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
1-1 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 passbands to combine the two estimation methods. Simulation results showed that: (1) the proposed method can estimate F0s for clean speech as accurate as the method using only instantaneous frequency, (2) the proposed method can robustly estimate F0s for speech with aperiodic noise in comparison with the other methods such as the cepstrum method, and (3) the proposed method had the capability of estimating F0s for speech with periodic noise.

1. Introduction

It is important for various speech signal processings to have characteristics of speech sounds. For application of speech signal processing such as speech segregation, in real environments, extraction of the characteristics of target speech is required. In speech segregation, the fundamental frequency (F0) as a characteristic is a significant factor characterizing differences between sounds and F0 can be used as a cue for segregation of concurrent speech. For example, Nakatani et al. proposed a computational model of sound stream segregation with a multi-agent paradigm. The agents extracted streams based on harmonics [1]. Unoki and Akagi proposed an auditory segregation model based on constraints related to the four regularities proposed by Bregman. The model used F0 for determining the concurrent time-frequency region of the desired signal [2]. Thus, accurate extraction of F0s from noisy speech is required. However, in noisy environments, it is difficult to estimate accurate F0s because of the interference of noise.

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 based on stability of instantaneous frequencies [3]. This method can estimate F0s for clean speech accurately, but it has difficulties in noisy environments, especially those below 10 dB signal-to-noise ratio (SNR). Unoki and Akagi proposed another method using instantaneous amplitude comb filtering in order to construct a sound segregation model [2]. This method can estimate F0s of 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.

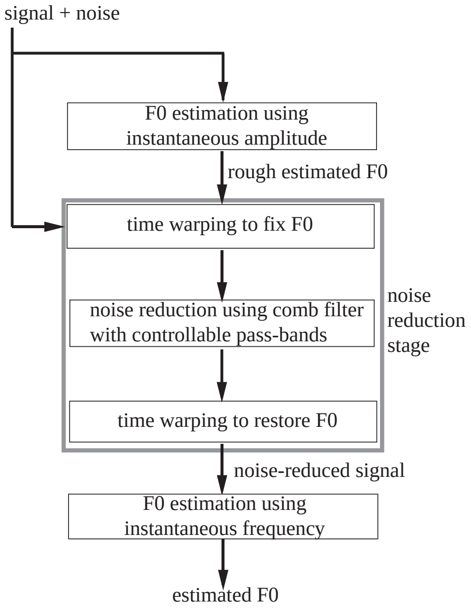
signal + noise
F0 estimation using instantaneous amplitude
rough estimated F0
time warping to fix F0
noise reduction using comb filter with controllable pass-bands
noise reduction stage
time warping to restore F0
noise-reduced signal
F0 estimation using instantaneous frequency
estimated F0

Figure 1: *Algorithm overview.*

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 passbands to combine the two estimation methods.

2. Algorithm

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, probabilities 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.

2.1. 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 [2] 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

$$m(A_k) = \frac{\displaystyle\sum_{A_1 \cap A_2 = A_k} m_1(A_{1i}) m_2(A_{2j})}{1 - \displaystyle\sum_{A_1 \cap A_2 = \phi} m_1(A_{1i}) m_2(A_{2j})}, \qquad (1)$$

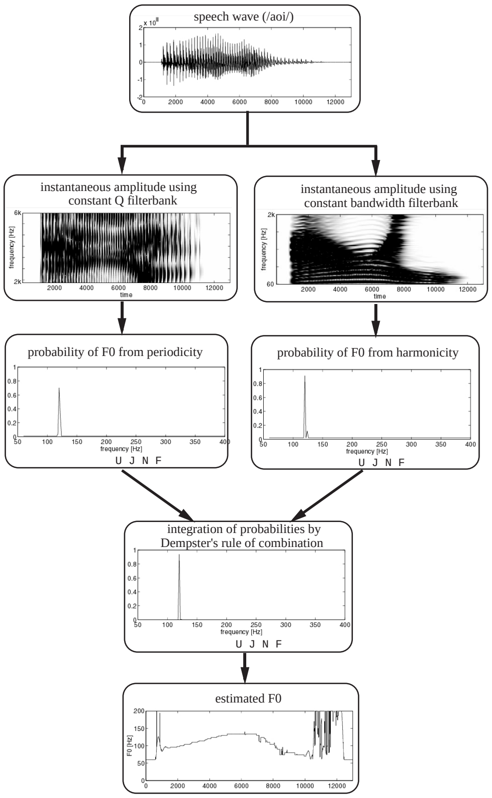
speech wave (/aoi/)
instantaneous amplitude using constant Q filterbank
instantaneous amplitude using constant bandwidth filterbank
frequency [Hz]
time
probability of F0 from periodicity
probability of F0 from harmonicity
frequency [Hz]
U J N F
integration of probabilities by Dempster's rule of combination
estimated F0
F0 [Hz]

Figure 2: *F0 estimation based on periodicity and harmonicity of instantaneous amplitude (PHIA).*

where m_1, m_2 are basic probability function and $A_{1i}, A_{2j} (i, j = 1, 2, 3, ...)$ 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.

2.2. 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.

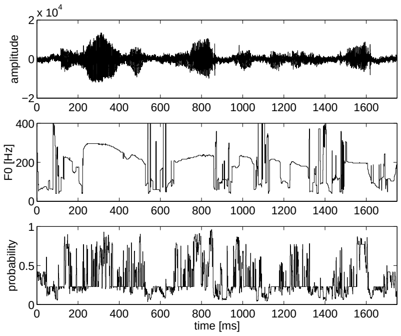

Figure 3: *An example of F0 estimation by PHIA: the speech wave with pink noise (top), the estimated F0s (middle) and probabilities of the F0s (bottom).*

2.2.1. Formulation

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

$$\begin{aligned} x(t) &= s(t)+n(t) \\ &= \sum_m a_m e^{j(m\omega_0(t)t+\theta_m)} + \sum_k b_k e^{j(\omega_k t+\theta_k)}, \quad (2) \\ \omega_0(t) &= 2\pi/T(t), \quad (3) \end{aligned}$$

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

$$\begin{aligned} g(t) &= \frac{2x(t)-x(t-T)-x(t+T)}{4} \quad (4) \\ &= \sum_k b_k e^{j(\omega_k t+\theta_k)} \sin^2\frac{\omega_k}{\omega_0}\pi. \quad (5) \end{aligned}$$

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

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

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

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

Since $N(\omega_k)$ becomes infinite when ω_k/ω_0 is an integer, $N(\omega_k)$ is actually calculated as

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

where ε 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 ε, although the proposed frequency filter is the same as a comb filter.

The value of parameter ε 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 ε is calculated as

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

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

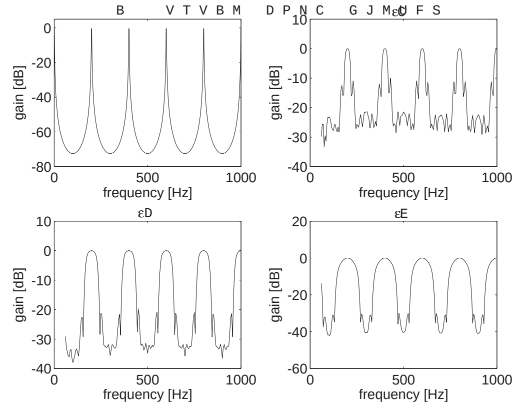

Figure 4: *Frequency response of the comb filter with controllable pass-bands (T=5 ms).*

2.2.2. 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. After this, noise is reduced. Following noise reduction, the speech waves are inversely time-warped once more.

2.3. 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. Simulations

3.1. Simulation 1 : Speech with aperiodic noise

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

The sound data consist of Japanese sentences presented by 14 male and 14 female speakers in the Speech and EGG (electro glottal graph) database [5]. 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.

Figure 5 shows percent-correct rate of the F0 estimation methods for speech with white noise. When speech signals are

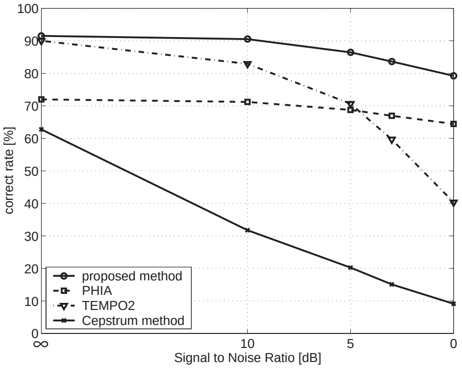

Figure 5: *The correct rates for speech with white noise.*

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, the proposed method can obtain the same accuracy as TEMPO2, while PHIA cannot obtain it. When 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.

3.2. Simulation 2 : Speech with periodic noise

Estimating F0s for speech with periodic noise is more difficult than that with aperiodic. In order to investigate PHIA's capability of estimating F0s for speech with periodic noise, we carried out a simulation using a vowel which was added a complex tone.

The sound data is a female vowel /a/ of the ATR Japanese speech database [6]. We added a complex tone, which has the F0 of 200 Hz, as interference noise to the vowel. The SNR of the noisy vowel is 0 dB. Figure 6 shows the noisy vowel and the probabilities made by PHIA for it. Black parts in Figure 6(b) indicate high probabilities of F0s. In speech section, the probabilities have some local maxima at frequency domain, as shown in Figure 6. One of these maxima corresponds to F0s of the vowel /a/ and the other correspond to F0s of the complex tone. In other word, the probabilities for speech with periodic noise have information about F0s of the target speech. PHIA can roughly estimate correct F0s for speech with periodic noise by tracking probabilities of the F0s using properties of F0 transition in time and frequency domains. When PHIA can roughly estimate F0s of the target signal from the mixed signal, the F0s are then used for noise reduction by the comb filter, i.e. the proposed method can estimate accurate F0s from speech with periodic noise.

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 and F0 estimation based on instantaneous frequency as the accurate method. The comb filter with controllable pass-bands is used for combining the two methods. The simulation results showed that the correct rate of the proposed method was equivalent to that of TEMPO2 for clean speech and it was over 20% higher than that of TEMPO2 when the SNR of speech with aperiodic noise was 0 dB. Other simulation results showed that this method had the capability of estimating F0s from speech with periodic noise.

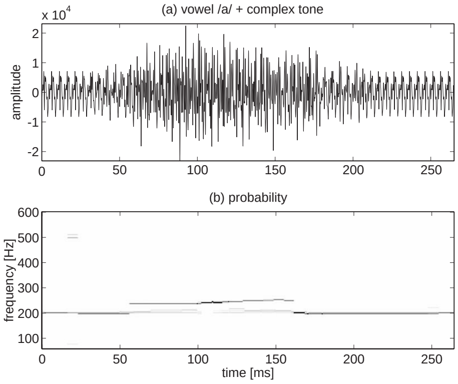

Figure 6: *The vowel /a/ with the complex tone and the probabilities by PHIA.*

Using the proposed method, F0s of target speech as a cue for segregation can be extracted accurately even in noisy environments. We conclude that the proposed method can be used for applications of speech signal processing such as speech segregation, support for automatic speech recognition and concurrent dialogue analysis in real environments.

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

[1] Nakatani, T. et al., "A computational model of sound stream segregation with multi-agent paradigm", Proc. ICASSP95, Vol.4, pp.2671-2674, 1995.

[2] Unoki, M. and Akagi, M., "Signal extraction from noisy signal based on auditory scene analysis", Proc. ICSLP98, Vol.4, pp.1515-1518, 1998.

[3] Kawahara, H. et al., "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity", Proc. Eurospeech99, pp.2781-2784, 1999.

[4] Shafer, G., a mathematical theory of evidence, Princeton University Press, 1976.

[5] Atake, Y. et al., "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components", Proc. ICSLP2000, Vol.2, pp.907-910, 2000.

[6] Takeda, K. et al., Speech Database User's Manual, ATR Technical Report TR-I-0028, 1988.