A Fundamental Frequency Estimation Method for Noisy Speech Based on Instantaneous Amplitude and Frequency

Yuichi Ishimoto, Masashi Unoki and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
I-1 Asahidai, Tatsunokuchi, Nomigun, Ishikawa 923-1292 Japan

Centre for the Neural Basis of Hearing, Dept. of Physiology, University of Cambridge
Downing Street, Cambridge, CB2 3EG United Kingdom

{y-ishi, akagi}@jaist.ac.jp, masashi.unoki@mrc-cbu.cam.ac.uk

Abstract

This paper proposes a robust and accurate F0 estimation method for noisy speech. This method uses two different principles: (1) an F0 estimation based on periodicity and harmonicity of instantaneous amplitude for a robust estimation in noisy environments, and (2) an F0 estimation based on stability of instantaneous frequency as an accurate estimation method. The proposed method also uses a comb filter with controllable pass-bands to combine the two estimation methods. Simulations were carried out to estimate F0s from real speech in noisy environments and to compare the proposed method with other methods. The results showed that this method can not only estimate F0s for clean speech with similar accuracy as the method using only instantaneous frequency but also robustly estimate F0s from noisy speech in comparison with the other methods such as the cepstrum method.

1. Introduction

Extraction of the fundamental frequency (F0) of target speech is an important problem not only in speech analysis/synthesis but also in various speech signal processings such as speech segregation. For example, in speech analysis/synthesis, F0 is a factor controlling the pitch of speech and extraction of accurate F0s from speech is necessary so that the synthesized speech has naturalness. In speech segregation, F0 is a significant factor characterizing differences between sounds and F0 can be used as a cue for segregation of concurrent speech [1, 2]. For application of speech signal processing, especially speech segregation, in real environments, an accurate extraction of F0s from noisy speech is required. However, it is difficult to extract accurately F0s of target speech in noisy environments because noises distort the fundamental component of target speech.

Various F0 estimation methods have been proposed, but the most of these methods have the drawbacks for estimating accurate F0s of target speech in noisy environments. Kawahara et al. proposed an F0 estimation method, TEMPO2, based on stability of instantaneous frequencies [3], in order to construct speech analysis/synthesis (VOCODER). This method can estimate F0s for clean speech accurately, but it has difficulties in noisy environments, especially those below 10 dB SNR. Another method was proposed by Unoki and Akagi using instantaneous amplitude comb filtering in order to construct a sound segregation model [1]. This method can estimate F0s of connected vowels in noisy environments. However, the estimated F0s are not accurate enough. Thus, the existing methods cannot satisfy being both accurate and robust in noisy environments.

This paper proposes a robust and accurate method for estimating F0s even in noisy environments. This method consists of two different type estimation methods, a robust F0 estimation method based on instantaneous amplitude and an accurate method based on instantaneous frequency. The proposed method also uses a comb filter with controllable pass-bands to combine the two methods.

2. Algorithm

2.1. Overview

Figure 1 shows a flow chart for the proposed method. This method first makes rough estimation of the F0s from noisy speech using instantaneous amplitude as robust information corresponding F0s. The F0 estimation is based on periodicity and harmonicity of instantaneous amplitude (PHIA). In PHIA, prob-
abilities of F0 are calculated from periodicity and harmonicity, then they are integrated by the Dempster’s rule of combination. Next, noise reduction is done using the comb filter with controllable pass-bands. Its center frequencies are calculated from the roughly estimated F0s. The pass-band widths are controlled not to reduce the harmonic components of speech. Before reducing the noise, time warping of the noisy environment speech wave is performed to fix the F0s, so that this can decrease errors in the noise reduction. Then, F0 estimation using instantaneous frequency is applied to the noise-reduced speech wave. Thus, accurate F0s can be obtained from the noisy speech.

In the following sections, F0 estimation based on periodicity and harmonicity of instantaneous amplitude, noise reduction using the comb filter with controllable pass-bands and F0 estimation based on instantaneous frequencies are explained.

2.2. F0 estimation based on Periodicity and Harmonicity of Instantaneous Amplitude (PHIA)

The first F0 estimation of the proposed method needs robustness in noisy environments. The F0 estimation method using instantaneous amplitude comb filtering [1] is capable of estimating F0s for connected vowels, even if the signal-to-noise ratio (SNR) of noisy speech is 5 dB. However it sometimes estimates half or double of F0s for sentences. This is because it uses only harmonicity of instantaneous amplitude. To get robustness of F0 estimation for sentences, the proposed method uses not only harmonicity but also periodicity of instantaneous amplitude. It calculates each probability from periodicity and harmonicity and estimates reliable F0s in noisy speech.

Figure 2 illustrates the F0 estimation based on periodicity and harmonicity of instantaneous amplitude (PHIA). It is processed as follows. A speech signal is analyzed by constant Q filterbank and constant bandwidth filterbank. Periodicity is represented in the high frequency region of instantaneous amplitude by using constant Q filterbank and harmonicity is represented clearly in the low frequency region by using constant bandwidth filterbank. In this paper, the filterbanks are constant Q gammatone filterbank and constant bandwidth gammatone filterbank. The constant Q gammatone filterbank is constructed with 256 channels and their center frequencies are from 2 kHz to 6 kHz. The constant bandwidth gammatone filterbank is constructed with 400 channels and their center frequencies are from 60 Hz to 2 kHz. Instantaneous amplitude by the constant bandwidth filterbank can be implemented by FFT instead of the filterbank.

This method calculates each probability from periodicity and harmonicity. For the instantaneous amplitude using constant Q filterbank, some candidates of F0s are extracted using autocorrelation in time domain for one channel of the filterbank. Similarly, for the instantaneous amplitude using constant bandwidth filterbank, some candidates of F0s are extracted using autocorrelation in frequency domain by changing the lag window length of autocorrelation. Each histogram of candidates is considered as probabilities of F0s from periodicity and harmonicity.

The probabilities are integrated by Dempster’s rule of combination. The Dempster’s rule of combination is

$$m(A_k) = \sum_{A_1 \cap A_2 = A_k} m_1(A_1) m_2(A_2) / \left( 1 - \sum_{A_1 \cap A_2 = \emptyset} m_1(A_1) m_2(A_2) \right) \quad (1)$$

where $m_1, m_2$ are basic probability function and $A_{ij}, A_j (i, j = 1, 2, 3, \ldots)$ are focal element [4]. Considered that each probability from periodicity and harmonicity is basic probability function and frequency (bin of the histogram) is focal element, the integrated probability is obtained by this rule. The frequency with the highest probability is the estimated F0s. Thus, PHIA is used for the first F0 estimation of the proposed method.

The probability of F0s is used as a coefficient of bandwidth of the comb filter in next stage. Figure 3 shows an example of the estimated F0s and the probabilities by PHIA. In voiced section, the probabilities are high. In unvoiced or noisy section, they are low.

2.3. Noise reduction using the comb filter with controllable pass-bands

The proposed method needs a comb filter that can decrease influence of F0 errors at the first F0 estimation and can reduce noises as much as possible. Therefore, it uses the comb filter with controllable pass-bands as follows.
2.3.1. Formulation

Assume that the target signal $s(t)$ is harmonic complex tone and $n(t)$ is noise. Thus, the observed signal is

$$x(t) = s(t) + n(t) = \sum_m a_m e^{j(\omega_m t + \theta_m)} + \sum_k b_k e^{j(\omega_k t + \phi_k)}; \quad (2)$$

$$\omega_0(t) = 2\pi / T(t), \quad (3)$$

where $T(t)$ is a fundamental period. If $T(t)$ is fixed to $T = 2\pi / \omega_0$, a signal $g(t)$, which is the subtracted signal shifted $x(t)$ to $\pm T$ in time, is

$$g(t) = \frac{2x(t) - x(t - T) - x(t + T)}{4} \quad \text{(4)}$$

$$= \sum_k b_k e^{j(\omega_k t + \phi_k)} \sin^2 \frac{\omega_k}{\omega_0} \pi. \quad \text{(5)}$$

$g(t)$ is transformed by using short-term Fourier transform (STFT). The result $G(\omega_k)$ is

$$G(\omega_k) = N(\omega_k) \sin^2 \frac{\omega_k}{\omega_0} \pi, \quad \text{(6)}$$

where $N(\omega_k)$ is the STFT of the noise $n(t)$. Then, the noise spectrum $N(\omega_k)$ is

$$N(\omega_k) = G(\omega_k) / \sin^2 \frac{\omega_k}{\omega_0} \pi. \quad \text{(7)}$$

Since $N(\omega_k)$ becomes infinite when $\omega_k / \omega_0$ is an integer, $N(\omega_k)$ is actually calculated as

$$\hat{N}(\omega_k) = \begin{cases} \frac{G(\omega_k)}{\sin^2 \frac{\omega_k}{\omega_0} \pi}, & |\sin \frac{\omega_k}{\omega_0} \pi| \geq \varepsilon \\ \frac{G(\omega_k)}{\sin^2 \frac{\omega_k}{\omega_0} \pi}, & |\sin \frac{\omega_k}{\omega_0} \pi| < \varepsilon \end{cases} \quad \text{(8)}$$

where $\varepsilon$ is a certain small value ($\varepsilon > 0$). Thus, in this method, noise is reduced by subtracting $\hat{n}(t)$, the inverse STFT of $\hat{N}(\omega_k)$, from the observed signal $x(t)$. Figure 4 illustrates the frequency response of the model, when $T = 5$ ms. As shown in Fig. 4, pass-bands are controllable as a function of $\varepsilon$, although the proposed frequency filter is the same as a comb filter.

The value of parameter $\varepsilon$ should be given according to features of target speech and noises in order to reduce noises effectively. In preliminary investigations, quality of noise-reduced speech deteriorated if $\varepsilon < 0.3$, effect of noise reduction was no change if $\varepsilon > 0.8$. In this paper, considering that the probabilities of F0s by PHIA are generally between 0.3 and 0.8 according to features of target speech and the SNR of noise, the value of $\varepsilon$ is calculated as

$$\varepsilon = 1.1 - \overline{P}, \quad \text{(9)}$$

where $\overline{P}$ is an average of the probabilities in one frame.

2.3.2. Time-warping to fix F0s

Although the fundamental period is assumed to be fixed in equation (4), real speech has fluctuating fundamental periods, which result in F0 estimation errors. In this method, therefore, speech waves are time-warped to fix their fundamental periods. Figure 5 shows an example of time-warped speech. After this, noise is reduced. Following noise reduction, the speech waves are inversely time-warped once more.

2.4. F0 estimation based on instantaneous frequency

The second F0 estimation of the proposed method needs accuracy for noise-reduced speech. F0 estimation using instantaneous frequency is used as the second, because the F0 estimation based on stability of instantaneous frequency, for example, TEMPO2 proposed by Kawahara et al. [3], can estimate accurate F0s. In this paper, TEMPO2 is used.
3. Evaluation

To compare the robustness of the proposed method with others (i.e., PHIA only, TEMPO2 only and the cepstrum method [5]), simulations are carried out using real speech added to white or pink noise. The evaluation measure is “a correct rate” that the estimated F0s are within ±5% of correct F0s in the voiced section.

3.1. Sound data

The sound data consist of Japanese sentences presented by 14 male and 14 female speakers in the Speech and EGG (electro glottal graph) database [6]. The sampling frequency is 20 kHz. Correct F0s of speech signals are regarded as equal to F0s extracted from EGG waves by TEMPO2. The SNRs of noisy speech are 10, 5, 3 and 0 dB.

3.2. Result

Figure 6 and Figure 7 show percent-correct rate of the F0 estimation methods for speech with white noise and with pink noise.

When speech signals are clean (the SNR is infinity), the correct rates by the proposed method and TEMPO2 are more than 90%. The rate by PHIA is about 72%. That is, if speech signal is clean, PHIA cannot obtain the same accuracy as TEMPO2, but the proposed method is still as accurate as well as TEMPO2.

When white noise is added to speech signals and the SNR is 0 dB, the correct rate by TEMPO2 declines about 50% compared with the rate for clean speech. The rate by PHIA declines only 10% compared with the rate for clean speech, so that PHIA is a robust F0 estimation method even in noisy environments. The rate by the proposed method is about 40% higher than TEMPO2 since the method uses PHIA as the first F0 estimation, and the proposed method is best of all. This result indicates that the proposed method is an accurate and robust F0 estimation method for speech in white noise.

When pink noise is added to speech signals, the correct rates by TEMPO2 and PHIA show the same tendency as for speech with white noise. The rate by the proposed method is over 20% higher than TEMPO2 when the SNR is 0 dB. Pink noise concentrates energy at low frequencies near F0. Therefore, the method suffers more from the influence of pink noise than white noise. This is because TEMPO2 uses the fundamental component and the comb filter cannot reduce noise whose bands are nearby F0s of target speech.

4. Conclusion

This paper proposed the robust and accurate F0 estimation method for noisy speech. This method consists of two different types of F0 estimation method, F0 estimation based on instantaneous amplitude as the robust method, F0 estimation based on instantaneous frequency as the accurate method, and the comb filter with controllable pass-bands to combine the two methods.

The results of simulations showed that the correct rate of the proposed method was approximately the same result of TEMPO2 for clean speech and that it was over 20% higher than TEMPO2, when the SNR of noisy speech was 0 dB.

We conclude that the proposed method can be used not only for speech analysis/synthesis in noisy environments but also for speech segregation using accurate F0s.

5. Acknowledgment

This work was supported by CREST and by Grant-in-Aid for Science research from the Ministry of Education (No. 10680374 and Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists).

6. References