FIXED POINT ANALYSIS OF FREQUENCY TO INSTANTANEOUS FREQUENCY MAPPING FOR ACCURATE ESTIMATION OF F0 AND PERIODICITY

Hideki Kawahara[1], *Haruhiro Katayose*[2], *Alain de Cheveigné*[3] *and Roy D. Patterson*[4]

[1]Wakayama University/ATR/CREST, 930 Sakaedani Wakayama, Wakayama, 640-8510 Japan
[2]Wakayama University/LIST, Japan, [3]IRCAM/CNRS, Paris, France
and [4]CNBH, University of Cambridge, Cambridge, United Kingdom
kawahara@sys.wakayama-u.ac.jp

ABSTRACT

An accurate fundamental frequency (F0) estimation method for non-stationary, speech-like sounds is proposed based on the differential properties of the instantaneous frequencies of two sets of filter outputs. A specific type of fixed points of mapping from the filter center frequency to the output instantaneous frequency provides frequencies of the constituent sinusoidal components of the input signal. When the filter is made from an isometric Gabor function convoluted with a cardinal B-spline basis function, the differential properties at the fixed points provide practical estimates of the carrier-to-noise ratio of the corresponding components. These estimates are used to select the fundamental component and to integrate the F0 information distributed among the other harmonic components.

1. INTRODUCTION

Accurate and reliable fundamental frequency (F0) extraction is important in general sound manipulation applications and as a research tool in speech production and perception. A combination of the wavelet-based instantaneous frequency analysis and a carrier-to-noise ratio (C/N ratio) estimation provides an integrated F0 extraction method that is especially suitable for high-quality speech manipulation [4, 5] based on the channel VOCODER type architecture [3]. In such applications, a temporal resolution comparable with one pitch period and tracking ability free from the moire effects (see Figure 4) are required. These requirements are difficult to fulfill solely by short-term Fourier transform-based (STFT-based) methods or autocorrelation-based methods.

1.1. Outline of the method

The proposed method consists of two processing stages. First, band-pass filters that are equally spaced and have the same shape on the log frequency axis are used to extract fixed points of mapping from the filter center frequency to the instantaneous frequencies of the filter outputs. These fixed points are evaluated in terms of the estimated C/N ratio to select a fixed point that corresponds to F0. This initial estimate of F0 already has a reasonable accuracy, but it can be improved by a refinement procedure in the second stage.

In the second stage, a parabolic time axis warping that uses F0 information and a derivative of F0 is introduced prior to performing a F0 adaptive STFT. A fixed point analysis based on this time warping STFT provides fixed points that correspond to harmonic components. Then, instantaneous frequencies of the fixed points are integrated, using their C/N information to provide the F0 estimate with the minimum estimation error. The estimated C/N ratios of the harmonic components also provide information to control the periodicity of the source signal that is applicable to speech resynthesis.

2. FIXED POINT ANALYSIS

Instantaneous (angular) frequency $\omega(t)$ of a non-stationary time series $x(t)$ is defined using its Hilbert transform $H[x(t)]$:

$$\begin{aligned} \omega(t) &= \frac{d\phi(t)}{dt} \ . \\ \phi(t) &= \arg\left[x(t) + jH[x(t)]\right] \ , \end{aligned} \tag{1}$$

where j represents $\sqrt{-1}$ and $\phi(t)$ represents a phase component.

2.1. Filter design

To use the instantaneous frequency for an F0 estimation, it is necessary for the fundamental component to be separated and selected prior to the calculation. This is performed by a band-pass filter bank that consists of filters equally spaced along the log frequency axis and a specially designed impulse response and a selection mechanism.

Assume a filter impulse response $w_s(t,\lambda)$ is designed using a Gabor function $w(t,\lambda)$. This function is convoluted with a second order cardinal B-spline basis function $h(t,\lambda)$ that is tuned to a hypothesized F0. Also, assume that the Gabor function has an equivalent relative resolution in both the time and the frequency domains in relation to the fundamental period and the F0.

$$w_s(t,\lambda) = w(t,\lambda) * h(t,\lambda) \ , \tag{2}$$

$$w(t,\lambda) = e^{-\frac{\lambda^2 t^2}{4\pi\eta^2}} e^{j\lambda t} \ ,$$

$$h(t,\lambda) = \max\left\{0, 1 - \left|\frac{\lambda t}{2\pi\eta}\right|\right\} \ , \tag{3}$$

where "$*$" represents convolution and η represents a time-stretching factor. Convolving with the B-spline selectively suppresses interference from neighboring harmonic components if $\lambda = 2\pi F_0$. This is the case for the channel tuned to the fundamental component, which is selected by a process to be described later on. These filters are allocated uniformly on the log frequency axis. In other words, a continuous wavelet transform is calculated.

2.2. Fixed point extraction

The instantaneous frequency $\omega_c(t;\lambda)$ of each filter output is represented as a function of time and the filter center frequency λ. A set of fixed points $\Lambda(t)$ of this center frequency to the output instantaneous frequency map (F-IF map) is defined as follows.

$$\begin{aligned} \Lambda(t) = \{\lambda \mid \omega_c(t;\lambda) = \lambda, \omega_c(t;\lambda-\varepsilon) - (\lambda-\varepsilon) \\ > \omega_c(t;\lambda+\varepsilon) - (\lambda+\varepsilon)\} \ , \end{aligned} \tag{4}$$

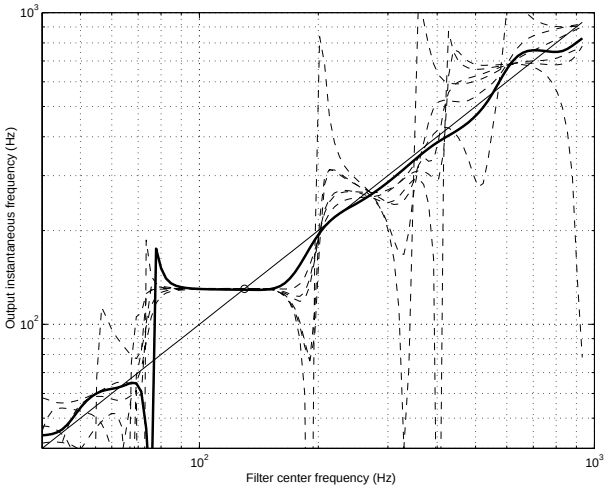

Figure 1. The filter center frequency to the output instantaneous frequency map. The thick solid line represents the mapping at 200 ms from the beginning of the Japanese vowel /a/ spoken by a male speaker. Broken lines represent mappings at different frames. The circle mark represents the fixed point corresponding to F0. η = 1.1 **was used.**

where ε represents an arbitrarily small positive constant.

Figure 1 illustrates the F-IF maps and a fixed point corresponding to F0. The map is calculated using a set of filters equally spaced along the log frequency axis. The spacing is 24 filters in an octave. Note that the F-IF map shows a stable plateau only around F0. This is because only the dominant sinusoidal component in a filter pass-band around F0 can be a fundamental component. Other fixed points corresponding to one of a harmonic component or a formant resonance frequency inevitably are affected by the neighboring (mainly lower) harmonic components within the same pass-band.

2.3. Carrier-to-noise (C/N) ratio estimation

The goodness of the plateau is objectively defined using differential parameters at each fixed point. Assume the following signal model holds around a fixed point that corresponds to a dominant carrier frequency $\omega_h(t)$. A small sinusoidal noise component with an amplitude $0 < \varepsilon \ll 1$ is then assumed to be added to the dominant sinusoid.

$$\begin{aligned} s(t) &= g(\lambda-\omega_h)e^{j\omega_h l} + \varepsilon g(\lambda-\omega_h+\delta)e^{j(\omega_h+\delta)l} \quad (5)\\ &= e^{j\omega_h l} g(\lambda-\omega_h)\left(1+\frac{\varepsilon g(\lambda-\omega_h+\delta)}{g(\lambda-\omega_h)}e^{j\delta l}\right), \end{aligned}$$

where $g(\lambda)$ represents the frequency response of a normalized filter which centers around λ = 0, and δ represents the frequency difference of the noise component from the dominant sinusoidal component.

Let $g(\lambda)$ have its maximum value 1 at λ = 0. Assume that the frequency weighting function $g(\lambda)$ is a smooth and continuous function, which does not have any singularity around ω = 0. Then the Taylor expansion of $g(\lambda)$ around 0 suggests that $g(\lambda) \simeq 1$, if $\omega \ll 1$. This yields the following approximation represented in an exponential form:

$$s(t) \simeq (1+\varepsilon g(\lambda-\omega_h+\delta)\cos\delta t)\, e^{j\omega_h l + j\varepsilon g(\lambda-\omega_h+\delta)\sin\delta l}. \quad (6)$$

The phase component $\phi(t)$ of the signal $s(t)$ is approximated as follows:

$$\phi(t) \simeq \omega_h t + \varepsilon g(\lambda-\omega_h+\delta)\sin\delta t\,. \quad (7)$$

This indicates that the noise component introduces a phase modulation. It is easily shown that the phase modulation due to the additional sinusoidal component is additive when it is small.

2.3.1. Differential property around a dominant component

The next step is to estimate the relative amplitude of the noise component in terms of the observable parameters. The derivative of the instantaneous frequency $\omega_c(t)$ in terms of the filter center frequency λ yields the following for $\varepsilon \ll 1$ and a smooth $g(\lambda)$:

$$\frac{\partial\omega_c(t,\lambda)}{\partial\lambda} \simeq \left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=\delta}\varepsilon\delta\cos\delta t\,, \quad (8)$$

which consists only of a component varying in the cosine phase.

Also, the time and frequency derivative yields the following:

$$\frac{\partial^2\omega_c(t,\lambda)}{\partial t\partial\lambda} \simeq -\left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=\delta}\varepsilon\delta^2\sin\delta t\,, \quad (9)$$

which consists only of a sine phase component.

2.3.2. Estimation of noise energy

For noise components with a uniform distribution and the relation $\sin^2 x + \cos^2 y = 1$, the following measure provides an approximate estimate of a C/N ratio as measured by the relative noise energy $\tilde{\sigma}^2(t)$:

$$\begin{aligned} \tilde{\sigma}^2(t) &= c_a\left(\frac{\partial\omega_c(t,\lambda)}{\partial\lambda}\right)^2 + c_b\left(\frac{\partial^2\omega_c(t,\lambda)}{\partial t\partial\lambda}\right)^2. \quad (10)\\ c_a &= \frac{1}{\int_{-\infty}^{\infty}\left(\delta\left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=\delta}\right)^2 d\delta}\,.\\ c_b &= \frac{1}{\int_{-\infty}^{\infty}\left(\delta^2\left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=\delta}\right)^2 d\delta}\,. \end{aligned}$$

This does not provide a good estimate in general, because the coefficients of the sinusoidal components in Equations 9 and 8 do not match exactly. However, an analysis that uses both the impulse response defined by Equation 2 and a temporal smoothing that uses the envelope of $w(t)$ effectively eliminates the artifacts from the second harmonic component and the interferences between noise components. Therefore, it can provide a practical estimate of the C/N ratio as $1/\bar{\sigma}$.

$$\bar{\sigma}^2(t,\lambda) = \int_{-T_w}^{T_w}|w(\tau,\lambda)|\tilde{\sigma}^2(t-\tau,\lambda)d\tau, \quad (11)$$

where $[-T_w, T_w]$ represents the effective support of $|w(\tau,\lambda)|$.

Figure 2 shows the test signal distribution of fixed points in the instantaneous frequency and noise energy plane. The signal is a mixture of a 200 Hz pulse train and an additive white noise (20 dB S/N). Note that the fixed points around 200 Hz correspond to the fundamental component, and that their ordinates are distributed around the value 0.1, as expected based on the 20 dB S/N ratio.

2.4. Fixed point selection

If the signal-to-noise ratio is not too low, the fixed point that corresponds to the F0 is the one that has the lowest C/N ratio. That is because of the filter design. A practical criterion for selecting an

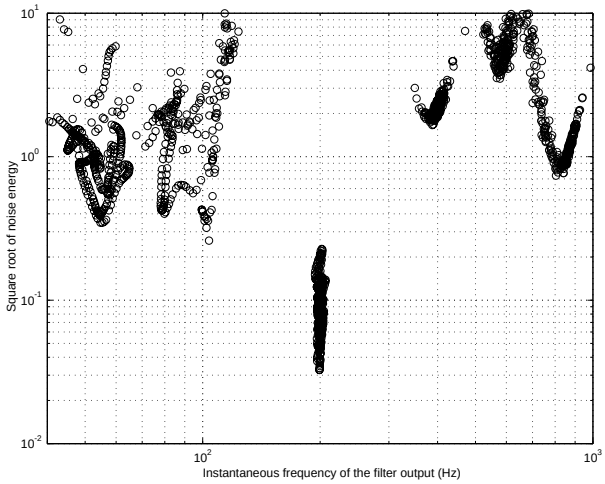

Figure 2. The instantaneous frequency and estimated noise energy distribution for a pulse train with 200 Hz F0 and 20 dB S/N.

F0 is to choose the fixed point that provides the maximum C/N ratio as the fundamental component when the maximum C/N ratio is 20 dB or higher. It is also easy to trace the closest fixed points forward and backward to extract the end points of the voiced portion.

Figure 3 shows an integrated display of extracted source information from a Japanese vowel sequence /aiueo/ spoken by a male speaker. Note that the fixed points corresponding to the F0 form a prominent smooth line in the top panel. Information on the total and higher frequency energy was also used in the V/UV decision making.

2.5. Performance of the first stage

A set of experiments using a speech database with simultaneously recorded EGG (electro-glottograph) data was conducted. The database was provided by Nick Campbell of ITL-ATR for prosody research. It consisted of 108 sentences spoken by a male speaker and 100 sentences spoken by a female speaker. Fixed points extracted from the EGG were used as the reference.

Only 712 frames (0.45%) of male speech out of 156,102 frames showed F0 errors greater than 20%, while 10,963 frames (7.0%) showed F0 errors greater than 5%. For the female speech, 181 frames (0.07%) out of 249,641 frames showed F0 errors greater than 20% and 2,577 frames (1.0%) showed F0 errors greater than 5%. In addition, 90% of frames were within 1.0% F0 errors in the female speech case. This performance for the female data is already competitive with or supersedes the conventional methods.

3. PERIODICITY ANALYSIS

The F0 estimate given in the previous section can be improved further by using F0 information distributed among other harmonic components. A harmonic component analysis also provides information about the aperiodicity at each harmonic frequency, which is useful for signal resynthesis.

3.1. Parabolic time warping

An adaptive STFT analysis using extracted F0 information is expected to enable separation of harmonic components. However, in a higher frequency region, the fixed point trajectories have discontinuities and irregularities. Sometimes trajectories in higher

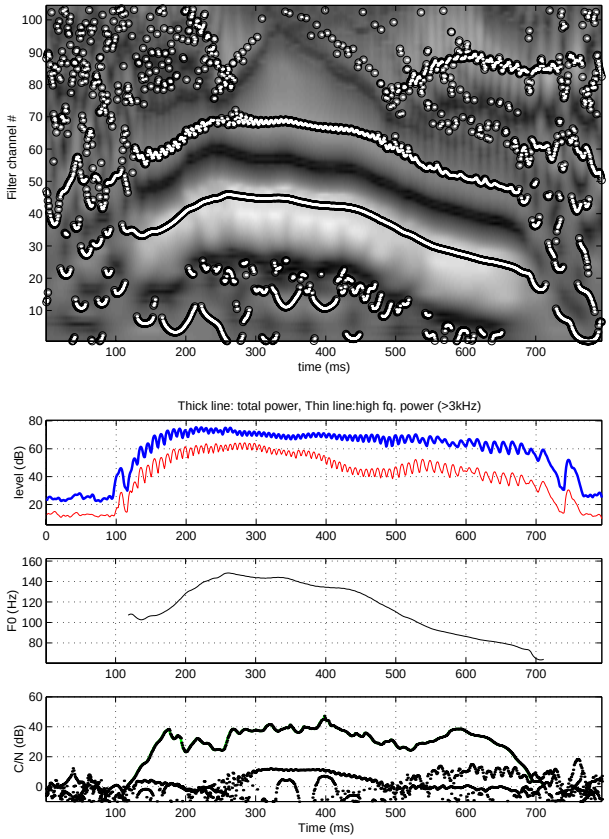

Figure 3. Extracted source information from a Japanese vowel sequence /aiueo/ spoken by a male speaker. The top panel represents fixed points extracted using a circle symbol with a white center dot. The overlayed image represents the C/N ratio. The lighter color indicates a higher C/N ratio. The middle panel shows the total energy (thick line) and the higher frequency (>3 kHz) energy (thin line). The next panel illustrates an extracted F0. The bottom panel shows the C/N ratio for each fixed point.

frequency regions move toward the opposite directions of the fundamental component trajectory. Figure 4 illustrates a typical example.

Figure 4 shows a time-frequency scatter plot of the fixed points extracted from a continuously pronounced Japanese vowel sequence /aiueo/ spoken by a male speaker. It is the same material shown in Figure 3. The first four harmonic components are proportional to the fundamental component. However, in the region from 450 ms to 500 ms and higher than 2 kHz, irregularities can be found. These are due to the moire effects described above.

These irregularities can be alleviated by introducing a nonlinear time warp to make the fundamental frequencies in the new temporal axis stay constant [1]. The nonlinear time warping function $u(t)$ is derived from the inverse of the phase function $\phi(t)$ of the fundamental component.

$$u(t) = c_0 \phi^{-1}(t)\,, \tag{12}$$

$$\phi(t) = \int_{\tau_0}^{t} \omega_0(\tau) d\tau \simeq \omega_0(t_c)t + \frac{1}{2}\frac{d\omega_0(t_c)}{dt}t^2\,. \tag{13}$$

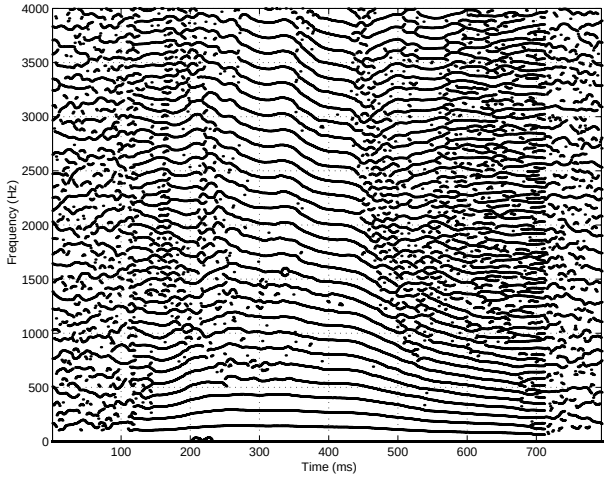

Figure 4. A time frequency plot of fixed points using a pitch-adaptive STFT without time warp. It was obtained in response to a Japanese vowel sequence /aiueo/ spoken by a male speaker. Note that the moire effects destroy higher harmonic structures when F0 changes rapidly.

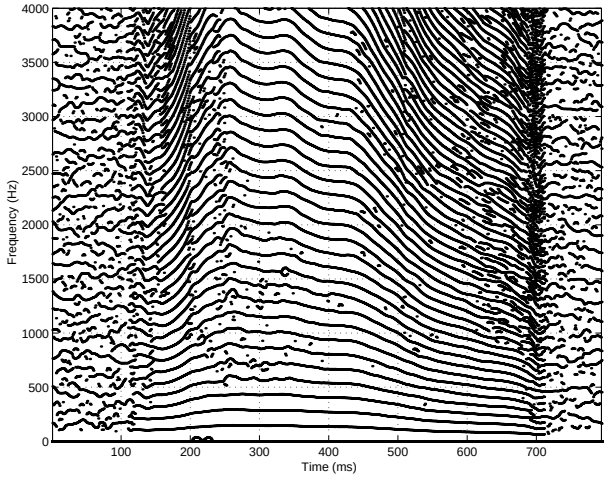

Figure 5. Same as Figure 4, with a parabolic time warping based on the F0 derivative. Note that the moire effects have vanished.

The last equation approximates the phase function when the derivative of $\omega_0(t)$ is approximately constant. This condition holds for a small segment when the F0 trajectory is smooth. In other words, a localized parabolic time warping approximates the original global nonlinear time warping function.

Figure 5 shows the time-frequency scatter plot of the fixed points. The fixed points are calculated using a parabolic time warping based on the F0 derivative information. Note that the moire effects found in the previous figure have vanished.

The C/N ratios for the extracted fixed points for both representations indicate that the deterioration of the C/N around the moire areas was an apparent effect. When the time warping was introduced, the C/N ratios were restored to normal levels. Therefore, a parabolic time warp based on F0 derivative allows C/N-based integration of F0 information from higher harmonics. The F0 information is distributed among different harmonic components. The C/N information of each harmonic component provides a variance estimate of the component frequency. Information from harmonics for which the C/N estimate is high enough can be integrated to minimize the error of the F0 estimate. A preliminary test indicated that this integration is especially useful for reducing F0 errors in male speech.

4. DISCUSSION

An F0 extraction based on the fixed points of the filter center frequency to the output instantaneous frequency map was first reported by [2] and reinvented by [1]. However, without the combination of a log linear filter bank, a special impulse response and a C/N ratio estimation introduced here, its usefulness was limited. For estimating a C/N ratio, the Kaiser energy separation operator was expected to provide a similar index [6]. However, it was found that the operator was numerically too sensitive to noise to be applied to fixed point analysis.

Based on preliminary observations using the proposed method, it is likely that F0 trajectories extracted from lower frequency components and from higher frequency components have systematic differences. A careful discussion about how to define various aspects of F0 is crucially important.

5. CONCLUSION

An instantaneous frequency-based F0 and a periodicity information extraction method was introduced. A combination of a special filter design and a C/N ratio estimation resulted in an accurate and reliable F0 estimation method. The C/N ratio was estimated with a fixed point analysis of a mapping from the filter center frequency to the output instantaneous frequency. Moreover, a parabolic time warping based on an F0 derivative allowed the aperiodicity of each harmonic component to be evaluated using C/N information, and provided a means to further refine the F0 estimate.

6. ACKNOWLEDGMENT

The authors appreciate the timely discussions with Dr. Masataka Goto of Electro Technical Laboratories and Dr. Parham Zolfaghari of CREST/ATR, among other colleagues.

REFERENCES

[1] Abe, T., Kobayashi, T. & Imai, S. (1997), The IF spectrogram: a new spectral representation, *in Proc. ASVA'97*, Tokyo, pp. 423-430.

[2] Charpentier, F. J. (1986), Pitch detection using the short-term phase spectrum , *in Proceedings of ICASSP'86*, pp. 113-116.

[3] Dudley, H. (1939), Remaking speech, *J. Acoust. Soc. Amer.* **11**(2), pp. 169-177.

[4] Kawahara, H. (1997), Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited, *in Proceedings of ICASSP'97*, Vol. 2, Muenich, pp. 1303-1306.

[5] Kawahara, H., Masuda-Katsuse, I. & de Cheveigné, A. (1999), Speech transformation using adaptive interpolation of time-frequency representation and all-pass filters , *Speech Communication* **27**(3-4), 187-207.

[6] Maragos, P., Kaiser, J. & Quatieri, T. F. (1993), Energy separation in signal modulations with application to speech analysis, *IEEE Trans. Signal Processing* **41**(10), 3024-3051.