AN INSTANTANEOUS-FREQUENCY-BASED PITCH EXTRACTION METHOD FOR HIGH-QUALITY SPEECH TRANSFORMATION: REVISED TEMPO IN THE STRAIGHT SUITE

Hideki Kawahara[1] *Alain de Cheveigné*[2] *Roy D. Patterson*[3]

[1]Wakayama University/ATR/CREST, Wakayama, Wakayama, Japan
[2]Paris 7 University/CNRS, Paris, France
[3]CNBH University of Cambridge, Cambridge, United Kingdom

ABSTRACT

A new source information extraction algorithm is proposed to provide a reliable source signal for a high-quality speech analysis, modification, and transformation system called STRAIGHT (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram). The proposed method makes use of instantaneous frequencies in harmonic components based on their reliability. A performance evaluation is conducted using a simultaneous EGG (Electroglottograph) recording as the reference signal. The error variance for F0 extraction using the proposed algorithm is shown to be about 1/3 that of the previous F0 extraction method used in STRAIGHT, although the previous algorithm is still competitive with conventional F0 extraction methods.

1. INTRODUCTION

STRAIGHT-suite [9, 8] is basically a VOCODER [6] with a collection of sophisticated algorithms to remove spectral interference caused by signal periodicity using a type of spline-based F0-adaptive time-frequency smoothing, (b) extract F0 information based on an instantaneous-frequency-related measure and (c) design excitation pulses using group delay manipulations [10]. Preliminary experiments have shown that our method provides high-quality synthetic speech that has equivalent naturalness to the original speech signal, when no parametric modification is introduced [11]. When there are no defects in the source information extraction, the transformed speech signals still sound natural, even with a large amount of parametric manipulations. At times, however, inaccurate source information has been found to cause deterioration of the re-synthesized speech.

Defects in source information extraction are categorized into three types; random variations in the extracted F0 (which reduce the accuracy and smoothness), difficulties in voiced/unvoiced decisions around phrase or sentence final parts due to their multi-band-excitation [7], and spectral secondary structures due to multiple excitations within a pitch period. The focus of this paper is on the reduction of random variations in extracted F0 information using cues from multiple harmonic components.

2. METHOD

Requirements There are several requirements for the source information to be used as a good base material for modification. **Resolution:** It is necessary for the F0 information to have a higher resolution than the usual limit determined by the sampling frequency. This requirement is crucial for high F0 speech like the speech of females. **Continuity:** The F0 trajectory has to be smooth and continuous. This condition is important for low F0 speech, where typical F0 extraction methods based on interval measurements between excitation epochs produce staircase F0 trajectories. **Graded value:** The voiced/unvoiced discrimination has to have a graded value in each different frequency band. This characteristic is desirable for applications like auditory morphing. **Graceful degradation:** It is desirable for the algorithm to deteriorate gracefully even when it fails to extract proper source information.

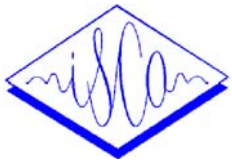
2.1. Instantaneous frequency and F0

F0 extraction methods based on interval measurements are not ideal to fulfill the above requirements, especially when dealing with real speech where F0 changes all the time. The instantaneous frequency of the fundamental component of the speech signal is a better alternative [1, 2, 3, 4]. The instantaneous frequency $\varphi(t)$ of a signal $s(t)$ is defined using its Hilbert transform $H[s(t)]$.

$$\varphi(t) = \frac{d \arg(s(t) + jH[s(t)])}{dt} \quad (1)$$

Using a band-pass filter that centers around F0 and excludes other components, this definition provides a continuous high-resolution F0 trajectory. In addition, using a band-pass filter that has an impulse response made from a quadrature signal replaces the Hilbert transform with simple filtering. However, it is somewhat contradictory to use the instantaneous frequency in such a manner, because F0 has to be known in advance to select the fundamental component. Accordingly, a measure called 'fundamentalness' associated with a sophisticated filter design is introduced to solve this apparent contradiction. The measure also provides graceful and graded behaviors.

Filter bank design A filter bank consisting of filters having the same shape on the log frequency axis is designed to produce a filter output signal that has minimum AM (amplitude modulation) and FM (frequency modulation) when the filter is centered on the fundamental component. Band-pass filters with a steep attenuation at the higher frequency side and a gradual attenuation at the lower frequency side provide the desired behavior. As shown in Figure 1, by using the described filter design, the filters centered around harmonic components other than the first one include multiple harmonic components within their pass-bands. This multiplicity of the harmonic components results



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

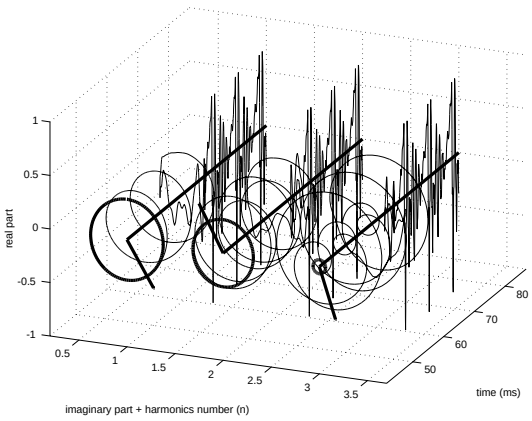


Figure 1: Illustration of the method for vowel 'a' pronounced by a male speaker (demonstration movie). Shown are channels centered on F0, 2F0 and 3F0. For each, the waveform input to the filter is plotted in perspective, followed by a polar representation of the complex output. The radius of the thick circle represents instantaneous amplitude, which is constant for the filter centered on F0 and pulsating for those centered on 2F0 and 3F0.

in a ''beating'' behavior, which modulates the instantaneous amplitude and the instantaneous frequency. The output of the filter centered around the first component does not show such variations due to the beating, as shown in Figure 1. The front half of the figure shows filter outputs and the rear half shows the input signal to each filter.

The conditions for the filter shape are not very strict. A differentiated Gabor function $w_{AG}(t;\eta)$ is one of the filters meeting the conditions.

$$w_{AG}(t;\eta) = w_g(t-1/4;\eta) - w_g(t+1/4;\eta) \quad (2)$$
$$w_g(t;\eta) = \frac{1}{\eta} e^{-\frac{\pi t^2}{\eta^2}} e^{-2\pi j t}$$

The original Gabor function $w_g(t;\eta)$ is designed to have a slightly finer resolution in frequency than in time, by selecting the parameter ($\eta > 1$).

Using scaled versions of this function $w_{AG}(t;\eta)$ as analyzing wavelets, the input signal $s(t)$ can be divided into a set of filtered complex signals $B(t;\tau_c)$. The frequency response of $w_{AG}(t;\eta)$ has a zero at $2f_c$ where $f_c = 1/\tau_c$ is the center frequency of the original Gabor function.

$$B(t;\tau_c) = |\tau_c|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t)\overline{w_{AG}\left(\frac{t-u}{\tau_c}\right)} du \quad (3)$$

Definition of 'Fundamentalness' A measure called 'fundamentalness' is defined to have the maximum value when the variations due to AM and FM are minimum. The following is used in the current implementation.

$$M(t;\tau_c) = -\log\left[\int_\Omega w\left(\frac{d|B(u)|}{du} - \mu_{AM}(u)\right)^2 du\right] - \log\left[\int_\Omega w\left(\frac{d^2\arg(B(u))}{du^2} - \mu_{FM}(u)\right)^2 du\right] + \log\left[\int_\Omega w|B(u)|^2 du\right] + 2\log\tau_c \quad (4)$$

$$\mu_{AM}(t) = \int_\Omega w(u-t;\tau_c)\left(\frac{d|B(u)|}{du}\right) du \quad (5)$$

$$\mu_{FM}(t) = \int_\Omega w(u-t;\tau_c)\left(\frac{d^2\arg(B(u))}{du^2}\right) du \quad (6)$$

$$w(t;\tau_c) = \frac{1}{\sqrt{2}\tau_c} e^{-\frac{\pi t^2}{2\eta^2}} \quad (7)$$

where the integration interval $\Omega = (t-T, t+T)$ is selected to cover the range where the weighting factor $w(u-t;\tau_c)$ is effectively non-zero (in Equation (4), the terms $w(u-t)$ are abbreviated as w for simplicity). The first two terms of Equation (4) measure AM and FM respectively, and the last term is a normalization factor that makes index $M(t;\tau_c)$ independent of the scale. Extracting $F0$ involves (a) finding the maximum index of $M(t;\tau_c)$ in terms of τ_c, and (b) calculating the average (or more specifically, interpolated) instantaneous frequency using the outputs of the channels neighboring τ_c. The method is called TEMPO (Time domain excitation Extraction based on a Minimum Perturbation Operator).

Implementation The current implementation of TEMPO uses 12 filters to cover one octave. The default coverage of a filter bank consisting of 52 filters is 40 Hz to 800 Hz. Each band-pass filter is implemented as an FIR filter using FFT-based convolution. Down-sampling of the input signal is employed to reduce the computational demand.

Evaluation and difficulties The proposed method was tested using a database prepared by Campbell [ITL-ATR] and modified by one of the authors [5]. Currently, this database consists of 208 sentences spoken by a male speaker and a female speaker. Each sentence has a speech signal part and a simultaneously recorded EGG (Electroglottograph) signal part. Both signals are analyzed using the proposed method. The F0 data extracted from the EGG signal is used as the reference.

The test results indicated that the gross error rate is less than 0.8% in total, where F0 discrepancies greater than 20% of the reference F0 are counted as gross errors. Moreover, more than 50% of the female data are within 0.3% of the EGG F0. This performance is competitive with current technical standards. However, unlike the usual F0 extraction methods, the F0 discrepancies are larger for the male speech [8]. Figure 2 shows a histogram of the F0 extraction accuracy for the male speaker. Here, F0 extraction was performed at a 1ms frame rate. The total number of frames tested was 155,939 for the male speech and 260,342 for the female speech. The standard deviation of the normalized F0 was 2.64% for the male speech and 1.09% for the female speech.

Similar difficulties were found in speech manipulation experiments using natural speech materials. It was observed that the initial and final transitions between voiced sounds and unvoiced

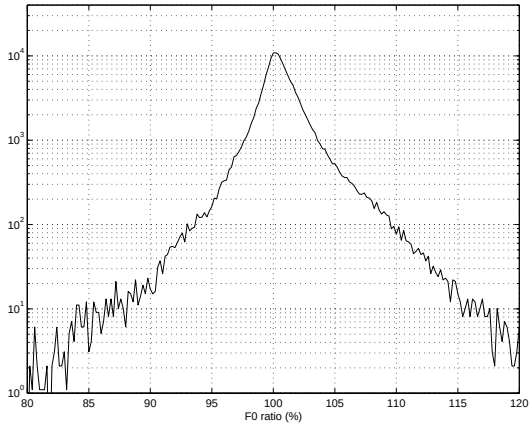


Figure 2: F0 extraction accuracy histogram for male speech. The horizontal axis represents the extracted F0 normalized by the reference F0 extracted from the EGG signal.

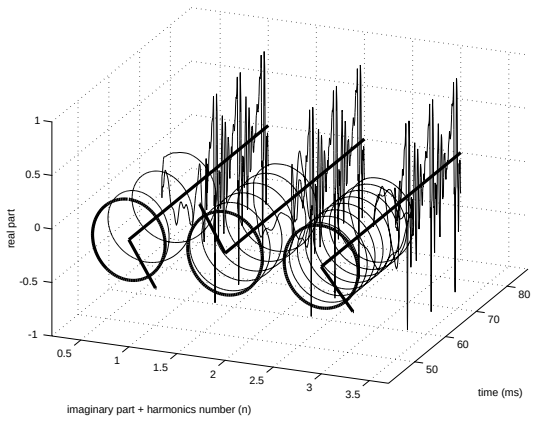


Figure 3: Illustration of the multi-harmonic scheme for vowel 'a' pronounced by a male speaker (demonstration movie). Shown are shifted bandpass filters centered on F0, 2F0 and 3F0. The input signal is multiplied by 1, $\exp(2j\pi F_0 t)$, and $\exp(4j\pi F_0 t)$ respectively. Note that the instantaneous amplitudes (radius of circle) of all filters are stable.

sounds tended to induce errors in the estimation process, especially for male speakers with very low F0s. In addition, the accuracy tended to be deteriorated for such speakers because of the relatively low signal to noise ratio in the fundamental component region where low-frequency noise, such as air conditioning noise, resides.

Such performance deterioration for male speech may reflect a fundamental limitation of our previous method. The limitation is that the method only relies on the first harmonic component, that is most susceptible to low-frequency environmental noise. The extracted F0 error is directly dependent on the signal to noise ratio of the corresponding filter output, which makes the method behave poorly when the fundamental component is weak or missing.

2.2. Use of multiple sources

Definition of 'n-th-ness' The difficulties mentioned in the previous section can be reduced by adding other sources of F0 information, i.e., distributed in other harmonic components. By using a modulation technique, it is possible to extend the definition of 'fundamentalness' to represent '2nd-harmonic-ness', 'third-harmonic-ness', and so on, based on a minimum perturbation criterion. A modulated filter output for the *n-th* harmonic component, $B_n(t;\tau_c)$, is defined as follows.

$$B_n(t;\tau_c) = |\tau_c|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t)\overline{w_{AGm}\left(\frac{t-u}{\tau_c};n\right)}du \quad (8)$$

$$w_{AGm}(t;n) = w_{AG}(t)e^{-2\pi j(n-1)t}$$

This modulation effectively replaces a constituent filter having a center frequency f_c, with a filter having n f_c for the *n-th* harmonic component. Note that the envelope is unchanged. Replacing $B(t;\tau_c)$ by $B_n(t;\tau_c)$ in the definition of 'fundamentalness' $M(t;\tau_c)$, yields the definition of 'n-th harmonic-ness' $M_n(t;\tau_c)$. Therefore, the 'n-th harmonic-ness' is designed to have the maximum value when the corresponding filter output has the minimum AM and FM for the modulated filter converted by an amount of n times the filter center frequency.

Figure 3 shows behaviors of modulated filter outputs for $n = 1, 2,$ and 3 with the same input signal as in Figure 1. Note that the *n-th* modulated filter outputs show a stable behavior similar to the filter output corresponding to the fundamental component.

Composite measure A new measure $M_s(t;\tau_c,n)$ is defined as a composite measure of these elementary measures. The variations of the instantaneous frequency and instantaneous amplitude determine the effective band-width of the signal. Since these variations result in F0 extraction errors, it is reasonable to combine elementary F0 information so as to minimize the errors in the final F0 estimation. Using variance measure $V_n(t;\tau_c)$ and elementary F0 estimation $f_n(t;\tau_c)$ at *n-th* step, variance estimation $V^{(n)}(t;\tau_c)$ and the minimum variance estimation of $f_0^{(n)}(t;\tau_c)$ are calculated recursively.

$$f_0^{(n)}(t;\tau_c) = \frac{V^{(n-1)}(t;\tau_c)f_n(t;\tau_c) + V_n(t;\tau_c)f_0^{(n-1)}(t;\tau_c)}{V_n(t;\tau_c) + V^{(n-1)}(t;\tau_c)}$$

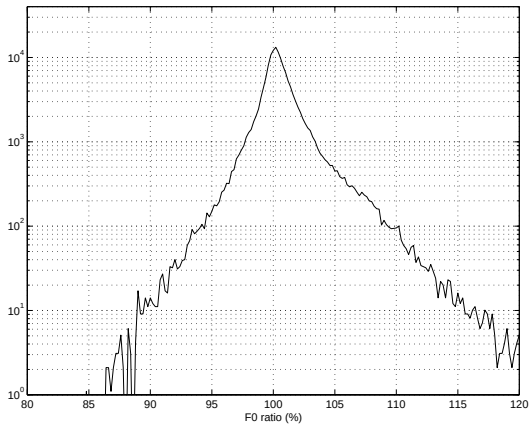
$$V^{(n)}(t;\tau_c) = \frac{V_n(t;\tau_c)V^{(n-1)}(t;\tau_c)}{V_n(t;\tau_c) + V^{(n-1)}(t;\tau_c)} \quad (9)$$

where $V_n(t;\tau_c)$ is calculated from total AM and FM. $M_s(t;\tau_c,n)$ is defined using this $V^{(n)}(t;\tau_c)$ with a heuristic compensation factor. The final F0 estimate is calculated by selecting the best filter using $M_s(t;\tau_c,n)$ and interpolating neighboring $f_0^{(n)}(t;\tau_c)$ values.

2.3. Evaluation

The same database was used to evaluate the new method. The reference F0 was extracted from the associated EGG recording; only the reliable portions were used in the evaluation. In the current evaluation, $n = 3$ was selected. The target F0 was extracted from a speech signal with $n = 1, 3$. The F0 with $n = 1$ was calculated for the purpose of comparison with our previous evaluation results.

(a)



(b)

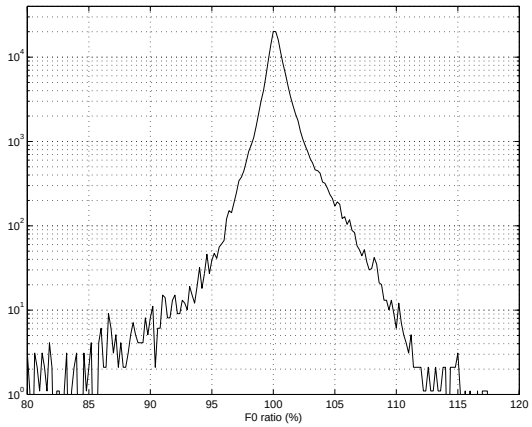


Figure 4: F0 extraction accuracy histogram for male speech based on the proposed method. (a) n = 1 and (b) n = 3. The horizontal axis represents the extracted F0 normalized by the reference F0 extracted from the EGG signal with n = 3.

Figures 4 shows F0 extraction accuracy histograms. The control condition n = 1 is slightly better than the previous implementation (Figure 2). This may be the result of the new reference F0 data, which is more reliable than the previous reference F0 data. This result also depends on the selection of the portions to be evaluated (154,364 frames were evaluated for the male data). The F0 data by the new method yields a more concentrated distribution. The standard deviations of the normalized F0s for n = 1 and n = 3 are 2.37% and 1.42%, respectively. In other words, the error variance for F0 extraction with n = 3 is 1/3 that with n = 1. It is also 1/4 that of the previous result shown in Figure 2.

3. DISCUSSION

The original implementation of TEMPO was computationally very demanding. The improved method proposed here offers an excellent F0 extraction performance, at a computational cost even greater than the original TEMPO method. But, by introducing appropriate modifications, the lag-window method of F0 extraction [12] may be shown to approximate it with a less-demanding implementation.

4. CONCLUSION

A highly accurate F0 extraction method that is specially designed for the high-quality speech analysis-modification-synthesis system STRAIGHT is presented. It is demonstrated that the method provides F0 estimation with a standard deviation of 1% from the reference F0 extracted using EGG. In other words, the error variance is reduced to 1/3 that obtained by our previous F0 extraction algorithm TEMPO, although TEMPO is still competitive with existing methods. The proposed method is computationally very demanding. It has been suggested that there is still a lot of room for improvement, e.g., introducing effective approximations, for implementation of this method. Manipulated speech examples and demonstration movies can be found at the following URL.

http://www.sys.wakayama-u.ac.jp/˜kawahara/straight/

5. REFERENCES

1. T. Abe, T. Kobayashi, and S. Imai. Harmonics estimation based on instantaneous frequency and its application to pitch determination. *IEICE Trans. Information and Systems*, E78-D(9):1188--1194, 1995.
2. T. Abe, T. Kobayashi, and S. Imai. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In *Proc. ICSLP 96*, pages 1277--1280, Philadelphia, 1996.
3. Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal -- part 1: Fundamentals. *Proc. of IEEE*, 80(4):520--538, 1992.
4. Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal -- part 2: algorithms and applications. *Proc. of IEEE*, 80(4):550--568, 1992.
5. Alain de Cheveigné. Speech fundamental frequency estimation. Technical Report TR-H-195, ATR-HIP, 1996.
6. H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, 11(2):169--177, 1939.
7. Daniel W. Griffin and Jae S. Lim. Multiband excitation vocoder. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(8):1223--1235, 1988.
8. Hideki Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proc. ICASSP'97*, volume 2, pages 1303--1306, Münich, 1997.
9. Hideki Kawahara and Ikuyo Masuda. Speech representation and transformation based on adaptive time-frequency interpolation. *Technical Report of IEICE*, EA96-28:9--16, 1996 (in Japanese).
10. Hideki Kawahara, Minoru Tsuzaki, and Roy D. Patterson. A method to shape a class of all-pass filters and their perceptual correlates. *Tech. Com. Psycho. Physio., The Acoust. Soc. Jpn.*, H-96-79:1--8, 1996 (in Japanese).
11. Hideki Kawahara and Reiko Akahane-Yamada. Perceptual effects of spectral envelope and F0 manipulations using the STRAIGHT method. In *Proceedings of the ICA/ASA'98 meeting*, 1aSC27, Seattle, 1998.
12. S. Sagayama and S. Furui. ''Pitch extraction using the lag window method,'' Proc. IECE Japan **1235-5**, 263, 1978 (in Japanese).