ROBUST FUNDAMENTAL FREQUENCY ESTIMATION
USING INSTANTANEOUS FREQUENCIES OF HARMONIC COMPONENTS

Yoshinori ATAKE 1  Toshio IRINO2,4  Hideki KAWAHARA3,4,2  Jinlin LU3
Satoshi NAKAMURA 1  Kiyohiro SHIKANO 1

1 Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma-shi, Nara, 630-0101 Japan
2 ATR Human Information Processing Laboratory
3 Faculty of Systems Engineering, Wakayama University
4 CREST

ABSTRACT

This paper proposes a noise-tolerant method for fundamental frequency (F0) extraction. This method includes several new ideas, including the estimation of the instantaneous frequencies of the higher harmonic components, and the design of an adaptive weighting function based on a bandwidth equation that combines the F0 information in the harmonic components. To evaluate the proposed method, we constructed a relatively large database of simultaneous recordings of speech waveforms and EGG (Electro Glottal Graphy). The database consists of 30 sentences pronounced by 14 male and 14 female normal subjects, i.e., 840 sentences in total. The duration of the sound is about 35 minutes including about 20 minutes of voicing. The experiments were performed with additive noise for four pitch extraction methods, i.e., the proposed method, the original TEMPO, an improved cepstrum method, and a common F0 extraction program in ESPS. The results were as follows: 1) the proposed method is always better than any of the other methods when the SNR is greater than about 2 dB; 2) for high SNR values (> 15 dB), the correct rates of the proposed method and the original TEMPO are about 95% and much better than the improved cepstrum method (92%); and 3) all of the methods degrade to less than 62% when the SNR is 0 dB. As a result, the proposed method improves the performance for low SNR values and also maintains high accuracy inherent from the original TEMPO for high SNR values.

1. INTRODUCTION

STRAIGHT [1] has been developed for high-quality sound manipulation as a sophisticated version of VOCODER [2]. The quality of the resynthesized sound is much more natural than that of conventional VOCODER-based methods (e.g., LPC vocoder). It enables the duration of speech sounds to be compressed and expanded more than 600% without undesired distortion. It has been reported, however, that the sound quality is largely degraded when the system is used in noisy environments [3]. The reason is because the “TEMPO” algorithm for fundamental frequency (F0) estimation does not maintain a high accuracy for low SNR values; the resulting error in the F0 estimation largely affects the accuracy of the time-frequency representation produced by the pitch-adaptive analysis.

This paper presents a new noise-tolerant method for F0 estimation from several harmonic components whereas the conventional TEMPO only uses a fundamental component. There are a lot of methods for fundamental frequency estimation [4] including the use of harmonic components [5, 6]. However, there are still arguments on a precise method to combine the F0 information from the harmonic components in noisy environments. We believe it is necessary to introduce a sort of reliability measure for each harmonic component because noise degrades the estimation of the harmonic frequency non-uniformly.

The new method has several stages: 1) rough F0 estimation using a revised TEMPO algorithm; 2) the design of an adaptive window for the STFT; 3) the creation of a map between the filter frequency and the instantaneous frequency; 4) the selection of fixed points as estimated instantaneous frequencies of harmonic components [7]; and 5) the utilization of a weighting function based on a bandpass equation [8] that combines the F0 information to estimate the final F0 value. Moreover, since there is a desire to make STRAIGHT a real-time system, it is necessary to exclude post-processing approaches such as DP-matching [6] for frequency smoothing.

2. METHOD

2.1. Frequency mapping

The initial part of the method is based on harmonic component estimation using the instantaneous frequency [5, 6, 7]. A Short-Term Fourier Transform (STFT) is applied to the input signal. The STFT can be interpreted as a bank of filters with equal spacings and equal bandwidths. The center frequency f_k of the k-th filter is readily calculated as (k − 1)f_s/N where f_s is the sampling frequency and N is the FFT length. It is possible to calculate the instantaneous frequency f_k(t) at the output of the k-th filter D_k(t) at time t using

\[ f_k(t) = \frac{1}{2\pi} \frac{\partial \arg(D_k(t))}{\partial t}. \]

(1)

As a result, two frequency components f_k and f_k(t) are obtained for each channel. Figure 1 shows the relationship (solid line) between the center frequency and the instantaneous frequency when analyzing a part of a male vowel ‘a’ as an example. The thin dashed line shows the mapping when f_k = f_k for all k. The solid line shows the staircase and crosses the dashed line at about the middle of the flat
as indicated by the circles. This is referred to as a stable “fixed-point” in this paper. The characteristics are produced as strong frequency components driving the filters in the flat range and producing the outputs concentrated near the particular instantaneous frequency. The strong components are most likely the fundamental component and its harmonics, at least in clean speech. Accordingly, the circles are good estimates for the frequencies of the fundamental and harmonic components.

2.2. Temporal window design

It is known that the accuracy for harmonic-component estimation is largely affected by the window size, particularly when the F0 values changes rapidly [9]. It is, therefore, essential to design a proper window for the STFT for any speech signal. With this in mind, we designed a window adaptive to the fundamental frequency. Although this might seem like a circular argument, it is not necessary to get the precise F0 value for the window design. In a preliminary simulation, we confirmed that the window works well even when errors are more than about 50%, i.e., obtainable in noisy speech data by the conventional TEMPO [1]. We, therefore, used a combination of the conventional TEMPO and a criterion based on a bandwidth equation [8] to estimate a rough F0 value [10] for each moment corresponding to the frame shift of the STFT.

2.2.1. Procedure

Initially, the speech signal is analyzed by the wavelet transform with a Gabor function having the bandwidth equal to 70% of the center frequency and a frequency range between 50 and 400 Hz. Then, we select a wavelet channel that maximally contains the fundamental component as in the original TEMPO [1]. The channel is specified as the instant of the minimum value of the bandwidth equation [8] in Eq. 3 (described shortly). At the selected channel, the fundamental frequency $f_{0}(t)$ is estimated as the instantaneous frequency of the channel output using Eq. 1.

The window function $w(t)$ is defined as a function of the fundamental frequency $f_{0}(t)$ as

$$w(t) = g(t) * h(t)$$

$$= \exp \left( -\pi \left( f_{0}(t) \right)^{2} \right) * \max \left( 0, 1 - \left| f_{0}(t) \right| \right)$$

i.e., a convolution of Gauss function $g(t)$ and cardinal B-spline function $h(t)$.

2.2.2. Channel selection based on the bandwidth

The channel selection is described in more detail here. When the signal is denoted as $A(t)e^{j\varphi(t)}$, the bandwidth equation is

$$B^{2}(t) = \frac{<A(t)A^{*}(t) >}{A^{2}(t)} \int \left| \varphi(t) - <\varphi> \right|^{2} A^{2}(t)dt$$

where $B$ is the bandwidth and

$$<\varphi> = \int \varphi(t)A^{2}(t)dt.$$  

The value of the first term increases as the fluctuation of amplitude $A(t)$ increases; the value of the second term increases as the fluctuation of angular frequency $\varphi(t)$ increases.

Bandwidth $B$ is calculated with the output from each channel of the gabor network for each frame cycle. For example, when the center frequency of the filter is equal to the frequency of the 3rd harmonic component, the lower harmonics components (i.e., 2nd and 1st) pass through the gabor filter since the bandwidth is sufficiently wide (70%) and the lower filter skirt is shallower than the upper one on the log-frequency scale. Then, the output of the filter fluctuates because of the combination of several harmonic components. It is the case for any harmonic component. In contrast, when the center frequency is equal to the fundamental frequency, the output mostly contains the fundamental component since there is no lower harmonic component. Consequently, the output does not fluctuate much and the bandwidth value is minimized in that channel. So, the channel selection with the minimum bandwidth is almost equivalent to the selection with “fundamentality” [1]. The wavelet transform is advantageous for this purpose.

2.3. Weighting of harmonic components

As described in the previous sections, the frequencies of the harmonic components are derived with the fixed-point method using an adaptive window3. When the SNR is low, the estimated frequencies of stable fixed-points are not necessarily reliable. Therefore, a weighting function based on a stability measure is necessary to combine the information from the harmonic components. Since the bandwidth in Eq. 3 evaluates the fluctuation of the signal, it must be also a good measure for the stability and reliability of the fixed-point. Accordingly, we used the following equation to calculate the weighting function to get the final result of the fundamental frequency.

$$f_{0}(t) = \sum_{i=1}^{K} w_{k} f_{k}(t)$$

$$w_{k} = \frac{\log_{2} \pi}{\sum_{k=1}^{K} \log_{2} \pi_{k}}$$

where $f_{k}$, $B_{k}$, and $n_{k}$ are the instantaneous frequency from Eq. 1, the bandwidth from Eq. 3, and the number of harmonics for the $k$-th stable fixed-point. The number of summations $K$ was determined so that the minimum value of $w_{k}$ would exceed 10% of the maximum value of $w_{k}$. Unreliable values were eliminated by this thresholding.

Figure 1. Relationship between center frequency $f_{c}$ and instantaneous frequency $f_{in}$ (solid line) at the output of the STFT of a male speech ‘a’. The circles show the frequencies for stable fixed-points.
In summary, Fig. 2 shows the flow chart of the procedure described above. For every frame shift (e.g., 1 ms), the fundamental frequency is derived with this procedure.

3. EXPERIMENTS

This section describes the evaluation of the proposed method in comparison with several conventional methods: the original TEMPO [1], an improved cepstral method [11], and the function get_F0 in the commonly-used ESPS (ver. 5.3.1). The original TEMPO and the function get_F0 were used with their default parameter settings. For the cepstral analysis, the harmonic components above 1.5 kHz were removed since they are generally irregular. This is known to result in an improvement of the estimation performance [11]. The window length was 1024 points and the sampling rate was 16 kHz.

3.1. Database of speech and EGG

It is essential to evaluate the estimation performance objectively using a large database. We therefore constructed a relatively large database of simultaneous recordings of speech sounds and Electro Glotto Graph (EGG). In the database, 30 short Japanese sentences were pronounced by 14 male and 14 female normal subjects between 22 and 36 years old [9]. The overall voiced duration was about 21 minutes within a total duration of about 30 minutes for 840 sentences. Since the frame shift in this experiment was 1 ms and the evaluation was performed in the voiced duration, the overall frame number used in the experiments was about 1.3 million.

3.2. Evaluation

It is virtually impossible to obtain the “true” value of the fundamental frequency from either speech data or EGG by using any known methods. Here, since robustness to noise was under consideration, we assumed the fundamental frequency derived from noise-free EGG data would be the most reliable. The correct rate for each method was calculated independently with the following procedure.

1. The fundamental frequency \( f_{0} \) is estimated by one of the methods from the original (or noise-free) EGG data for each frame. This is assumed to be the “true” frequency.
2. The fundamental frequency \( f_{0} \) is estimated by the same method from the speech signals after noise is added in accordance with the desired SNR. The noise is either white noise or pink noise.

3. When the absolute error between \( f_{0} \) and \( f_{0} \) is within 5%, the frame is counted as the “correct” one. The correct rate is calculated as the ratio of the number of the correct frames to all frames.

3.3. Results

3.3.1. White noise

Figure 3 shows overall correct rates obtained from various methods as a function of the SNR between speech signals and additive white noise. When no noise is added to the speech signals, the correct rates by the proposed method and the original TEMPO [1] are about the same at 95%. This value is higher than those for the cepstral method (92%) and the ESPS (80%). When the SNR is between 3 dB and 10 dB, the correct rate for the proposed method is the highest. The method maintains 80% even when the SNR is 3 dB. When the SNR is 0 dB, the cepstral method becomes the best. The rate is, however, only about 62%, which is useless in realistic applications. None of the methods work well when the SNR is 0 dB.

Figure 4 shows overall correct rates obtained from various methods as a function of the SNR between speech signals and additive pink noise. When no noise is added to the speech signals, the correct rates by the proposed method and the original TEMPO [1] are about the same at 95%. This value is higher than those for the cepstral method (92%) and the ESPS (80%). When the SNR is between 3 dB and 10 dB, the correct rate for the proposed method is the highest. The method maintains 80% even when the SNR is 3 dB. When the SNR is 0 dB, the cepstral method becomes the best. The rate is, however, only about 62%, which is useless in realistic applications. None of the methods work well when the SNR is 0 dB.
3.3.2. Pink noise

Figure 4 shows overall correct rates obtained from various methods as a function of the SNR between speech signals and additive pink noise. The proposed method becomes the best when the SNR is between 3 dB and 15 dB. Since pink noise concentrates energy at low frequencies, the fundamental component used by the original TEMPO comes to be more degraded than the harmonic components used by the proposed method.

3.3.3. Performance at an SNR of 5 dB

Figure 5 shows the correct rates for all of the above methods in both white and pink noise at 5 dB. The proposed method is better than any of the other methods. The differences between the proposed method and the original TEMPO for the white and pink noise is 5.5 and 7.6 percentage points. For all of the methods, the performance under the white noise condition is always better than under the pink noise condition. The proposed method is also affected by error in the initial rough estimate of the fundamental frequency.

4. SUMMARY

This paper proposes a noise-tolerant method for fundamental frequency estimation to improve the performance of STRAIGHT under noisy conditions. The method uses harmonic components to estimate the fundamental frequency. A reasonable weighting function is derived on the basis of Cohen’s bandwidth equation to combine the F0 information. A relatively large database containing simultaneous recordings of speech sounds and EGG is produced for the performance evaluation. The proposed method is compared to the original TEMPO, an improved cepstral method, and a common F0 extraction program in ESPS. The results show that 1) the proposed method is always better than any of the other methods when the SNR is greater than about 2 dB; 2) for high SNR values (> 15 dB), the correct rates of the proposed method and the original TEMPO are about 95% and much better than the improved cepstral method (92%) and ESPS function (80%); 3) all of the methods degrade to less than 62% when the SNR is 0 dB. As a result, the proposed method improves the performance for low SNR values and maintains a high accuracy inherent from the original TEMPO.

Acknowledgments: The authors wish to thank N. Masaki of ATR Human Information Research Labs. and Y. Tohkura of NTT Communication Science Labs. for their advice in constructing the EGG database. This work was partially supported by CREST.

REFERENCES


