PAPER

A pitch detection method based on continuous wavelet transform for harmonic signal

Yoshifumi Chisaki¹, Hidetoshi Nakashima², Shuuichi Shiroshita²,
Tsuyoshi Usagawa² and Masanao Ebata³

¹*Faculty of Engineering, Kumamoto University,
2-39-1 Kurokami, Kumamoto, 860-8555 Japan*

²*Graduate School of Science and Technology, Kumamoto University,
2-39-1 Kurokami, Kumamoto, 860-8555 Japan*

³*Kumamoto National College of Technology,
2659-2 Suya, Nishigoshi, Kikuchigun, Kumamoto, 861-1102 Japan*

(Received 3 October 2001, Accepted for publication 2 July 2002)

Abstract: In order to track a rapid transient of pitch, a required frame length of some conventional pitch detection methods is too long. Although there are wavelet based pitch detection methods which require only a few periods of pitch for a frame, they are not robust enough against noise. This paper proposes a new pitch detection method which can work properly under noisy environments even if a frame duration is short. The proposed method consists of a power level detector, a signal analyzer, an autocorrelator, a voiced-unvoiced detector and a lag time interpolator. The signal analyzer is based on the continuous wavelet transform using a harmonic analyzing wavelet. Usage of the harmonic analyzing wavelet gives us more information about a pitch in a scalogram. Simulations of pitch detection for a harmonic chirp signal and speech signals are performed. Performances are compared with two conventional pitch detection methods, cepstrum and modified correlation methods. As a result, a performance of a pitch detection by the proposed method under a noisy environment is better than that of the other two conventional methods. In particular, the largest improvement of performance is obtained for male voices.

Keywords: Pitch detection, Continuous wavelet transform, Harmonic signal

PACS number: 43.72.Ar, 43.66.Hg

1. INTRODUCTION

Pitch information plays an important role for pitch-based speech processing systems, such as a speech rate converter, speech morphing, speech enhancer for a telephone line and so on [1-6]. A pitch detection based on a short duration is required to improve performance of the systems because a pitch frequency has a rapid transient in frequency. Some references mention that the duration for steady-state of speech is 10-20 ms and a dynamic range of pitch frequency in a short time is wider than an octave in frequency [1,2]. Simultaneously, a robustness against noise is also one of the essential issues in order to improve performance of pitch-based applications.

There are many papers which are related to pitch detection of speech since the 1960s [2,7]. In the literature, results of simulation show that over seven or eight pitch periods are required as a fixed frame length to perform

stable pitch detection for male speech; hence at least 70 ms as a duration for a frame is required when a pitch is 100 Hz.

Moreover, other references describe a pitch detection method based on a short duration not only for clean speech but also under noisy environments [1,3,8,9]. In particular, a pitch detection method based on instantaneous frequency proposed by Atake [9] improves a robustness against noise with a short duration over other methods based on instantaneous frequency. However, a duration for pitch detection requires over 40 ms when a pitch is 100 Hz, since a frame length for these methods corresponds to an analysis frequency. In view of a frame length for high performance pitch-based systems, a shorter duration is preferred to reduce a group delay. Thus, a development of a pitch detection method which can work with a short frame length and also has a robustness against noise is an important issue.

In studies based on autocorrelation [2,7], performance

of pitch detection is degraded under noisy environments because a peak of autocorrelation corresponding to a pitch is not enhanced against other peaks corresponding to noise. There are various techniques to decrease effects of noise in autocorrelation, e.g. low pass filtering in the cepstrum method and the modified correlation method. However, a robustness against noise for those techniques is insufficient when a frame length becomes shorter. In order to obtain further robustness against noise, we consider to use both harmonic components of speech and temporal information corresponding to a pitch in scalogram obtained by a continuous wavelet transform. The continuous wavelet transform is a suitable transform method because we can obtain much information about a pitch in a scalogram. It is expected that information about a pitch along each time and frequency axis can be obtained by using the following two major flexibilities for speech analysis. The first advantage is flexibility with respect to arrangements for resolution of both time and frequency. A second one is a flexible selection of a transform kernel which is called an analyzing wavelet or a mother wavelet. The analyzing wavelet plays an important role to decide characteristics of wavelet transform. The only restriction for an analyzing wavelet is admissible condition. Thus, there is a large variety in selection of mathematical function as an analyzing wavelet when inverse wavelet transform is neglected. In order to extract pitch information, a function based on a harmonic structure is selected as an analyzing wavelet because our target signal, a speech signal, has a harmonic structure. In a pitch detection method based on an autocorrelation, pitch candidates are selected from local peaks in autocorrelation. As studies for a pitch detection of a music signal [10,11] show, a manipulation of relative power enhancement for the local peak corresponding to a pitch is important even though each harmonic component of speech has a perturbation of frequency. Power for each high order harmonic component is summed up at a pitch frequency in the scalogram because the characteristics of frequency for an analyzing wavelet which has a harmonic analyzing wavelet are the same as that for a comb filter. Thus, it is expected that a relative power level at a fundamental frequency obtained by wavelet transform using the harmonic analyzing wavelet is higher than that obtained by wavelet transform using a Gabor function [3,9].

In addition, another characteristic of continuous wavelet transform based on the harmonic analyzing wavelet gives one more benefit to get a robustness against noise. The continuous wavelet transform brings some series of local peaks with an interval of a pitch period along the time axis in a scalogram. Therefore, gathering pitch period information from pitch interval series contributes to enhanced power of a local peak in a short time. Moreover,

the series of local peaks along the time axis contributes to the detection of a pitch when the power level at a fundamental frequency is insufficient; namely, a missing fundamental signal, such as a speech signal over telecommunication lines.

In this paper, a pitch detection method based on the continuous wavelet transform is proposed. The proposed method consists of a power level detector, a signal analyzer, an autocorrelator, a voiced-unvoiced detector and a lag time interpolator. In the signal analyzer block, the continuous wavelet transform is used to obtain both a pitch detection based on a short duration and a robustness against noise. According to both a relative enhancement of power level and some series of local peaks, a pitch detection based on a short duration and a robustness against noise is expected simultaneously.

First, a procedure for the proposed method is described. In particular, detailed advantages for the continuous wavelet transform using a harmonic analyzing wavelet are mentioned. Secondly, parameters for the proposed scheme are discussed by a computer simulation. The proposed method is compared with the conventional modified correlation method and the cepstrum method in all simulations. In order to confirm a performance of pitch detection, a harmonic chirp signal is used as an input signal. Moreover, results of simulations with respect to a robustness against noise in case of speech with the addition of white or pink noise are shown. Finally, a performance of the proposed method is concluded by those results compared with the modified correlation method and the cepstrum method.

2. PROCEDURE OF PITCH DETECTION

In this section, a procedure of pitch detection is mentioned. Figure 1 shows a block diagram of the proposed method. The pitch detection method consists of 5 blocks; a power level detector, a signal analyzer, an autocorrelator, a voiced-unvoiced detector and a lag time interpolator. The pitch detection is performed frame by frame with T_{shift} .

STEP 1:

In the first stage, the power level detector works as a switch for turning the pitch detection process on and off. The pitch detection is performed when the following equations are satisfied,

$$\begin{aligned} 10 \log(\bar{P}) &> P_{\text{TH}}, \\ P &> 0, \end{aligned} \quad (1)$$

where $\bar{P}$ is average power level with respect to time in a frame, and P_{TH} is a threshold derived from average power level of background noise.

Fig. 1 Block diagram of pitch detection.

STEP 2:

Wavelet transform is performed in this stage. Since the admissible condition is the only restriction for the analyzing wavelet, there is a large variety in selection of mathematical functions when the inverse wavelet transform is neglected. The Gabor function is often used as the analyzing wavelet for the continuous wavelet transform [3,9]. Figures 2(a) and 2(b) show an example of a frequency response for the Gabor function and the proposed analyzing wavelet, respectively. As shown in Fig. 2(a), the frequency response of the Gabor function is regarded as that of a band pass filter. Since the lag time for a pitch candidate in an autocorrelation is chosen by searching for a peak of maximum value, a relative enhancement of a local peak gives us a robustness against noise. In order to bring information about a pitch in a scalogram, we propose to make use of an analyzing wavelet which has information of high order harmonic components. Since a frequency response of the proposed

Fig. 2 Example of frequency response. (a): Gabor function. (b): Harmonic Analyzing Wavelet (Proposed).

Fig. 3 Illustration of characteristics of harmonic wavelet transform.

analyzing wavelet, as shown in Fig. 2(b), is the same as that of a comb filter, the power at harmonic components is summed up at the pitch frequency by the proposed analyzing wavelet, as illustrated in Fig. 3. P_{g0} and P_{h0} at f_0 shown in Fig. 3 denote power levels by the Gabor function and the proposed analyzing wavelet, respectively. Therefore, it is expected that the relative enhancement of power by the summation of power for harmonic components brings a robustness against noise.

Moreover, the proposed method gives other useful information about a pitch to a scalogram. The information is that a series of local peaks at an interval of pitch period appear along a time axis (i.e. shift axis) in the scalogram. The ripples which correspond to a pitch period along a time axis appear in a scalogram not only at a pitch frequency but also at other frequencies. Mismatching of phase at each harmonic component gives only a modification of a pattern

for the ripple, and an error ratio for pitch estimation caused by the mismatching of phase is less than 2% when a duration for an analyzing wavelet is longer than a duration corresponding to 3 fundamental periods [12].

Hereafter, the proposed analyzing wavelet and the proposed method is called Harmonic Analyzing Wavelet and Harmonic Wavelet Transform (HWT), respectively.

A general form of the wavelet transform is expressed as follows,

$$(W_{\psi}f)(b, a) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{|a|}} \psi^*\left(\frac{x-b}{a}\right) f(x) dx, \quad (2)$$

where

$\psi(x)$: analyzing wavelet,

$f(x)$: input signal,

a : scale parameter,

b : shift parameter,

x^* : conjugate of x .

And the analyzing wavelet $\psi(x)$ must satisfy the following admissible condition,

$$\int_{-\infty}^{\infty} \psi(x) dx = 0. \quad (3)$$

In order to apply a harmonic function to the analyzing wavelet, let us assume a part of Eq. (2) as follows,

$$g(b, a, x) = \frac{1}{\sqrt{a}} \psi^*\left(\frac{x-b}{a}\right). \quad (4)$$

Since the inverse of the scale parameter $1/a$ and shift parameter b correspond to parameters for frequency f and time t in scalogram, respectively, Eq. (2) can be expressed by using Eq. (4) with t and f as follows,

$$W(t, f) = \int_{-\infty}^{\infty} g(t, f, x) f(x) dx. \quad (5)$$

A harmonic structure is adopted to the function $g(t, f, x)$. In addition, a factor $1/\sqrt{a}$, shown in Eq. (4), should also be taken into account. The function $g(t, f, x)$ is derived as follows: we define a basic function $h(t, f, x)$ which has the characteristics of a harmonic structure as follows.

$$h(t, f, x) = \left(\sum_{k=1}^n \alpha_k e^{j(2\pi f k(x-t) + \phi_k)} \right) \cdot w(x), \quad (6)$$

$$-\frac{N}{2f} \leq x \leq \frac{N}{2f}, \quad (7)$$

where α_k and ϕ_k denote amplitude and phase at k -th order harmonic component, respectively. f is a fundamental frequency for analysis and n is a number of harmonics. As higher harmonic components do not keep a strict harmonic structure [13], a value of n is set up to 20 or a number where a frequency of the highest harmonic components does not exceed the Nyquist frequency. $w(x)$ is Gaussian

Fig. 4 Examples of analyzing wavelets.

window function. The window is designed to give a localization in a duration of N periods for a fundamental frequency. In order to obtain the highest performance, α_k and ϕ_k should be the same as those of a target signal. However, it is difficult to obtain suitable values of α_k and ϕ_k frame by frame because the target signal is speech. Therefore, each fixed value is set to α_k and ϕ_k in this paper. It is confirmed that the performance is sufficient for pitch detection, as shown in this paper and other literature [12]. In addition, a combination of values of α_k and ϕ_k gives a variety of patterns of ripple waveform, a period of ripples corresponding to a pitch period always appears with the highest power in any values of α_k and ϕ_k [12].

Finally, an analyzing wavelet $g(t, f, x)$ is expressed with a factor for normalization of the function $h(t, f, x)$ and a factor $\sqrt{f}$ instead of $1/\sqrt{a}$ shown in Eq. (4) as follows.

$$g(t, f, x) = \frac{\sqrt{f}}{\|h\|^2} h(t, f, x) \quad (8)$$

where $\| \|^2$ is a L_2 norm.

Examples of the proposed analyzing wavelet are shown in Fig. 4. Each waveform of the analyzing wavelet, from the left side in Fig. 4, has a relation of doubled analysis frequency. As the figure shows, Eq. (8) satisfies the admissible condition shown in Eq. (3).

STEP 3:

To determine a pitch period candidate, a scalogram is converted by two autocorrelations. First, the autocorrelation is calculated along a time axis in a scalogram to synchronize a lag time at each frequency. Second, the autocorrelation is executed to determine a final candidate for pitch period.

This scheme is similar to that of human pitch perception models, such as the Meddis-Hewitt model which is based on psychophysical studies [14,15]. As a first process, a conversion from a scalogram into an autocorrelation domain is executed by Eq. (9).

$$R(\tau, f) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \{W(t, f)W(t + \tau, f)\} dt, \quad (9)$$

where $W(t, f)$ is a set of transformed values by wavelet transform. For calculations on a computer, a null is padded to $W(t, f)$ in a duration except for a frame of signal.

As a second process, a conversion from $R(\tau, f)$ to $R(\tau)$ is performed by integration to enhance local peaks for pitch candidates relatively. $R(\tau)$ is defined as integrated autocorrelation, and can be expressed as follows.

$$R(\tau) = \int_{F_L}^{F_H} R(\tau, f) df. \quad (10)$$

where F_L and F_H are the lowest and the highest frequencies for pitch analysis interval, respectively. Parameter τ is a lag time.

A value of the lag time τ where a maximum value of $R(\tau)$ locates in a range of $1/F_H \leq \tau \leq 1/F_L$ corresponds to a pitch period.

STEP 4:

In the process of searching for a local peak for a pitch candidate, a peak of maximum value in an integrated autocorrelation domain is detected in both cases of a voiced or an unvoiced frame. Therefore, a process for judgment of voiced-unvoiced is required. As values of $R(\tau)$ at local peaks correspond to the power level of harmonic components of the observed signal, the following condition is defined to distinguish whether a frame is voiced or unvoiced.

$$20 \log \left\{ \frac{\max(R(\tau); 1/F_H \leq \tau \leq 1/F_L)}{R(0)} \right\} \geq L_{TH}, \quad (11)$$

where $\max(\cdot)$ is a function which gives a maximum value, and L_{TH} is a threshold. L_{TH} is defined by a preliminary experiment.

STEP 5:

In a final step, an interpolation of $R(\tau)$ with respect to a lag time is performed by cubic spline function to improve the resolution of lag time. The resolution of lag time is 10 times higher than that of the original. A final estimation of pitch is obtained by inverting the interpolated lag time.

3. SIMULATIONS

In this section, three simulations are performed. Scalogram, correlogram, and integrated autocorrelation are shown by using pseudo vowel in the first simulation. A harmonic chirp signal is used to show a basic performance in the second simulation. Finally, performance of pitch detection for speech is examined by using words and sentences uttered by females and males. Sampling frequency and resolution of quantization are 10 kHz and 16 bits, respectively. T_{shift} for a frame shift is set to 10 ms.

3.1. Confirmation of Each Process

In this subsection, each result is shown as scalogram, correlogram, and integrated autocorrelation. Pseudo vowel signals are used as an input signal whose fundamental frequency and order of harmonic components are set to 200 Hz and 10, respectively. Lower and higher frequencies for searching range, F_L and F_H , are set to 50 Hz and 500 Hz respectively, because a pitch for test signals is 200 Hz. A duration for frame length is set to 20.0 ms. N , as a number of fundamental periods to satisfy the admissible condition, is set to 3 and n , as a maximum number of harmonics for the analyzing wavelet, is set to 20. α_k and ϕ_k are set to 1 and 0, respectively.

A threshold parameter L_{TH} for judgment of voiced or unvoiced is obtained as follows. An unvoiced frame signal is simulated by adding noise to a frame of a voiced speech. White and pink noise are added to the pseudo vowel. Values of the left-hand side in Eq. (11) are calculated in combinations of vowel and noise. A value of the left-hand side corresponds to a degree of distortion of harmonic structure. The value becomes smaller as the harmonic structure is distorted. As a result, a range of the values is from -26.3 dB, for incorrect estimation, to -3.6 dB, for clean speech, in case of white noise. In case of pink noise, the left-hand side values are 1 dB greater than that in the case of white noise. According to the preliminary experiment, parameter L_{TH} is set to a value -24 dB for both white and pink noise.

Figure 5 shows a scalogram for the pseudo vowel /a/ by the proposed method without noise. Vertical and horizontal axes represent frequency and time (i.e. scale and shift in the wavelet transform), respectively. Peaks are located with around 5 ms intervals along the time axis. This interval corresponds to the fundamental frequency.

For picking up the information about a pitch, the autocorrelation in the scalogram is calculated. Figure 6 shows an autocorrelation defined by Eq. (9). Vertical and horizontal axes represent frequency and lag time for autocorrelation. In this figure, peaks are aligned along the frequency axis around 5 and 10 ms which corresponds to a pitch and a half pitch.

To obtain a pitch period candidate, the integrated autocorrelation is calculated. Figure 7 shows a result of integrated autocorrelation defined in Eq. (10). Vertical and horizontal axes represent the magnitude of integrated autocorrelation and lag time, respectively. As shown in Fig. 7, a maximum local peak in an interval $[1/F_H, 1/F_L]$ is located at 5.00 ms; namely, the estimated fundamental frequency is 200 Hz. This is the pitch period of the test signal. Figure 8 shows a result of integrated autocorrelation when white noise is added at SNR = 0 dB. As the figure shows, magnitudes of local peaks are degraded. In this case, a maximum magnitude is located at 2.50 ms; namely

Fig. 5 Scalogram for pseudo vowel /a/.

Fig. 6 Autocorrelation along time (shift) axis in scalogram.

Fig. 7 Integrated autocorrelation for pseudo vowel without noise.

Fig. 8 Integrated autocorrelation for pseudo vowel when a white noise is added at SNR = 0 dB.

an estimated pitch is 400 Hz.

3.2. Pitch Detection for Chirp Signal

In this subsection, basic characteristics of pitch detection are examined by using a harmonic chirp signal. The proposed method is compared with two conventional methods, the cepstrum method [7] and the modified correlation method [2]. The cepstrum method determines a pitch based on cepstral peaks. The modified correlation method is based on autocorrelation of LPC residual signal. These methods show better performances for clean speech; however, a long frame length, such as over 40 ms, is required to perform accurate pitch detection.

A condition of simulation using a harmonic chirp signal is as follows. Fundamental frequencies at onset and offset time are 50 Hz and 600 Hz, respectively. The fundamental frequency sweeps exponentially to simulate a rapid change in higher frequency. A number of harmonic components for the signal is set to four.

In the proposed method, parameters F_L , F_H in Eq. (10) are set to 30 Hz, 700 Hz, respectively, because a range of frequency for the harmonic chirp signal is set from 50 Hz to 600 Hz. L_{TH} is set to -24 dB from the result shown in the previous simulation. N and n are 20 and 3, respectively. Frame length is set to 20.0 ms.

In the conventional methods, a frequency range for pitch detection and length for the frame signal obtained from the observed signal is on the same conditions as that for the proposed method, respectively. Hanning window is used for pre-processing. In the cepstrum method, null padding is performed in order to improve frequency resolution. The padded frame length is 4 times longer than that of the original. Coefficients of cepstrum correspond to over 1.5 kHz is flattened by using the mean value of coefficients under 1.5 kHz, as shown in other papers [1,9,15].

Figure 9 illustrates the results of simulation by each method. Figures 9(a), 9(b) and 9(c) are obtained by the cepstrum method, the modified correlation method and the proposed method, respectively. Circle marks denote an estimated pitch, and dotted lines show a fundamental frequency of chirp signal. Horizontal axis represents frequency in logarithmic scale.

In case of the cepstrum method shown in Fig. 9(a), pitches are not detected correctly when a frequency is under 400 Hz. It is also shown that the resolution of the frequency is insufficient in higher frequency caused by a sampling rate.

As for the modified correlation method in Fig. 9(b), quadplex pitch frequency is estimated when the true pitch is under 147.8 Hz.

In case of the proposed method shown in Fig. 9(c), an error of a pitch detection occurs when a pitch is under

Fig. 9 Result of pitch detection when a duration for a frame is 20 ms. (a) Cepstrum method (CEP). (b) Modified correlation method (MOC). (c) the proposed method (HWT).

88.9 Hz.

By comparing these results, the proposed method performs better than that of the other two conventional methods. In particular, an improvement is apparent at lower frequencies. This is one of the benefits for the proposed method.

3.3. Simulation for Speech Signals

In this subsection, performance of a pitch detection is evaluated by both gross pitch error and fine pitch error. After the simulation, performance of a pitch detection with various frame lengths is discussed. Moreover, a robustness against noise is also discussed. Finally, processing time is examined.

3.3.1 Conditions for simulation and evaluation method

Speech signals, as input signals, are recorded in an anechoic room. A glottal vibration is recorded as a reference pitch simultaneously. Sampling rate is 48 kHz because a resolution of frequency for a standard pitch estimation depends on the sampling rate. A reference pitch is obtained by autocorrelation of a glottal waveform. As doubled and half pitches are sometimes estimated, final reference pitches are determined by human inspection. A Japanese female and male utter 2 sentences and 7 words. Sentences are /ashitano tenki wa nani desu ka?/ and

/fukuoka made no kippu ga hoshii no desuga/. Words are /ohayou/, /kumamoto/, /kagoshima/, /meiru/, /keikeitii/ (KKT), /tiikeiyuu/ (TKU), and /keieibii/ (KAB). A sampling rate of speech signal is converted from 48 kHz to 10 kHz for simulation.

Performance is evaluated by both gross pitch error and fine pitch error. In evaluation of gross pitch error, an evaluation of absolute frequency difference between estimated pitch and reference pitch is used in some references [16,17]. However, the range of pitch frequency for pitch detection is wide, such as from 50 Hz to 500 Hz. Therefore, gross pitch error based on a ratio is used as shown in some studies [1,18].

Let us assume that reference and estimated pitch frequencies are f_k and $\hat{f}_k$ for k -th frame. Then, a ratio of relative error is defined as follows,

$$\hat{e}(k) = \frac{|f_k - \hat{f}_k|}{f_k}, \quad (12)$$

where k denotes a frame index. Error frames are defined that the ratio of relative error $\hat{e}(k)$ is greater than 0.05, hence 5%. Gross pitch error (GPE) is calculated as follows,

$$GPE = \frac{N_{\text{error}}}{N_{\text{all}}} \times 100 (\%), \quad (13)$$

where N_{error} and N_{all} are the number of error frames and whole frames, respectively.

In evaluation of fine pitch error, standard deviation and mean error is calculated when the ratio of relative error $\hat{e}(k)$ is less than 0.05 as shown in some studies [19,20].

3.3.2 Length for analysis frame

This subsection examines how a frame length affects to errata of pitch analysis is examined. A frame length is varied as 15.0, 20.0, 25.6, 30.0, and 51.2 ms. In simulations, white and pink noise are added at SNR = 5 dB. In preliminary experiments, as for several values of SNR, performance of pitch detection is degraded extremely under SNR = 5 dB. Therefore, this value of SNR is selected as the lowest SNR in this paper. SNR is calculated in overall duration. The proposed method is compared with the cepstrum and the modified correlation methods. Figures 10(a) and 10(b) show results in case of female and male speech, respectively. Filled and open marks denote the results for white noise and pink noise, respectively. Circles, triangles and squares show results of the proposed (HWT), modified correlation (MOC) and cepstrum (CEP) methods, respectively. Vertical and Horizontal axes are gross pitch error and frame length, respectively. According to the results shown in both (a) and (b), performance of the proposed method is better than both those of the modified correlation method and the cepstrum method. In all cases, the worst performance is shown when the frame length is 15.0 ms. The reason for degradation of the performance is

(a) GPE vs. frame length in case of female speech.

(b) GPE vs. frame length in case of male speech.

Fig. 10 Results of pitch detection evaluated by Gross pitch error (GPE). (a) and (b) are in case of female and male, respectively. Circle, triangle and square represent harmonic wavelet transform (HWT), modified correlation (MOC) and cepstrum (CEP) methods, respectively. Filled and open marks denote results for white noise and pink noise, respectively.

that the power level for harmonic components in the lower frequency range cannot be obtained sufficiently as a frame length becomes shorter. Moreover, gross pitch error increases slightly until frame length is 20.0 ms even though SNR is 5 dB.

From a point of view at the type of noise, performance of the proposed method is a quite a bit better than that of the conventional methods, except for a case of pink noise for females. The reason why the performance for pink noise is the worst is that a power level of pink noise in a range of a pitch frequency is higher than that in a range of other frequencies.

According to these results, the proposed method shows the best performance among three methods. It is also shown that performance of the proposed method is not degraded when a frame length is over 20.0 ms.

3.3.3 Performance against noise

Figures 11(a) and 11(b) show performance of pitch detection in case of white and pink noise when a frame length obtained from observed signal is 20 ms, respectively. Three methods; the proposed (HWT), the modified correlation (MOC) and the cepstrum (CEP) are compared. Each result is denoted by circle, triangle and square mark, respectively. Filled and open marks denote the result for

(a) GPE vs. SNR in case of white noise.

(b) GPE vs. SNR in case of pink noise.

Fig. 11 Performance of pitch detection evaluated by Gross pitch error (GPE). (a) and (b) are in case of white noise and pink noise, respectively. Circle, triangle and square represent harmonic wavelet transform (HWT) method, modified correlation (MOC) method and cepstrum (CEP) method, respectively. Filled and open marks denote results for female and male, respectively.

females and males, respectively. Vertical and horizontal axes are gross pitch error and SNR, respectively. In both Figs. 11(a) and 11(b), performance of each method is degraded as SNR becomes smaller, except for the result of the cepstrum method for male. In each SNR, performance of the proposed method is better than other two conventional methods.

Furthermore, performance of the proposed method is examined by gross pitch error with some frequency bands each when a frame length obtained from observed signal is 20.0 ms. Figure 12 shows gross pitch error at each frequency when SNR is 15 dB and 30 dB, respectively. Gross pitch error is calculated at each center frequency from 100 Hz to 300 Hz with 50 Hz steps. The bandwidth at each frequency is 50 Hz. In Fig. 12, filled and open marks are for female and male speech, respectively. Circle and triangle marks represent the result for the proposed (HWT) and modified correlation (MOC) methods, respectively. Solid and dotted lines denote in case of SNR = 15 dB and 30 dB, respectively.

Performance cannot be compared between each frequency band because SNR at each frequency band is not the same. Performance at 100 Hz is not better than that

Fig. 12 Gross pitch error (GPE) at each center frequency from 100 Hz to 300 Hz with 50 Hz step for female and male speech in case of SNR = 15 dB and 30 dB. Filled and open marks are in case of female and male, respectively. Circle and triangle marks represent the proposed method and modified correlation method, respectively.

at other frequencies at each SNR in Fig. 12 because segmental SNR at 100 Hz is the lowest among SNR at each frequency band. In each case except for 300 Hz at SNR = 30 dB, performance of the proposed method is better than that of the modified correlation method at both SNR = 15 dB and 30 dB. Moreover, it is confirmed that the difference of performance between the proposed method and the modified correlation method at SNR = 15 dB is greater than that at SNR = 30 dB.

In addition, a reason for difference of performance between female and male in Figs. 11(a) and 11(b) can be considered as follows. The performance in Figs. 11(a) and 11(b) can be considered as sum of performance at each frequency band in Fig. 12. Therefore, performance of female is better than that of male because pitch frequency for female is higher than 187.6 Hz in this simulation.

Next, fine pitch error is examined. The condition of simulation for fine pitch error is the same as that for gross pitch error. Standard deviation and mean error for fine pitch error is calculated when the ratio of relative error $\hat{e}(k)$ is less than 0.05, as shown in some studies [19,20]. Table 1 shows the result of standard deviation of fine pitch error. Mean of fine pitch error is shown in Table 2. Performance of the proposed method with respect to standard deviation is not always better than that of the modified correlation method at each SNR, except for male. In a point of view from the mean value, it is also shown that performance of the proposed method is not always better than that of the conventional method.

According to the results, performance of the proposed method is better than conventional pitch detection methods with respect to gross pitch error.

3.3.4 Performance of processing time

Performance of processing time is examined. Conditions for hardware are as follows. CPU and memory size

Table 1 Standard deviation of fine pitch error when a frame length obtained from observed signal is 20.0 ms (Hz).

(a) female						
SNR	∞	30	20	15	10	5
HWT	3.03	3.15	3.22	3.30	4.10	4.71
MOC	2.85	3.18	3.35	3.36	3.95	4.91
(b) male						
SNR	∞	30	20	15	10	5
HWT	2.48	2.39	2.45	2.43	2.55	2.68
MOC	1.94	1.96	2.05	2.22	2.07	1.69

Table 2 Mean value of fine pitch error when a frame length obtained from observed signal is 20.0 ms (Hz).

(a) female						
SNR	∞	30	20	15	10	5
HWT	-0.57	-0.56	-0.50	-0.64	-0.40	-0.30
MOC	-0.84	-0.88	-0.72	0.31	1.45	3.54
(b) male						
SNR	∞	30	20	15	10	5
HWT	0.58	0.60	0.52	0.50	0.49	0.63
MOC	-0.29	-0.14	-0.10	0.18	0.74	0.46

are Pentium III (550 MHz) and 128 MB, respectively. In softwares, Operating System is linux (kernel 2.2.18) and gcc (version egcs-2.91.66) is used for the compiler. Performance is measured by off-line processing. Chirp signal, whose duration is 2.0 s, is used as an input signal. As a result, processing time is 120 s for the modified correlation method. In case of the proposed method, processing time is 8.6 times longer than the modified correlation method.

4. CONCLUSION

A pitch detection method based on the continuous wavelet transform for a harmonic signal is proposed. Basic characteristics of pitch detection are confirmed by using a harmonic chirp signal. In addition, a simulation of a pitch detection with added noise is performed. According to the results of all simulations, it is confirmed that the proposed method has advantages with respect to a frame length and a robustness against noise. In addition, it is mentioned that a long processing time is required when comparing to the conventional method. We plan to use temporal information and auditory phenomena of human in order to improve performance. Furthermore, we will discuss the adoption of this algorithm to parallel processing in order to improve performance of processing time.

ACKNOWLEDGMENTS

We would like to thank Prof. Adrianus J. M. Houtsma, IPO at the Eindhoven University of Technology, for comments, suggestions, and review.

Part of this research was carried out by the Ono Acoustics Research Fund (2001).

REFERENCES

- [1] T. Takagi, N. Seiyama and E. Miyasaka, "A method for pitch extraction of speech signals using autocorrelation functions through multiple window-lengths," *IEICE A*, **J80-A**, 1341–1350 (1997).
- [2] F. Itakura and S. Saito, "Speech information compression based on the maximum likelihood spectral estimation," *J. Acoust. Soc. Jpn. (J)*, **9**, 463–472 (1971).
- [3] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: VOCODER revisited," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1303–1306 (1997).
- [4] Y. Chisaki, T. Usagawa and M. Ebata, "Speech enhancements in multiple speaker condition using adaptive comb filter," *Proc. Inter-Noise 95*, Vol. II, pp. 1125–1128 (1995).
- [5] M. Nakai, H. Shimodaira and S. Sagayama, "Prosodic phrase segmentation based on pitch-pattern clustering," *Trans. IEICE*, **J77-A**, 206–214 (1993).
- [6] T. Yoshimura, S. Hayamizu, H. Omura and K. Tanaka, "Pitch pattern clustering of user utterances in human-machine dialogue," *ICSLP '96*, Vol. 2, pp. 837–840 (1996).
- [7] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, **22**, 293–309 (1967).
- [8] T. Abe, T. Kobayashi and S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," *Proc. ICSLP '96*, Vol. 2, pp. 1277–1280 (1996).
- [9] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura and K. Shikano, "Robust estimation of fundamental frequency using instantaneous frequencies of harmonic components," *IEICE D-II*, **J83-D-II**, 2077–2086 (2000).
- [10] R. Abe and N. Kambayashi, "Frequency estimation based on harmonic product spectrum method," *J. Acoust. Soc. Jpn. (J)*, **53**, 691–697 (1997).
- [11] R. Abe and N. Kambayashi, "An algorithm for fundamental frequency estimation using the harmonic spectral model," *J. Acoust. Soc. Jpn. (J)*, **54**, 296–304 (1998).
- [12] Y. Chisaki, T. Usagawa and M. Ebata, "Spectral analysis for inharmonic signal by signal specific analyzing wavelet," *Tech. Rep. IEICE*, EA98-95, pp. 29–36 (1998).
- [13] S. Kato and J. Miwa, "Pitch detection using moving average and band-limitation in cepstrum method and its application," *Tech. Rep. IEICE*, SP94-95, pp. 29–36 (1995).
- [14] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Am.*, **89**, 2866–2882 (1991).
- [15] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, **91**, 233–245 (1992).
- [16] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-24**, 399–418 (1976).
- [17] H. Ohmura and K. Tanaka, "Fine pitch contour extraction by voice fundamental wave filtering method," *J. Acoust. Soc. Jpn. (J)*, **51**, 509–518 (1995).
- [18] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-24**, 399–418 (1976).
- [19] N. Kunieda, T. Shimamura and J. Suzuki, "Pitch extraction by using autocorrelation function on the log spectrum," *IEICE A*, **J80-A**, 435–443 (1997).
- [20] A. Sasou and S. Nakamura, "A pitch extraction method using wavelet transform," *IEICE A*, **J80-A**, 1848–1856 (1997).