit of the fundamental period, and \( M \) represents the total number of frequency bands. The weighting function \( w_{LAG}(\tau; T_0) \) is defined as follows in the current implementation.
\[ w_{LAG}(\tau; T_0) = 0.5 + 0.5 \cos \left( \frac{\pi \log_2 \left( \frac{\tau}{T_0} \right) }{4} \right), \]

Figure 2 shows equal probability contour of peak frequencies of \( A(\tau) \) for random signal input. The F0 detectors were allocated from 40 Hz to 600 Hz with a half octave spacing. Please note that the contour is frequency independent within the region of interest. In other words, it is possible to represent the probability distribution of \( A(\tau) \) as a function of peak value.

Figure 3 shows relation between peak level and the probability. This representation enables risk evaluation in periodicity.
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 because of 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 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 belongs to 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 speech signals.
5.1. 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 higher frequency range. Converting the time axis to \( \tau(t) \) using the instantaneous frequency of the fundamental component \( f_0(t) \) and target F0 \( f_{fix}^0 \) in Equation \( \tau(t) = \int_0^t f_{fix}^0 / f_0(\lambda)d\lambda \), the F0 of the signal converted onto the new time axis has constant value \( f_{fix}^0 \).

This F0 stabilization procedure eliminates the amplitude modulation of \( P_C(\omega) \) on the frequency axis mentioned in section 4. Therefore periodicity can be evaluated locally on the frequency axis irrespective to frequency position.

5.2. 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.

\[
h_N(\omega) = w_{0,N}(\omega) \exp(2\pi j \omega / \omega_0) ,\]

where a signal envelope function \( w_{0,N}(\omega) \) defines spectral resolution of aperiodicity calculation. In terms of time-bandwidth (TB) 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.

\[
w_{0,N}(\omega) = c_0(1 + \cos(\pi \omega / N \omega_0)) ,\]

where constant \( c_0 \) is used to normalize \( \int w_{0,N}(\omega)d\omega = 1 \). Using this quadrature signal, initial evaluation of periodicity is defined as follows.

\[
Q_C(\omega; T_0) = \int_{-\infty}^{\infty} h_N(\lambda; T_0)P_C(\omega - \lambda; T_0) d\lambda
\]

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

5.3. 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. A Japanese vowel sequence /iaue/ spoken by a male speaker sampled at 22050 Hz was used. The sample was also used in Figure 4.

Figure 5 shows \( Q_c \) map. It is observed that even in silence period there are positive values in \( Q_c \) reflecting statistical fluctuation due to randomness. It also is observed that in the voiced part it does not completely reach the value 0.56, which is for periodic pulse trains.

6. Conclusions

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.

7. Acknowledgements

This work is partly supported by CrestMuse project and Aids for Scientific research (A)19200017.

8. References