Fundamental frequency estimation for noisy speech based on instantaneous amplitude and frequency

Yuichi ISHIMOTO, Masashi UNOKI, and Masato AKAGI

28 March 2005
IS-RR-2005-006

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
y-ishi@jaist.ac.jp, unoki@jaist.ac.jp, akagi@jaist.ac.jp

ISSN 0918-7553

Fundamental frequency estimation for noisy speech based on instantaneous amplitude and frequency

Yuichi Ishimoto, Masashi Unoki, Masato Akagi

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

Abstract

This paper proposes a robust and accurate method of estimating the fundamental frequencies (F0s) for noisy speech. In general, it is difficult to directly estimate accurate F0s from noisy speech. This method combines two different methods of F0 estimation. One is based on the periodicity and harmonicity of instantaneous amplitude of speech; it is robust against noise, but it does not allow for accurate F0 estimation. The other is based on the stability of instantaneous frequency, and it enables accurate F0 estimation, but this method is not robust against noise. To combine these two methods, the proposed method makes use of noise reduction by using a comb filter with controllable pass-bands. Experiments were carried out to estimate F0s of real speech in noisy environments and to compare the proposed method with other methods such as an autocorrelation method and a cepstrum method. The results showed that this method was more robust than the other methods. This method could estimate F0s of noisy speech with accuracy similar to that in clean speech F0 estimation by using only the stability of instantaneous frequency.

Key words: Periodicity; Harmonicity; Instantaneous amplitude; Instantaneous frequency; Comb filter; Noise reduction

1 Introduction

Extracting fundamental frequencies (F0s) of target speech signals is important in various areas of speech signal processing such as speech recognition, speech analysis/synthesis, and speech segregation. For example, in speech recognition, F0 information can be used for spectrum normalization to improve the accuracy of speech recognition (Singer and Sagayama, 1992). In speech analysis/synthesis, F0 is a factor controlling the pitch of speech, so extracting accurate F0s is important if the synthesized speech is to sound natural (Kawahara et al., 1999). In 'Auditory Scene Analysis', F0 is a significant element characterizing the difference between sounds that can be used as cues to segregate concurrent speech signals (Bregman, 1990). There have been many studies in the field of 'Computational Auditory Scene Analysis', for example, on speech segregation, which used F0s from target speech in noisy environments (e.g., Nakatani et al., 1995; Unoki and Akagi, 1999; see also, Cooke and Ellis, 2001, for a detailed review). In order to use speech signal processing in real environments, we must be able to extract accurate F0s. However, accurately extracting F0s from target speech in noisy environments is difficult, because noise distorts fundamental harmonic components of the target speech.

Various methods of F0 estimation have been proposed (for further details, see Hess, 1983, 1992; Hermes et al., 1993), most of which make use of periodic features of speech in the time domain or harmonic features in the frequency domain. F0s from periodic features of speech in the time domain are currently

extracted using an autocorrelation function of speech waveforms (Takagi et al., 1997; de Cheveigné and Kawahara, 2002). This autocorrelation method is relatively robust against noise, but it is not very accurate. F0s from harmonic features in the frequency domain are extracted using a cepstrum method (Noll, 1967) or by comb filtering of the amplitude spectrum (e.g., Nishi and Ando, 1998). The cepstrum method can extract relatively accurate F0s in noiseless speech, but it is not very robust against noise. The method based on comb filtering of the amplitude spectrum is more robust against noise than the cepstrum method. These methods, however, cannot estimate F0s in noiseless speech with accuracy similar to that of the F0 estimation method using instantaneous frequency described below.

The time-frequency representation of speech obtained by time-frequency analysis can also represent harmonic components of speech (Cohen, 1995). The instantaneous amplitude of speech signals, which are analyzed by filterbanks and are represented in the time-frequency domain, has harmonic features that are robust against noise (Unoki and Akagi, 1998; Ishimoto et al., 2001). A method of F0 estimation based on the comb filtering of instantaneous amplitude was proposed by Unoki and Akagi (1998) to construct a sound segregation model. This method can estimate F0s of vowels in noisy environments. However, the estimated F0s of sentences are not sufficiently accurate.

Instantaneous frequency of speech has also been used to accurately estimate F0s (Qiu et al., 1995; Abe et al., 1996; Kawahara et al., 1999). For example, Kawahara et al.(1999) proposed a method of F0 estimation, STRAIGHT-TEMPO, based on the stability of instantaneous frequencies to construct a speech analysis/synthesis system. This method can accurately estimate F0s in clean speech, however, it is sensitive to noise. Other, more robust meth-

ods using instantaneous frequency in noisy environments have been proposed (Atake et al., 2000; Nakatani and Irino, 2002); some of these methods are more robust than STRAIGHT-TEMPO.

Nevertheless it appears that most existing methods have certain drawbacks in estimating F0s of target speech in noisy environments and cannot be both accurate and robust.

This paper proposes a robust and accurate method for estimating F0s even in noisy environments. The proposed method (1) estimates rough F0s from noisy speech, (2) reduces noise by using the roughly estimated F0s, and (3) estimates accurate F0s from the noise-reduced speech.

2 The model of F0 estimation for noisy speech

In general, it is difficult to directly estimate accurate F0s from noisy speech. The proposed method reduces the noise and then estimates accurate F0s from the noise-reduced speech. The noise is reduced by using a comb filter that eliminates noise components in the noisy speech, leaving harmonic components intact (de Cheveigné, 1993, 1997). The center frequencies of the comb filter must be the same as those of the harmonics of the target speech. Hence, to construct a comb filter, we first need to extract the F0s of the target speech. We cannot obtain accurate F0s directly from noisy speech, but we can obtain rough one. In this method, we first try to obtain F0s as accurately as possible in order for the roughly estimated F0s not to reduce the harmonics of the target speech as a result of comb filtering.

Figure 1 shows a block diagram of the proposed method. First, the method roughly estimates F0s from noisy speech using instantaneous amplitude, which has robustly periodic and harmonic features corresponding to F0s, in the time-frequency domain. The F0 estimation is based on the periodicity and harmonicity of instantaneous amplitude of speech signals by using time-frequency analysis. Then noise reduction is performed using a comb filter with controllable pass-bands, whose center frequencies are calculated from the roughly estimated F0s. The pass-band width is controlled depending to the amount of noise present. Before reducing the noise, this method performs time warping of the noisy-speech waveform in order to flatten the F0 contours, because the comb filter is constructed by assuming that the F0s in a frame are constant. This procedure decreases the amount of error in noise reduction. After that, accurate F0s are estimated from the noise-reduced speech waveform using the stability of instantaneous frequency.

In the following sections, we describe the proposed method in detail. Section 2.2 outlines F0 estimation based on the periodicity and harmonicity of instantaneous amplitude; section 2.3 describes noise reduction by using a comb filter with controllable pass-bands; and section 2.4 presents F0 estimation based on the instantaneous frequency.

2.2 F0 estimation based on the periodicity and harmonicity of instantaneous amplitude

The first stage of the proposed method requires that it be robust in noisy environments, although the estimated F0s do not need to be very accurate. Time-frequency representation of the instantaneous amplitude of speech can contain information about F0s even in noisy environments. To enable robust estimation of F0s in noisy speech, the proposed method uses harmonicity, which appears as fluctuation in instantaneous amplitude with intervals corresponding to the F0 in the frequency direction, and periodicity, which appears as fluctuation in instantaneous amplitude with intervals corresponding to the fundamental period in the time direction. Figure 2 illustrates F0 estimation based on the periodicity and harmonicity of instantaneous amplitude (called PHIA), which is the first stage of the proposed method.

2.2.1 Time-frequency representation of instantaneous amplitude

The time-frequency representation of speech signals is obtained as follows. An input signal, $x(t)$, is analyzed by using band-pass filters $h_k(t)$. The outputs of filterbank, $y_k(t)$, are given by

$$y_k(t) = x(t) * h_k(t), \tag{1}$$

where k is the channel index, and * denotes the operation of the convolution. The analytic signal, $\tilde{y}_k(t)$, is obtained as

$$\tilde{y}_k(t) = \mathcal{F}^{-1}\left[2Y_k(\omega)U(\omega)\right], \tag{2}$$

$$U(\omega) = \begin{cases} 1, & \omega > 0 \\ 1/2, & \omega = 0 \\ 0, & \omega < 0 \end{cases} \tag{3}$$

where $Y_k(\omega)$ is the Fourier spectrum of $y_k(t)$ and $\mathcal{F}^{-1}[\cdot]$ is the inverse Fourier transform. Then, the instantaneous amplitude, $s_k(t)$, and instantaneous frequency, $\lambda_k(t)$, are given in Eqs. (4) and (5), respectively.

$$s_k(t) = |\tilde{y}_k(t)|, \tag{4}$$

$$\lambda_k(t) = \frac{\partial}{\partial t} \arg \tilde{y}_k(t). \tag{5}$$

In this time-frequency analysis, the proposed method uses two filterbanks as the band-pass filters. One is a constant-Q filterbank, and the other is a constant narrow bandwidth filterbank. In this paper, the filterbanks are constructed by using the gammatone filter (Patterson and Holdworth, 1991), which is represented by

$$gt(t) = At^{N-1} \exp(-2\pi b_f ERB(f_c)t) \cos(2\pi f_c t), \qquad (t > 0) \tag{6}$$

where N and b_f are the parameters related to bandwidth, f_c is the center frequency and ERB(f_c) is the Equivalent Rectangular Bandwidth of f_c (Glasberg and Moore, 1990).

In constructing a constant-Q gammatone filterbank (for details, see Unoki and Akagi, 1999),

$$h_k(t) = \frac{1}{\sqrt{|a|}} gt\left(\frac{t}{a}\right), \tag{7}$$

$$a = 10^{(2/K)(k-1)-1}, \tag{8}$$

where $N = 4$, $f_c = 600$ Hz, $b_f = 1$, $1 \leq k \leq K + 1$, and $K = 64$. Then Eq. (1) is the wavelet transform. In practice, filters whose center frequencies are in the range of 2 to 6 kHz, that is, which have 16 channels, are used.

In constructing a constant-bandwidth filterbank, $h_k(t)$ is given by

$$h_k(t) = At^{N-1} \exp(-2\pi b_f t) \cos(2\pi f_k t), \tag{9}$$
$$f_k = 60 + 5(k-1) \text{ Hz}, \tag{10}$$

where $N = 4$, $b_f = 20$ Hz and $1 \leq k \leq 389$. f_k means that the center frequencies of these filters are ranging from 60 to 2000 Hz. The boundary between the two filterbanks is decided as 2 kHz because the expedient periodic feature represented by a constant-Q filterbank did not appear below 2 kHz in our preliminary investigation, and the frequency region below 2 kHz is used to extract the harmonic feature represented by a constant-bandwidth filterbank.

When $s_k(t)$ is allocated by the channel index in the time-frequency plane by using a constant-Q filterbank, periodic fluctuations appear in the instantaneous amplitude in the time direction. The peak intervals of the fluctuations equal the fundamental period, which is the reciprocal of the F0. These fluctuations are called 'periodicity'. Periodicity is especially noticeable in the high-frequency region because the constant-Q filterbank has high temporal resolution in the high-frequency region.

Similarly, when a constant-bandwidth filterbank is used, fluctuations in the instantaneous amplitude appear in the frequency direction. These fluctuations are called 'harmonicity' and their peak intervals equal the F0. Harmonicity is especially prominent in the low-frequency region, because speech harmonics

tend to swerve as a result of harmonicity in the high frequency region. The instantaneous amplitude obtained by using a constant-bandwidth filterbank can alternatively be implemented by using fast Fourier transformation.

The upper panel in Fig. 3 shows the instantaneous amplitude calculated by using a constant-Q filterbank for male vowel /a/. The bottom panel in Fig. 3 is the instantaneous amplitude on one channel (#13) of the filterbank, which is indicated by the dashed line in the upper panel. The periodicity is clearly visible in the time direction. The left panel in Fig. 4 shows the instantaneous amplitude obtained by using a constant-bandwidth filterbank for the same vowel. The right panel in Fig. 4 is the instantaneous amplitude at one time (320 ms), which is indicated by the dashed line in the left panel. The harmonicity is clearly visible in the frequency direction.

2.2.2 Calculating F0 distributions derived from periodicity and harmonicity

Next, this method calculates F0 distributions from the periodicity and harmonicity that indicate the presence of F0s. In the time-frequency representation of periodicity, F0 candidates are extracted by using an autocorrelation function in the time domain. The autocorrelation function of lag τ in the time direction is given in Eq. (11).

$$r_{k,t}(\tau) = \sum_{i=t}^{t+w_t} \bar{s}_k(i)\bar{s}_k(i+\tau), \tag{11}$$

where w_t is 35 ms, and $\bar{s}_k(t)$ is center-clipped $s_k(t)$ using the average amplitude. $r_{k,t}(\tau)$ is calculated in 1-ms steps for every filter channel. Then, F0 candidates are estimated at time t and channel index k from $r_{k,t}(\tau)$. A histogram, $p_t(f)$, is constructed using the F0 candidates gathered from all filter

channels within 20 ms, where f is the bin of the histogram that contains frequency.

Similarly, in the representation of harmonicity, F0 candidates are extracted by using an autocorrelation function in the frequency domain. The autocorrelation function of lag ζ in the frequency direction is given by

$$r'_{w_k,t}(\zeta) = \sum_{j=1}^{w_k} \tilde{s}_j(t)\tilde{s}_{j+\zeta}(t), \tag{12}$$

where w_k is the filter index that contains 800, 1100, 1400, 1700 or 2000 Hz, and $\tilde{s}_k(t)$ is center-clipped $\log(s_k(t))$ using the average amplitude. $r'_{w_k,t}(\zeta)$ is also calculated in 1-ms steps. Then, F0s are estimated as candidates from each $r'_{w_k,t}(\zeta)$. A histogram, $q_t(f)$, is constructed using the F0 candidates gathered from all filter channels within 20 ms.

Then, 1 is added to all the bins of the histogram as offset. If there is no offset, the correct F0 information may be lost at integration of the F0 distributions described in the next section. The obtained histograms are then normalized so that their sum is 1. We call the resultant histogram 'F0 distribution'.

2.2.3 Integration of F0 distributions and F0 estimation

If both distributions obtained from the periodicity and harmonicity of instantaneous amplitude have a peak at the same point, the peak is assumed to indicate the correct F0. However, if the distributions have peaks at different points, they may not indicate the correct F0. To improve F0 estimation in noisy environments, the method integrates the two F0 distributions as

$$d_t(f) = p_t(f)q_t(f). \tag{13}$$

Finally, the frequency with the peak in the integrated F0 distribution $d_t(f)$ is extracted as the F0.

Figure 5 shows an example of estimated F0s and the peak values of F0 distributions at each time. In the voiced section, the peak values are relatively close to 1. In the unvoiced, or noisy section, they are close to 0. This means that the peak value of the F0 distributions shows whether or not the signal is noisy. The line close to 0.2 in the peak values is due to the offset of the histogram described in section 2.2.2.

2.3 Noise reduction by using a comb filter with controllable pass-bands

2.3.1 Comb filter construction

The second stage of the proposed method requires a comb filter with controllable pass-bands that can both decrease the amount of error in the F0 estimation obtained in the first stage and minimize the amount of noise. This comb filter is constructed as follows.

We assume that the target signal $s(t)$ is a harmonic complex tone, and $n(t)$ is noise. We obtain the observed signal, $x(t)$, as

$$\begin{aligned} x(t) &= s(t) + n(t) \\ &= \sum_l a_l e^{j(l\omega_0(t)t+\theta_l)} + \sum_m b_m e^{j(\omega_m t+\theta_m)}, \end{aligned} \tag{14}$$

$$\omega_0(t) = 2\pi/T_0(t), \tag{15}$$

where $T_0(t)$ is a fundamental period. To simplify the construction of the comb filter, we assume that $T_0(t)$ is fixed to

$$T_0 = 2\pi/\omega_0. \tag{16}$$

If we eliminate the two signals, which are shifted $x(t)$ to $t \pm T_0$, from $x(t)$, then $c(t)$ is calculated as follows.

$$c(t) = \frac{2x(t) - x(t - T_0) - x(t + T_0)}{4} \tag{17}$$

$$= \sum_m b_m e^{j(\omega_m t + \theta_m)} \sin^2 \frac{\omega_m}{\omega_0}\pi. \tag{18}$$

Signal $c(t)$ is transformed by using a short-term Fourier transform (STFT). The result is $C(\omega_m)$:

$$C(\omega_m) = N(\omega_m) \sin^2 \frac{\omega_m}{\omega_0}\pi, \tag{19}$$

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

$$N(\omega_m) = C(\omega_m) / \sin^2 \frac{\omega_m}{\omega_0}\pi. \tag{20}$$

By transforming $N(\omega_m)$ using the inverse STFT, we estimate noise $\hat{n}(t)$. Then, the noise is reduced by subtracting $\hat{n}(t)$ from the observed signal, $x(t)$. Since $N(\omega_m)$ becomes infinite when ω_m/ω_0 is an integer, $N(\omega_m)$ is calculated as

$$\hat{N}(\omega_m) = \begin{cases} C(\omega_m) / \sin^2 \frac{\omega_m}{\omega_0}\pi, & |\sin \frac{\omega_m}{\omega_0}\pi| \geq \varepsilon \\ C(\omega_m), & |\sin \frac{\omega_m}{\omega_0}\pi| < \varepsilon \end{cases} \tag{21}$$

where $0 \leq \varepsilon \leq 1$. This equation means that the pass-bands of the comb filter can be easily changed by ε. Figure 6 illustrates the frequency response in the noise reduction model by using a comb filter when T_0 is 5 ms. As shown in Fig. 6, the pass-bands are controlled as a function of ε. That is, if ε is small, the pass-bands are narrow, and if ε is large, the pass-bands are wide.

2.3.2 *Determining ε*

To reduce noise effectively and minimize the effect of error in the F0 estimation in the first stage, the value of ε should be given according to the features of the target speech and noise. Therefore, to construct a comb filter, we need to determine whether or not the observed signal is noisy. As shown earlier, the peak values of the F0 distributions obtained with PHIA vary with noise, hence these values are used to determine ε. In our preliminary investigation, we found that the quality of noise-reduced speech deteriorated when ε was less than 0.3, and the noise could not be reduced when ε was greater than 0.8. Given that the peak values of the F0 distributions obtained by PHIA generally fall between 0.3 and 0.8 depending on the features of the target speech and SNR (see Fig. 5), the value of ε is calculated as follows.

$$\varepsilon = 1.1 - \overline{P}, \tag{22}$$

where $\overline{P}$ is the average of the peak values of the F0 distributions in 20 ms. When $\overline{P}$ is close to 0 (i.e. the error of F0 estimation in the first stage is large), the pass-bands of the comb filter are made wide in order for the comb filter to reduce harmonic components of the target speech. When $\overline{P}$ is close to 1 (i.e., the error is small), the pass-bands are made narrow in order for the comb filter to reduce noise components more effectively.

2.3.3 *Time-warping of speech waveforms to flatten the F0*

In the above formulation of the comb filter, we assumed that the fundamental period is constant in Eq. (17). However, in real speech, the fundamental periods vary with time. This causes error in F0 estimation. Therefore, before

reducing the noise by using a comb filter, the speech waveforms are time-warped to flatten their fundamental periods using the roughly estimated F0s.

The time-warping of speech waveforms to flatten the F0 contours is performed by varying the sampling interval and resampling. The sampling interval is varied by using fundamental periods $T_0(t)$ estimated in the first stage. The sampling interval, $\widetilde{T}_s(t)$, is calculated as

$$\widetilde{T}_s(t) = \frac{\overline{T}_0}{T_0(t)} \times T_s, \tag{23}$$

where $\overline{T}_0$ is the average of $T_0(t)$, and T_s is the sampling period of the original speech waveform. The speech waveform represented by using $\widetilde{T}_s(t)$ is resampled by using interpolation in T_s steps. The speech waveform, then, has a constant F0. Figure 7 shows an example of time-warped speech. After the noise has been reduced by the comb filter, the speech waveforms are once again inversely time-warped to restore their estimated fundamental periods.

2.4 F0 estimation based on instantaneous frequency

The third stage of the proposed method is estimating the F0s of the noise-reduced signals. This estimation must be highly accurate. In mapping from the center frequencies of the band-pass filters to the instantaneous frequencies of the filter outputs, the fixed points that contain F0 information appear at all times if the signal is periodic. The F0 estimation method based on this stability of instantaneous frequency can give accurate F0s (for details, see Kawahara et al., 1999). In this paper, therefore, we consider STRAIGHT-TEMPO proposed by Kawahara et al. as the third stage of our method.

3 Experiments

To compare the robustness and accuracy of the proposed method with those of other methods, we carried out experiments using real speech to which we added white and pink noise. The evaluated methods were the proposed method, PHIA, STRAIGHT-TEMPO, YIN (de Cheveigné and Kawahara, 2002), an autocorrelation method, and a cepstrum method. It should be noted that here are many autocorrelation and cepstrum methods; in this study, we used an autocorrelation method with multiple window lengths (Takagi et al., 1997) and an improved cepstrum method developed by Kato and Miwa (1995). The evaluation measures were 'the gross F0 error' and 'the fine F0 error'. The former was defined as a more than 20% difference from the reference F0s in the voiced section. The latter was defined as a standard deviation of the error within the threshold of the gross F0 error. Hence, the gross F0 error indicated the robustness of the methods in noisy environments. The fine F0 error indicated the accuracy of the methods within the threshold of the gross F0 error.

To evaluate the methods, we used a database of simultaneous recordings of speech sounds and electroglottographs (EGGs) (Atake et al., 2000). This database contains 30 short Japanese sentences pronounced by 14 male and 14 female speakers, together with voiced-unvoiced labels, which were used to detect the voiced sections in these experiments. The reference F0s of the speech signals were F0s extracted from EGG waves by STRAIGHT-TEMPO. This is because STRAIGHT-TEMPO had the least bias in F0 estimation from the EGG waves in the preliminary experiment (see Appendix). The sampling frequency was 16 kHz. The SNRs of the noisy speech were 10, 5, 3, and 0 dB.

3.1.1 White noise

Figures 8 and 9 show the gross F0 error and the fine F0 error obtained for the six F0 estimation methods for speech signals containing white noise.

When the speech signals were clean (the SNR was infinite as shown in Figs. 8 and 9), the gross F0 error obtained with almost all the methods was about 5%. YIN had the best performance. The fine F0 error obtained with STRAIGHT-TEMPO was about 1.8 Hz and that of the proposed method was about 2.1 Hz, while that of PHIA was 6.2 Hz. That is, when the speech signal was clean, PHIA could not provide the same accuracy as the other methods, however, the proposed method's accuracy was close to that of STRAIGHT-TEMPO. The fine F0 error of YIN was larger than those of the other methods. One reason for this is that YIN reduces only those errors that are too-high or too-low, and does not reduces fine F0 errors.

When white noise was added to the speech signals and the SNR was 0 dB, the gross error obtained with the proposed method had the best performance of all the methods. The gross F0 error of STRAIGHT-TEMPO increased more than 40% compared with the error for clean speech. The gross F0 error of PHIA increased only 11% compared with the error for clean speech. PHIA, therefore, has high robustness against white noise. The proposed method could have the same robustness as that of PHIA, which is the first stage of the proposed method, and it was about 30% lower than that of STRAIGHT-TEMPO, which is the third stage of the proposed method. The results showed that the proposed method can greatly improve the robustness of STRAIGHT-TEMPO.

The autocorrelation and the cepstrum methods were also relatively robust, however, their gross F0 error was low compared with that of the proposed method under noisy conditions.

The fine F0 error obtained with STRAIGHT-TEMPO was very small even in noisy speech. This result means that the estimated F0s were accurate within the threshold of the gross F0 error, although most of the F0 estimation errors exceeded the threshold. The fine F0 error of the proposed method was much smaller than the error of PHIA, and it was close to that of STRAIGHT-TEMPO. This result indicates that the second and third stages of the proposed method can reduce the error of PHIA in the first stage and that the method can estimate accurate F0s under noisy conditions. The fine F0 error of the autocorrelation and cepstrum methods was large compared with that of the proposed method.

Based on the results, we conclude that STRAIGHT-TEMPO and YIN are not robust against noise. The autocorrelation and cepstrum methods are relatively robust for noisy speech, but they are not accurate. The proposed method is the most robust against white noise of all the methods, and it is as accurate for noisy speech as STRAIGHT-TEMPO for clean speech.

3.1.2 Pink noise

Figures 10 and 11 show the gross F0 error and the fine F0 error for speech signals containing pink noise.

When pink noise was added to the speech signals, the gross F0 error of all the methods showed the same tendency towards increasing. However, the gross

F0 error obtained with the proposed method was more than 20% lower than that of STRAIGHT-TEMPO when the SNR was 0 dB, and it was more than 10% lower than that of STRAIGHT-TEMPO when the SNR was 10 dB. The proposed method had the best performance in terms of the gross F0 error when the SNR was more than 3 dB, and its fine F0 error was also small compared with that of the autocorrelation and cepstrum methods. When the SNR was 0 dB, the gross F0 error of the proposed method was a little higher than that of the cepstrum method. However, the fine F0 error of the proposed method was smaller than that of the cepstrum method. The proposed method suffered more from the effect of pink noise than it did from the effect of white noise. This is because pink noise has higher energy in the low-frequency region around F0s than in the high-frequency region. STRAIGHT-TEMPO, which is the third stage of the proposed method, uses fundamental components of speech signals, and it is sensitive to pink noise. The comb filter, which is used in the second stage of the proposed method, cannot effectively reduce noise in the fundamental components. The proposed method, therefore, cannot perform as robustly as the proposed method against white noise.

4 Conclusion

This paper described a robust and accurate method of estimating the F0s of noisy speech. The proposed method consists of two different methods of F0 estimation. One is based on the time-frequency representation of instantaneous amplitude, which is relatively insensitive to noise. Therefore, F0 estimation based on instantaneous amplitude enables robust F0 extraction from noisy speech, although the extracted F0s are not highly accurate. The other method

is F0 estimation based on instantaneous frequency. Although F0 estimation using instantaneous frequency is sensitive to noise, this method can extract accurate F0s from noiseless speech. The proposed method combines these two methods of F0 estimation, reducing the noise by using a comb filter with controllable pass-bands, which enables the method to be both robust and accurate.

Experimental results showed that the gross F0 error of the proposed method was approximately the same as that of STRAIGHT-TEMPO for clean speech and that it was about 30% lower than that of STRAIGHT-TEMPO when the SNR of the speech with white noise was 0 dB. The proposed method was most effective when the SNR of noisy speech was more than 3 dB.

Thus, the proposed method can be used not only for speech analysis/synthesis in noisy environments but also for speech segregation that requires estimating accurate F0s.

Acknowledgements

This work was supported by CREST (Core Research for Evolutional Science and Technology) of the Japan Science and Technology Corporation (JST) and by a grant-in-aid for scientific research from the Ministry of Education (No. 14780267).

Table 1 Fine F0 error for clean speech.

F0 estimation method	from EGG (Reference)			
	AC	**Cepstrum**	**YIN**	**TEMPO**
Autocorrelation (AC)	7.52 Hz	4.42 Hz	6.57 Hz	3.08 Hz
Cepstrum method	4.93	3.62	7.30	3.92
YIN	8.17	8.84	6.91	7.27
STRAIGHT-TEMPO	4.06	3.90	6.70	1.64
average	**6.17**	**5.20**	**6.87**	**3.98**

Appendix A

To investigate which method was the most effective in extracting the F0s from EGGs, we evaluated the fine F0 error for clean speech using four F0 estimation methods. The methods were an autocorrelation method using multiple window lengths (Takagi et al., 1997), an improved cepstrum method (Kato and Miwa, 1995), YIN (de Cheveigné and Kawahara, 2002), and STRAIGHT-TEMPO (Kawahara et al., 1999). The same database was used as in the experiments in section 3. The evaluation measure was the fine F0 error. The results are shown in table 1. The columns are the methods for extracting the F0 from EGG waves (used as reference F0s), and the rows are the methods for extracting F0s from clean speech. The bottom of the table shows the average fine F0 error: the smaller the error, the less biased the method. As can be seen in the table, although all the methods have a certain amount of bias, STRAIGHT-TEMPO appears to have the least bias of the four methods. In this study,

therefore, STRAIGHT-TEMPO was used to extract the reference F0s from EGGs.

References

Abe, T., Kobayashi, T., Imai, S., 1996. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. Proc. ICSLP96, Vol.2, pp.1277-1280.

Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S., Shikano, K., 2000. Robust fundamental frequency estimation using instantaneous frequencies of harmonic components. Proc. ICSLP2000, Vol.2, pp.907-910.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, MA.

Cohen, L., 1995. Time-frequency analysis. Prentice Hall PTR, New Jersey.

Cooke, M., Ellis, D.P.W., 2001. The auditory organization of speech and other sources in listeners and computational models. Speech Communication 35, pp.141-177.

de Cheveigné, A., 1993. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Am., 93(6), pp.3271-3290.

de Cheveigné, A., 1997. Concurrent vowel identification III: A neural model of harmonic interference cancellation. J. Acoust. Soc. Am., 101(5), pp. 2857-2865.

de Cheveigné, A., Kawahara, H., 2002. Yin, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am., 111(4), pp. 1917-1930.

Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Research 47. pp.103-138.

Hermes, D.J., 1993. Pitch analysis. In: Cooke, M., Beet, S., Crawford, M. (Eds.), Visual representations of speech signals., John Wiley & Sons, Chichester.

Hess, W., 1983. Pitch determination of speech signals. Springer-Verlag, Berlin.

Hess, W.J., 1992. Pitch and voicing determination. In: Furui, S., Sondhi, M.M. (Eds.), Advances in speech signal processing. Marcel Dekker, New York.

Ishimoto, Y., Unoki, M., Akagi, M., 2001. A fundamental frequency estimation method for noisy speech based on instantaneous amplitude and frequency. Proc. Eurospeech2001, Vol.4, pp.2439-2442.

Kato, S., Miwa, J., 1995. Pitch detection using moving average and band-limitation in cepstrum method and its application. Technical report of IEICE, SP94-95.

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication 27, pp.187-207.

Kawahara, H., Katayose, H., de Cheveigné, A., Patterson, R.D., 1999. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. Proc. Eurospeech99, pp.2781-2784.

Nakatani, T. Kawabata, T., Okuno, H.G., 1995. A computational model of sound stream segregation with multi-agent paradigm. Proc. ICASSP95, Vol.4, pp.2671-2674.

Nakatani, T., Irino, T., 2002. Robust fundamental frequency estimation against background noise and spectral distortion. Proc. ICSLP2002, pp.1733-1736.

Nishi, K., Ando, S., 1998. An optimal comb filter for time-varying harmonics

extraction. IEICE Trans. Fundamentals, Vol.E81-A, NO.8, pp.1622-1627.

Noll, A.M., 1976. Cepstrum pitch determination, J. Acoust. Soc. Am., Vol.41, pp.293-309.

Patterson, R.D., Holdsworth, J., 1991. A functional model of neural activity patterns and auditory images. In: Ainsworth, W.A., Evans, E.F., Hackney, C.M. (Eds.), Advances in speech, Hearing and language processing. Vol.3, JAI Press, London.

Qiu, L., Yang, H., Koh, S.N., 1995. Fundamental frequency determination based on instantaneous frequency estimation. Signal Processing 44, pp.233-241.

Singer, H., Sagayama, S., 1992. Pitch dependent phone modeling for HMM based speech recognition. Proc. ICASSP92, Vol.1, pp.273-276.

Takagi, T., Seiyama, N., Miyasaka, E., 1997. A method for pitch extraction of speech signals using autocorrelation function through multiple window-lengths. IEICE vol.J80-A, No.9, pp.1341-1350.

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

Unoki, M., Akagi, M., 1999. A method of signal extraction from noisy signal based on auditory scene analysis. Speech Communication 27, pp.261-279.

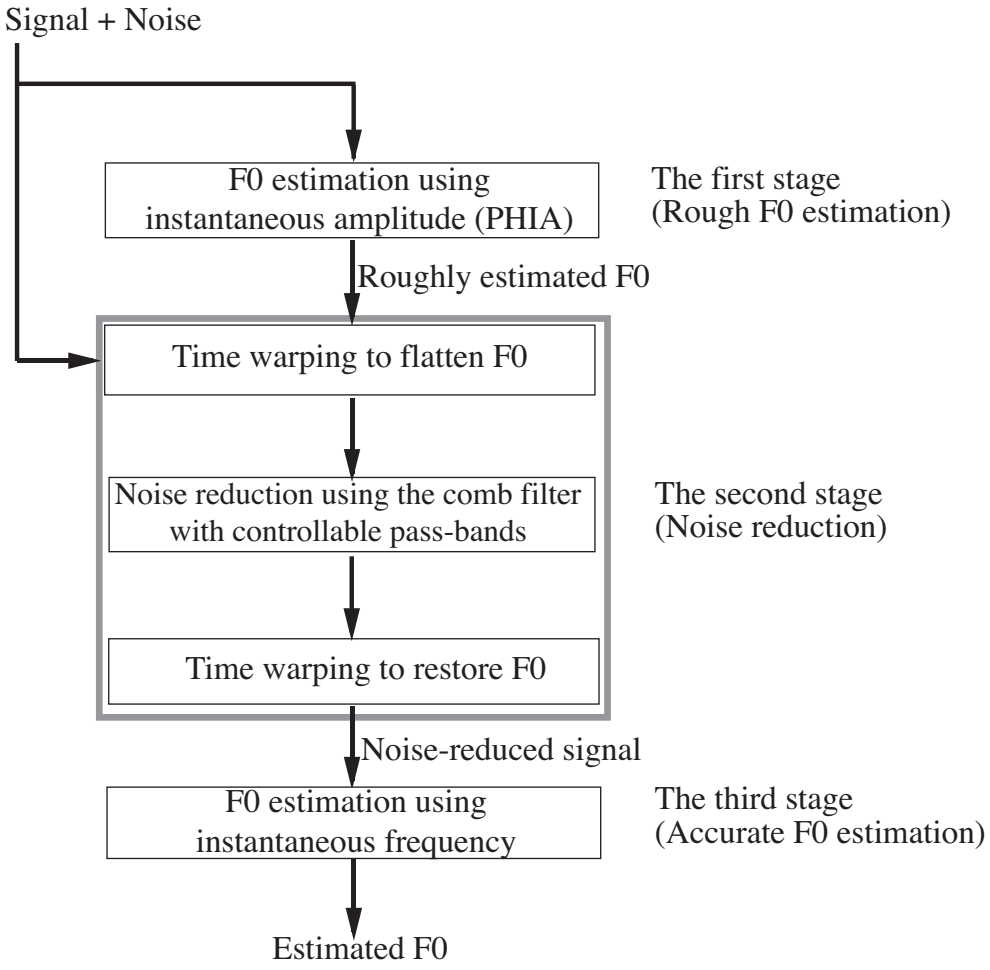

Fig. 1. Block diagram of the proposed method.

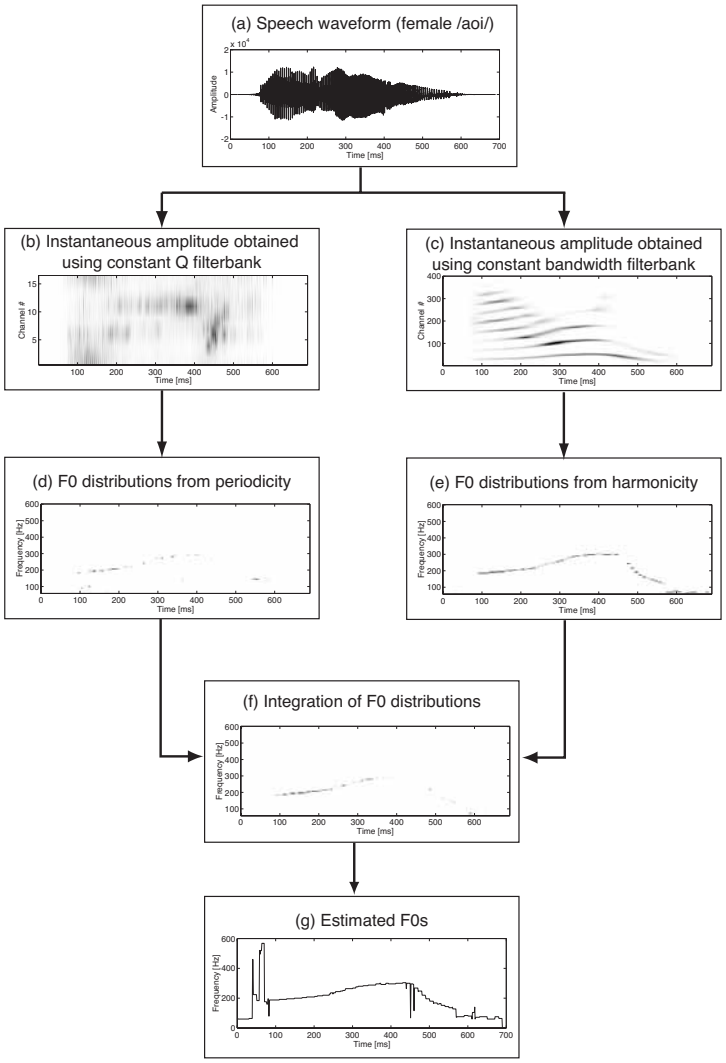

Fig. 2. F0 estimation based on the periodicity and harmonicity of instantaneous amplitude (PHIA). (a) A female speech waveform /aoi/. (b) Time-frequency representation of the instantaneous amplitude of speech by using a constant Q filterbank. The periodicity appears in the time domain. (c) Time-frequency representation of the instantaneous amplitude of speech by using a constant bandwidth filterbank. The harmonicity appears in the frequency domain. (d) Time-frequency plot of F0 distributions obtained from the periodicity by allocating the F0 distributions in temporal order. The shading indicates peaks. (e) Time-frequency plot of F0 distributions obtained from the harmonicity. (f) Time-frequency plot of the integrated F0 distributions. (g) Estimated F0s.

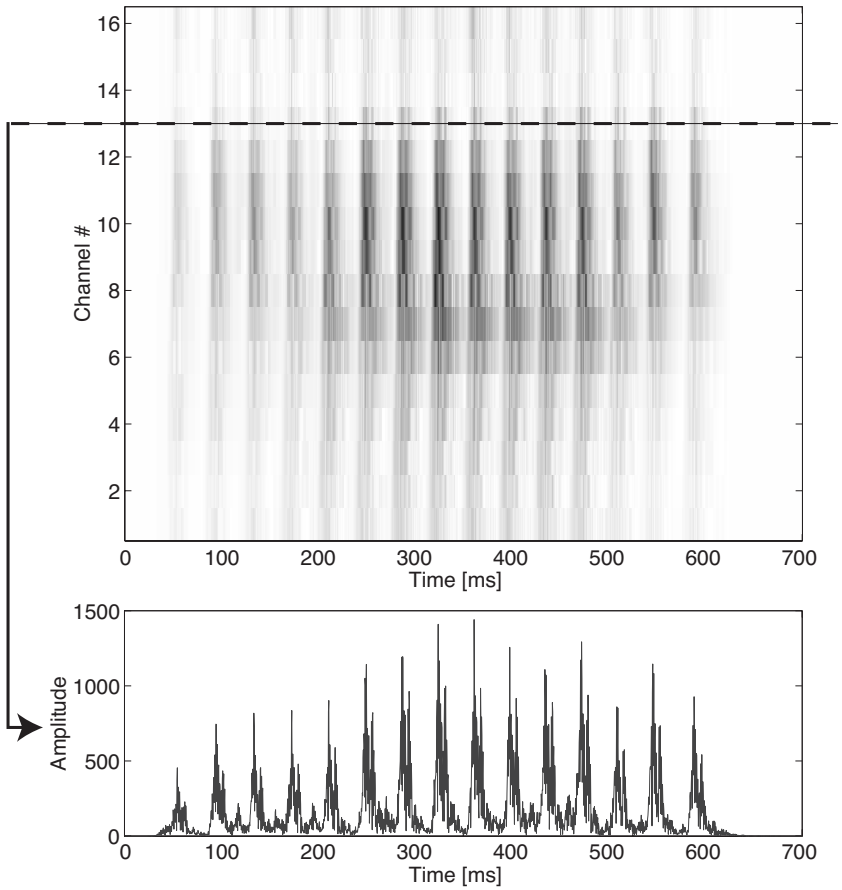

Fig. 3. Periodicity of instantaneous amplitude for male vowel /a/ in the time-frequency domain. Bottom panel shows the instantaneous amplitude in the time domain on channel #13 (dashed line in top panel).

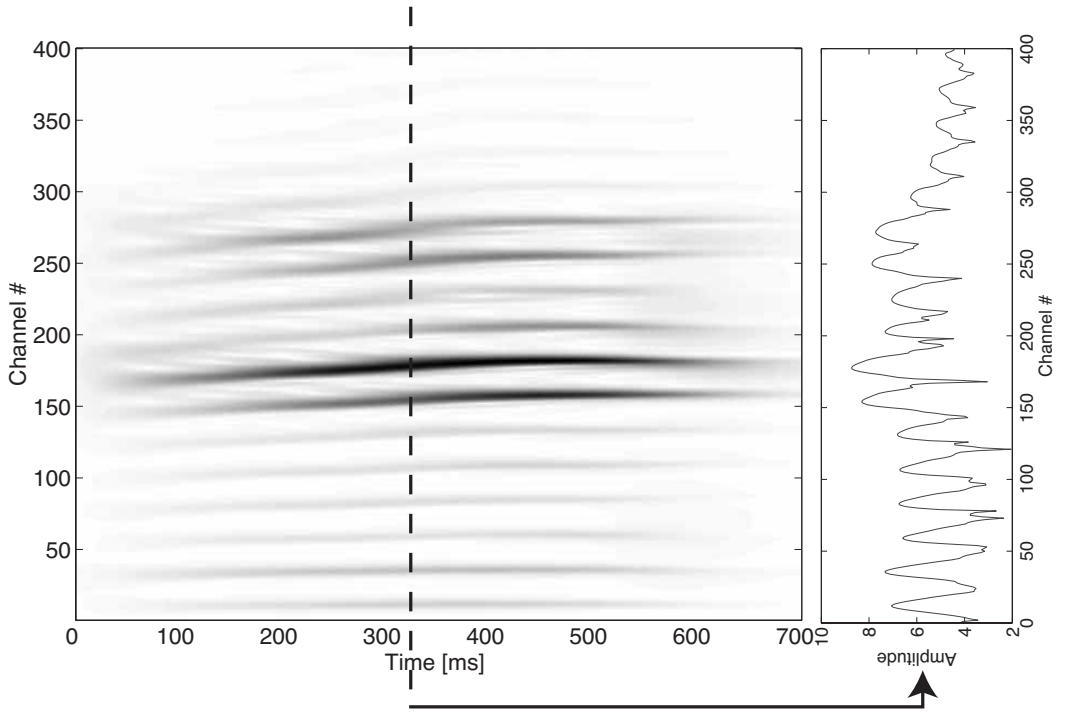

Fig. 4. Harmonicity of instantaneous amplitude for the male vowel /a/ in the time-frequency domain. Right panel shows the instantaneous amplitude in the frequency domain on 320 ms (dashed line in left panel).

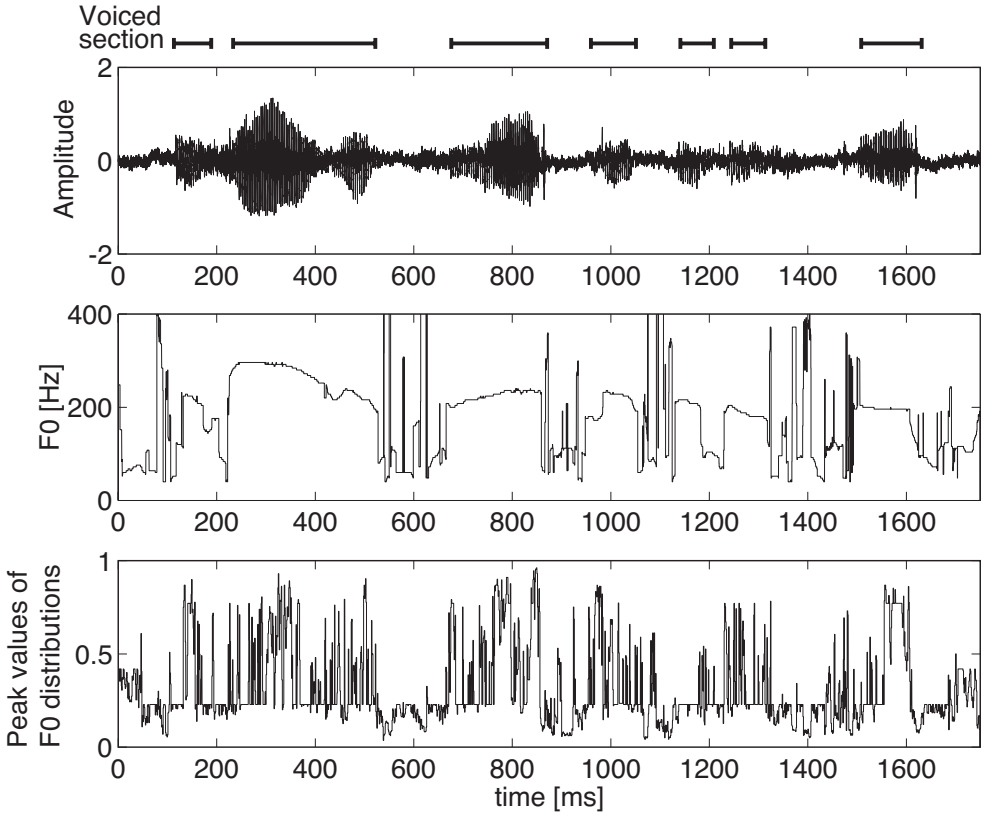

Fig. 5. F0s estimated by using PHIA. Top panel: female speech with pink noise (SNR is 10 dB); middle panel: F0s estimated by using PHIA; bottom panel: the peak values of F0 distributions obtained by PHIA. The line that is close to 0.2 in the peak values is due to the offset of the histogram described in section 2.2.2.

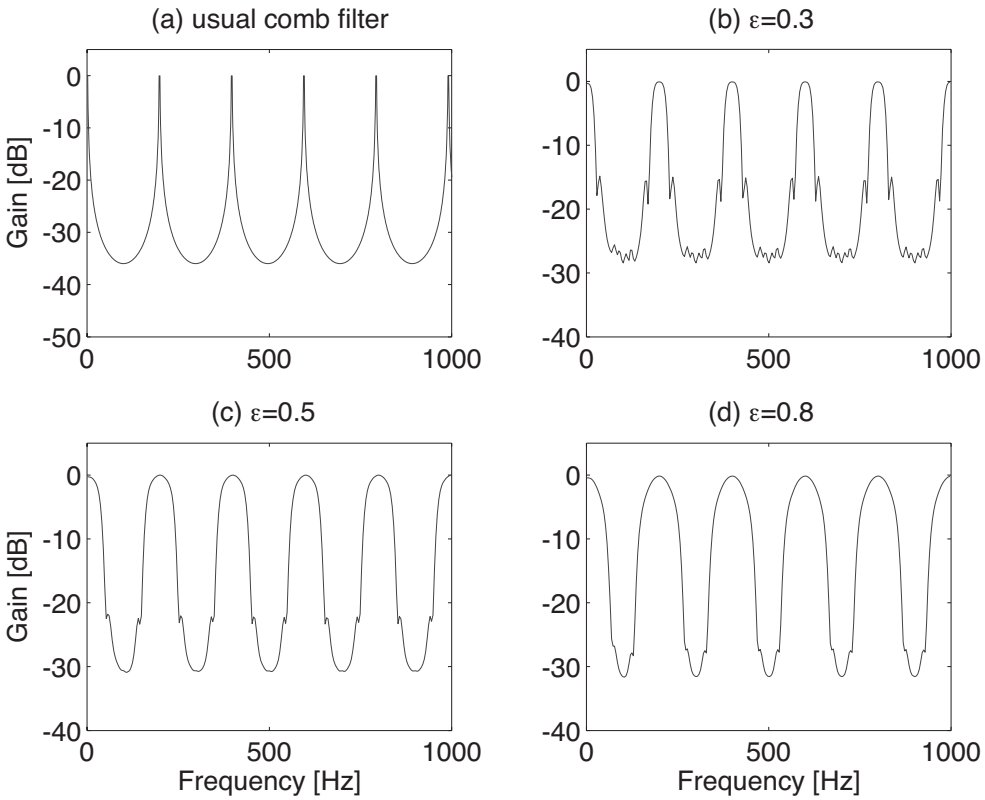

Fig. 6. Frequency responses in noise reduction by using a comb filter with controllable pass-bands.

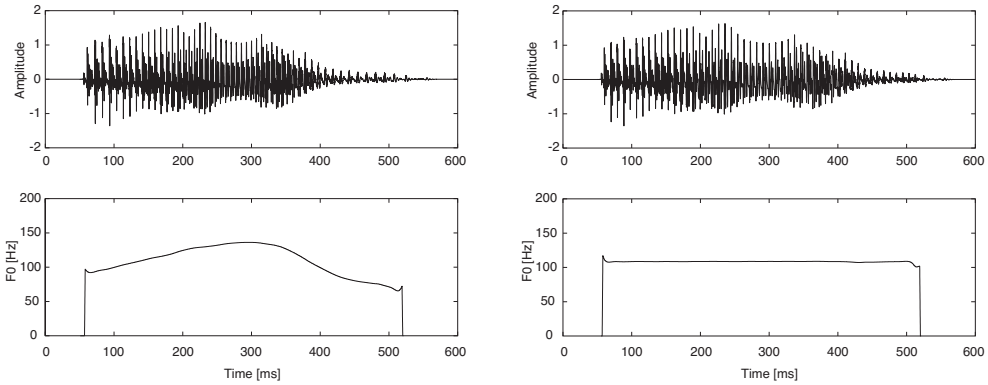

Fig. 7. Time-warping of speech waveforms to flatten F0s. Left panel: original speech and its F0s; right panel: time-warped speech and its F0s.

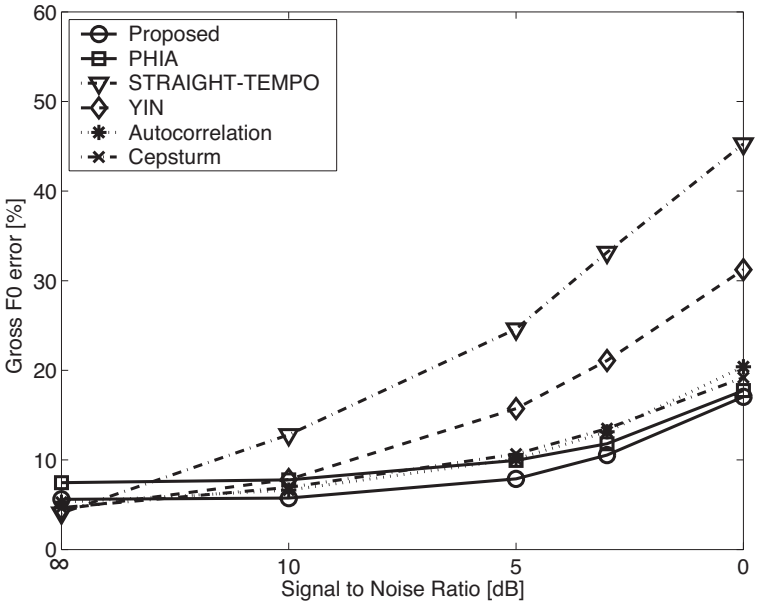

Fig. 8. Gross F0 error for speech with white noise.

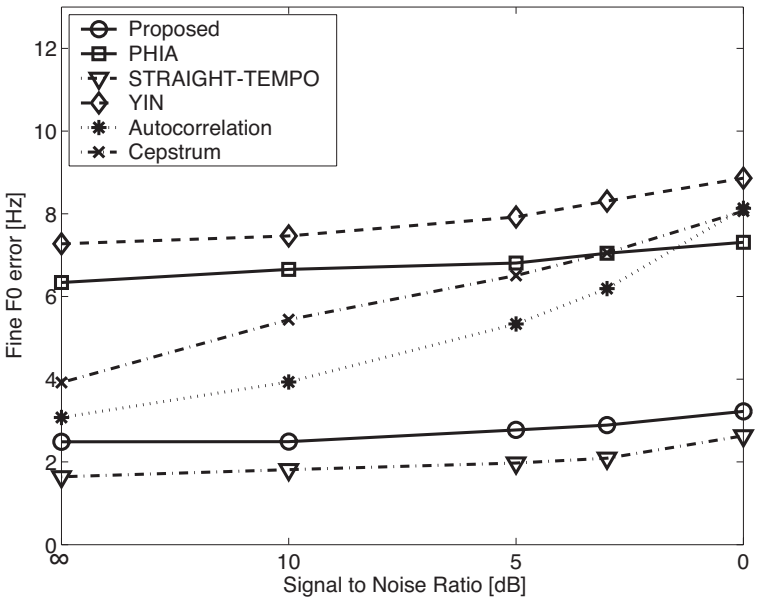

Fig. 9. Fine F0 error for speech with white noise.

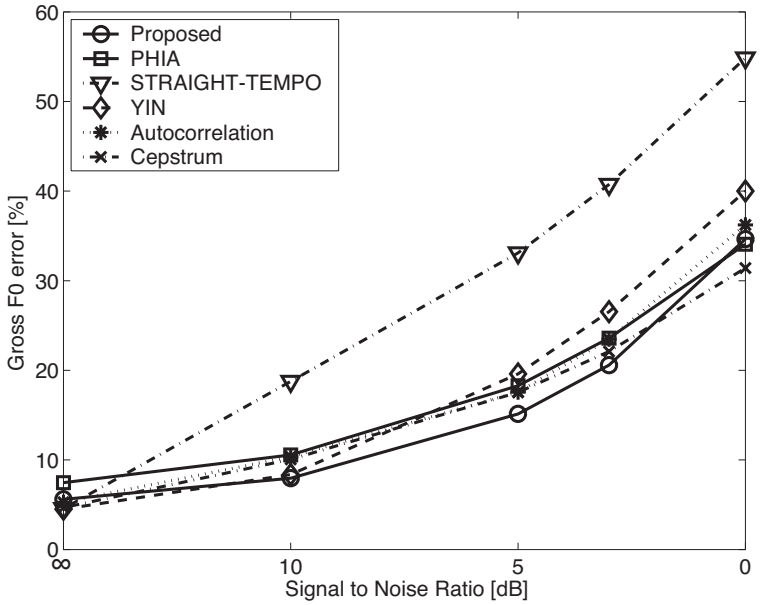

Fig. 10. Gross F0 error for speech with pink noise.

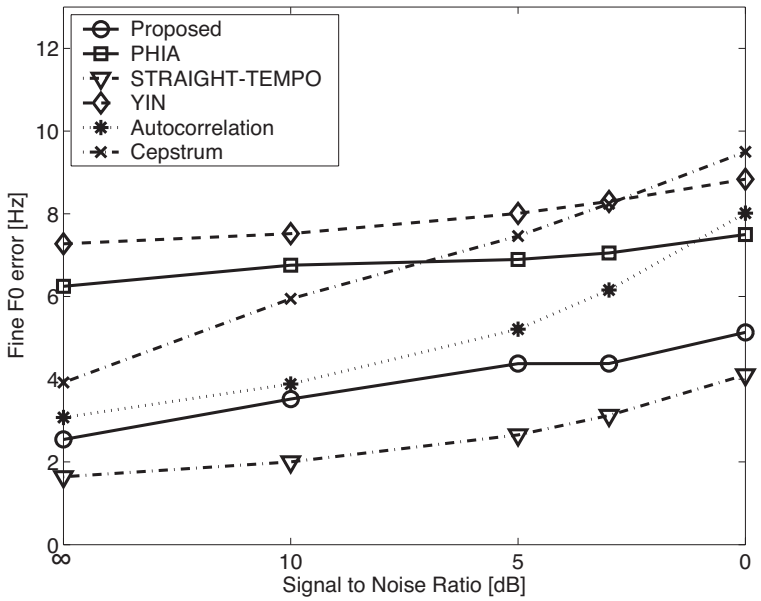

Fig. 11. Fine F0 error for speech with pink noise.