ELSEVIER

Signal Processing 44 (1995) 233-241

SIGNAL
PROCESSING

Fundamental frequency determination based on instantaneous frequency estimation

Lunji Qiu, Haiyun Yang, Soo-Ngee Koh*

School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Ave., Singapore 2263

Received 15 October 1993; revised 31 March 1994, 4 July 1994 and 14 November 1994

Abstract

For certain types of speech coding algorithms, it is important to determine accurately the pitch period or the fundamental frequency of the speech signals. A new fundamental frequency determination method based on the instantaneous frequency estimation is presented in this paper. A bank of two bandpass filters is used to eliminate and attenuate the harmonics of the speech signals. The fundamental frequency can be determined from the estimated instantaneous frequencies of the filtered speech signals. This new method can be used to obtain a continuous fundamental frequency as a function of time. Also, it does not seem to have the problem of occasional incorrect frequency estimation, such as frequency doubling or halving.

Zusammenfassung

Für gewisse Typen von Sprachkodieralgorithmen ist es wichtig, die Grundfrequenz oder die Grundperioden von Sprachsignalen genau zu bestimmen. In dieser Arbeit wird eine neue Methode zur Bestimmung der Grundfrequenz vorgestellt, die sich auf die Schätzung der Momentanfrequenz abstützt. Es wird eine Filterbank von zwei Bandpaßfiltern zur Eliminierung und Dämpfung der Harmonischen des Sprachsignals benutzt. Die Grundfrequenz kann aus den geschätzten Momentanfrequenzen des gefilterten Sprachsignals bestimmt werden. Diese neue Methode kann gebraucht werden, um eine stetige Grundfrequenz als Zeitfunktion zu erhalten. Zusätzlich scheint es keine Probleme gelegentlicher inkorrekter Frequenzschätzung, wie Frequenz-verdopplung oder -halbierung, zu geben.

Résumé

Pour certains types d'algorithmes de codage de la parole, il est important de déterminer de manière précise la période de vibration ou la fréquence fondamentale des signaux de parole. Une méthode nouvelle de détermination de la fréquence fondamentale basée sur l'estimation de la fréquence instantanée est présentée dans cet article. Un banc de deux filtres passe-bande est utilisé pour éliminer et atténuer les harmoniques des signaux de parole. La fréquence fondamentale peut être déterminée à partir des fréquences instantanées estimées des signaux de parole filtrés. Cette méthode nouvelle peut être utilisée pour obtenir une fréquence fondamentale continue fonction de temps. De plus, cette méthode ne semble pas

* Corresponding author. Tel.: (65) 7995429. Fax: (65) 7912687.

SSDI 0165-1684(95)00027-5

présenter de problème d'estimation de fréquence incorrecte occasionnelle tel que doublement ou division par deux de la fréquence.

Keywords: Instantaneous frequency; Fundamental frequency

1. Introduction

The accurate representation of the pitch period or fundamental frequency is required in many applications, such as speech coding, speech synthesis, speech and speaker recognition. Most of the fundamental frequency determination algorithms [3, 6] in either the time or frequency domain are based on the assumption that the speech signals are locally stationary. As a result, the conventional pitch period or fundamental frequency determination algorithms sometimes fail to correctly estimate the pitch period since speech signals are strictly speaking naturally nonstationary in characteristics.

To represent the nonstationarity of the speech signals, the time-frequency analysis techniques are suitable tools since the time-frequency signal analysis techniques represent signals simultaneously in both time and frequency, and there is no assumption of stationarity. Instantaneous frequency (IF) is an important concept in time-frequency analysis. The IF is a function of time and is a measure of the location in frequency corresponding to a particular time component of the signal.

Research results so far show that the IF can only be estimated for monocomponent signals. Before we can estimate the IF of the speech signals, the speech signals have to be somehow transformed to a monocomponent or nearly monocomponent signal which must contain the fundamental frequency of the speech signals. In other words, the harmonics of the speech signals have to be largely attenuated or even eliminated. In general, it is difficult to use a fixed bandpass filter to perform this function satisfactorily because the speech fundamental frequency changes with time and gender. A time-varying bandpass filter is necessary to preserve the fundamental frequency, while significantly attenuating the harmonics of speech signals. To design a satisfactory time-varying filter, the a priori knowledge of the fundamental frequency of the speech signal is needed. However, such an a priori knowledge is not available. Therefore, the use of a bank of bandpass filters to cover the complete range of pitch frequency of speech signals is considered instead of using a time-varying filter.

In this paper, a new fundamental frequency determination method based on the IF estimation is presented. A bank of two bandpass filters has been used in our study to filter speech signals to monocomponent or nearly monocomponent signals. The IF estimation techniques may then be used to estimate the fundamental frequency to obtain the fundamental frequency of the speech signals.

2. Instantaneous frequency estimation

The IF of a signal can be uniquely defined by using its corresponding analytic signal. The IF of a time-varying signal can be estimated in various ways [1, 4]. In this paper, the smoothed central finite difference (CFD) discrete IF (DIF) estimator is used.

For a given real signal, $s(t)$, a unique representation may be obtained if the Hilbert transform is used to generate the complex signal [7]. The IF of a signal $s(t)$ can then be uniquely defined by using the analytic signal. This is formally expressed as

$$f_i(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t}, \tag{1}$$

where $f_i(t)$ is the IF of the signal, $\phi(t)$ is the phase of the analytic signal, $z(t) = a(t)\mathrm{e}^{\mathrm{j}\phi(t)}$, which is formed from the real signal, $s(t)$.

The definition of IF given by Ville [7] may be extended to the discrete time signal case, with the discrete version of the IF referred to as DIF. The corresponding estimator, based on the CFD of the phase, was defined by Claasen and

Mechlenbrauker [1] as

$$f_i(n) = \frac{f_s}{4\pi}\left\{\arctan\frac{\mathrm{Im}[z(n+1)]}{\mathrm{Re}[z(n+1)} - \arctan\frac{\mathrm{Im}[z(n-1)]}{\mathrm{Re}[z(n-1)]}\right\}_{\mathrm{mod}\,2\pi}, \tag{2}$$

where f_s is the sampling frequency. The notation mod 2π represents a modulo 2π operation. $\mathrm{Im}[z(n)]$ and $\mathrm{Re}[z(n)]$ are respectively the imaginary and real parts of $z(n)$, and arctan is the principal value of the inverse tangent.

A smoothing window may be applied to the CFD DIF estimation to reduce the dispersion at the expense of time resolution. Since the CFD DIF estimation is a periodic phase signal (it is modulo $f_s/2$, where f_s is the sampling frequency), a modulo convolution with a smoothing window must be used [4]. Let $f_i(n)$, which is modulo $f_s/2$, be the DIF estimator and let $h(n)$ be a smoothing window function of odd length P. Then the smoothed DIF estimators are defined by the modulo convolution operation given below:

$$f_i^{\mathrm{s}}(n) = \frac{f_s}{4\pi}\arg\left[\sum_{p=-(P-1)/2}^{(P-1)/2} h(p)\,\mathrm{e}^{\mathrm{j}4\pi f(n-p)/f_s}\right]_{\mathrm{mod}\,f_s/2} \tag{3}$$

3. Fundamental frequency determination algorithm

In this section, the fundamental frequency determination algorithm based on IF estimation is presented. The algorithm involves the following steps.

(1) elimination and attenuation of the harmonics,
(2) estimation of the DIFs,
(3) voiced or unvoiced and silence decisions,
(4) decision of the fundamental frequency from two estimated DIFs.

3.1. Elimination and attenuation of harmonics

A bank of two bandpass filters with different bandwidths and center frequencies is used. To attenuate the harmonics of speech signals while preserving the time-varying pitch frequency, the $L^2(R)$ space is decomposed into a set of orthogonal spaces corresponding to differences in scale to achieve the filters.

A single prototype function is constructed from the sigmoidal function [5] by the following equation:

$$\psi(t) = \varphi(t+1) - \varphi(t-1), \tag{4}$$

where $t \in R$, $\varphi(t) = s(t+1) - s(t-1)$ and $s(t) = (1 + \mathrm{e}^{-2t})^{-1}$.

The basis filter functions are obtained from (4) as $\psi_{2^j}(t) = 2^{-j/2}\psi(t/2^j)$. The filter functions $\psi(t/2^j)$ are shown in Fig. 1. The Fourier transform of the basis functions, $\psi_{2^j}(t)$, is

$$\Psi_{2^j}(\omega) = -\mathrm{i}\frac{2\pi 2^{j/2}\sin^2(2^j\omega)}{\sinh(2^{j-1}\pi\omega)}, \tag{5}$$

which is shown in Fig. 2.

As the fundamental frequency of most speech signals is within the frequency range of 50–500 Hz, the sampling frequency for fundamental frequency determination could be as low as 1 KHz to reduce the density of the computation in the proposed algorithm. Since the information above 500 Hz is not needed for the detection of the fundamental frequency, a lowpass filter with cutoff frequency $f_c = 500$ Hz may be used to avoid the problem of alising before down sampling.

It can be observed in Fig. 2 that the Fourier transform of the bandpass filter at scale $j = 3$, has the frequency range of approximately 10–500 Hz when the sampling frequency, f_s, is 8000 Hz. This bandpass filter can eliminate the harmonics which are higher than 500 Hz and attenuate the other harmonics below 500 Hz. It was found through simulations that for speech signal of which fundamental frequency is in the range of 110–500 Hz, the bandpass filter at this scale is a suitable one.

However, if the fundamental frequency of the speech signal is very low, say less than 110 Hz or around 110 Hz, the first harmonic of such speech signal is less than or around 220 Hz. In this case, the first harmonic cannot be eliminated, nor attenuated by the use of the bandpass filter at scale

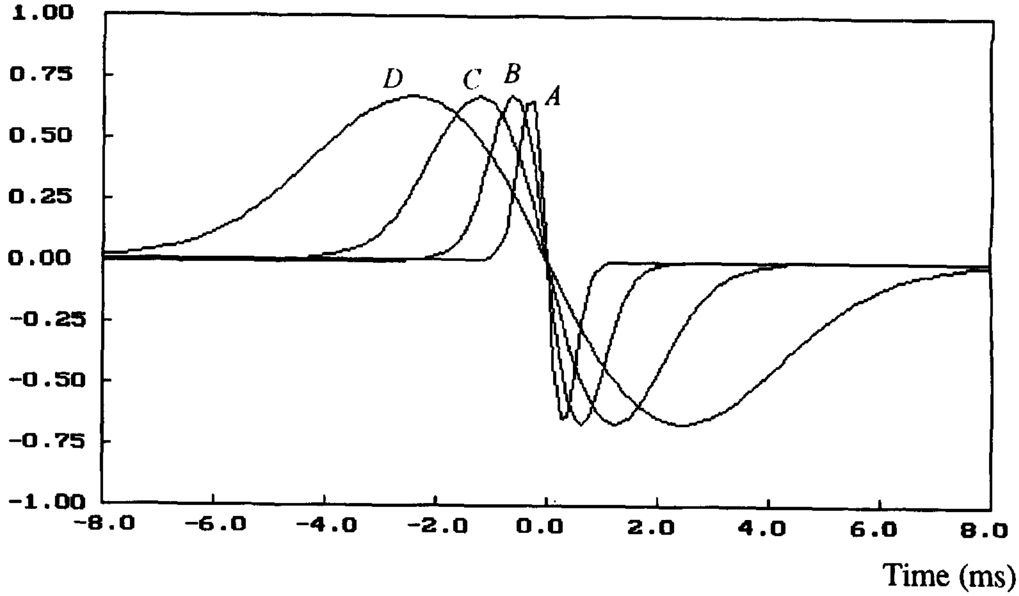

Fig. 1. Filter function $\psi_{2^j}(t)$ (A: $j = 1$; B: $j = 2$; C: $j = 3$; D: $j = 4$).

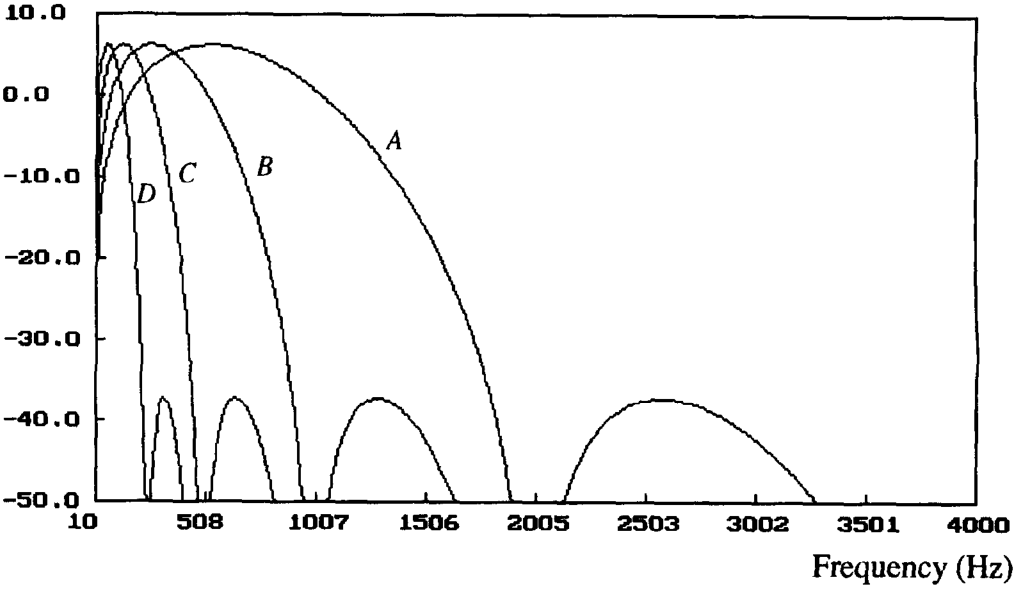

Fig. 2. The Fourier transform of filter function $\psi_{2^j}(t)$ (A: $j = 1$; B: $j = 2$; C: $j = 3$; D: $j = 4$).

$j = 3$. This will result in estimation errors in the DIF estimation since the filtered speech signal is not approximately monocomponent. To overcome this problem, the bandpass filter at scale $j = 4$ should be used to attenuate the first harmonic of the speech signal. From Fig. 2, we can see that the filter at scale $j = 4$ has the pass band from 10 to 250 Hz.

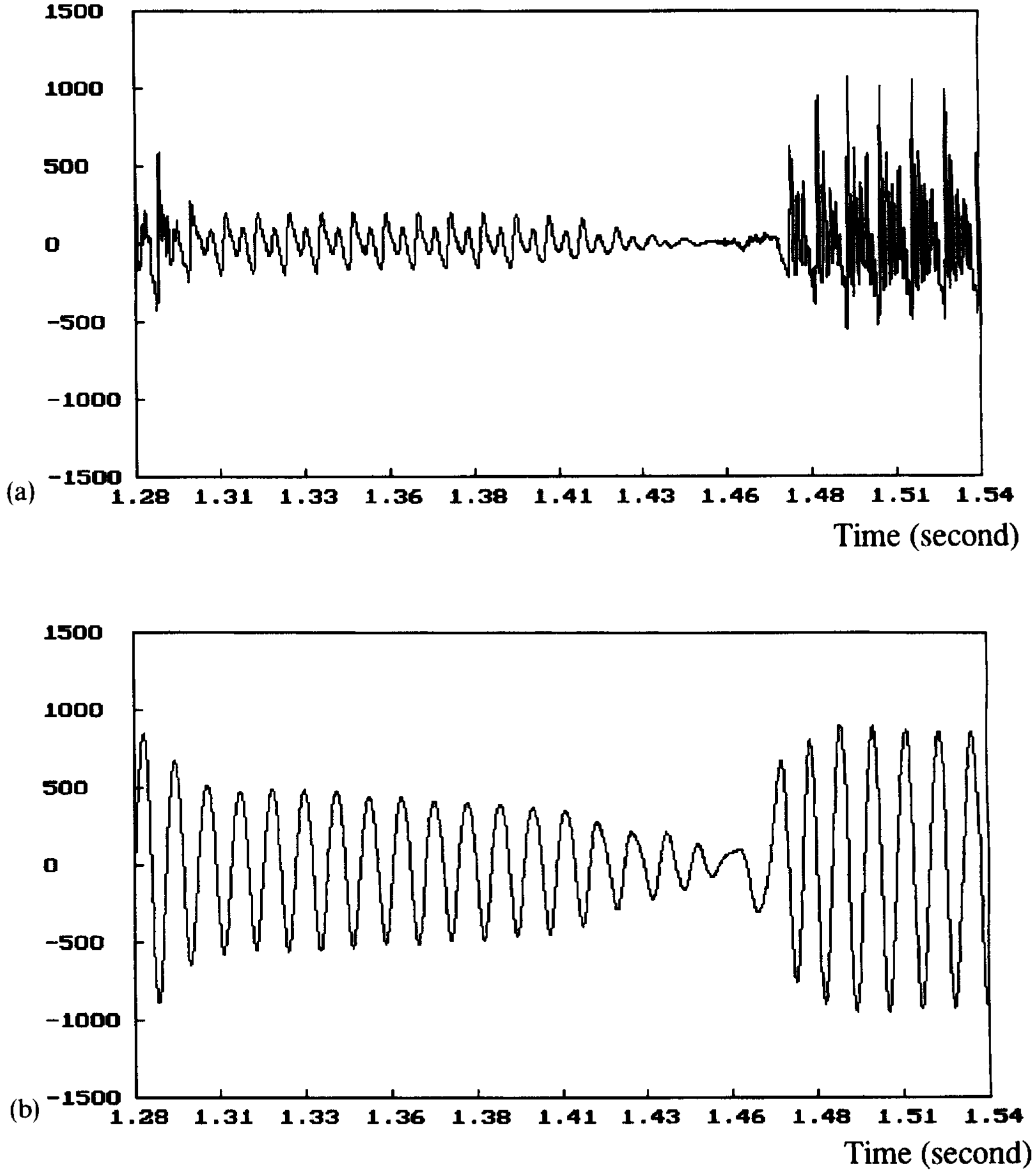

Fig. 3. (a) A frame of female speech ($f_s = 8000$ Hz); (b) filtered speech signal of Fig. 3 at scale $j = 3$.

When one deals with an actual human speech, it is necessary to filter the speech in parallel at the scales of $j = 3$ and $j = 4$. Based on the two filtered speech signals, two DIFs will be estimated separately. The correct fundamental frequency can then be determined from the two DIFs.

As an example, Fig. 3(a) and (b) shows the original and filtered signal of a segment of female speech. It is clear that the harmonics of the speech signal are eliminated or largely attenuated after filtering.

3.2. *Estimation of the DIFs*

The smoothed CFD DIF estimator is used in our investigation. When the CFD DIF estimator

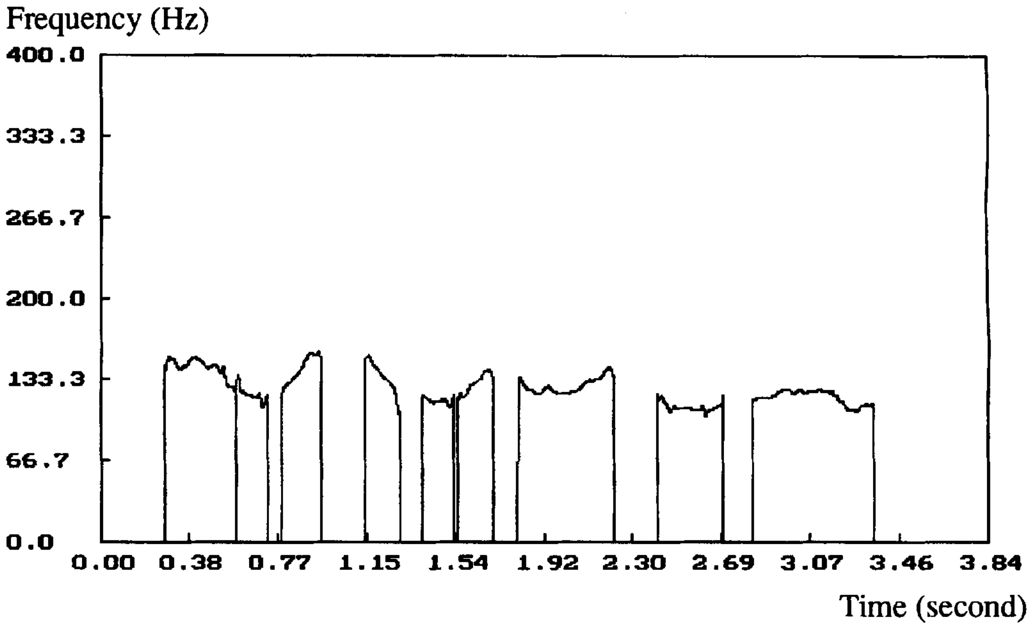

Fig. 4. Fundamental frequency estimation of timit.wav.

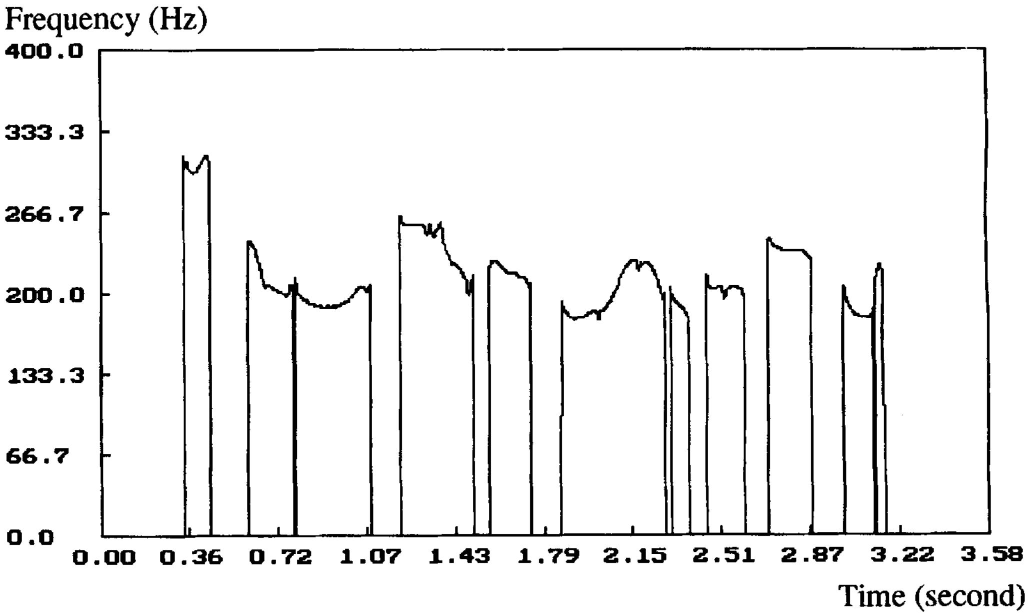

Fig. 5. Fundamental frequency estimation of clean.wav.

is applied to estimate the DIF of the wavelet transformed speech signal, there are distortions in the IF of the signal due to a small number of attenuated harmonics in the transformed speech signal. The experimental results show that the smoothed CFD DIF estimator can greatly reduce the variance of the CFD DIF estimator.

3.3. *Voiced or unvoiced and silence decisions*

After obtaining the DIF of the filtered speech signals, it can be found that there are very strong irregularities and large variations in the estimated DIF for the unvoiced or silent speech signal segments, whereas there is little variation in the voiced speech segments. We can therefore use a set of criteria to identify the voiced or unvoiced and silence segments as follows:

(1) The variations between two neighboring frequency samples are greater than 1.4 Hz.
(2) The frequencies are higher then 500 Hz.
(3) The frequencies are less than 50 Hz.
(4) The duration of a sustained frequency is less than 20 ms.

We set all the DIFs which satisfy the above criteria to zero. As a result, only the voiced fundamental frequencies remain.

3.4. *Decision of the fundamental frequency from the two estimated DIFs*

From the two estimated DIFs, we have to decide which one is the actual fundamental frequency. For the fundamental frequency range of 50–110 Hz, the estimated DIFs using scale $j = 4$ are the correct fundamental frequencies, whereas the estimated DIFs using scale $j = 3$ are the doubles of the actual fundamental frequencies. This is because the wavelet transform at scale $j = 3$ preserves the first harmonic rather than the fundamental frequency, whereas the wavelet transform at $j = 4$ preserves the fundamental frequency in this frequency range. In the range of 110–250 Hz, the estimated DIFs at both scales result in the same correct fundamental frequencies. For the range over 250 Hz the estimated DIF using scale $j = 3$ has the correct result and the estimated DIF using scale $j = 4$ has large variations and is classified as unvoiced speech or silence. This is because the fundamental frequency is outside the range of the wavelet transform at $j = 4$ so that no frequency can be estimated in this range using scale $j = 4$. The frequency halving error will never occur in this method because there is no spectral peak below the fundamental frequency. Based on these observations we can determine the fundamental frequency according to the following rule.

(1) Retain the DIF which is nonzero.
(2) Choose the smaller DIF in the case of frequency doubling.

4. Experimental results and discussions

As examples, the Sheffield signals, timit.wav and clean.wav, which were analyzed by participants in the European Speech Communication Association (ESCA) tutorial, are used. They include a male utterance, timit.wav, "she had your dark suit in greasy wash water all years" and a female utterance, clean.wav, "Fred can go, Susan can't go, and Linda is uncertain". The fundamental frequency estimations of timit.wav and clean.wav based on the smoothed CFD DIF estimator with window length 21 samples are illustrated in Figs. 4 and 5, respectively.

We also compared the experimental results obtained by using the DIF-based fundamental frequency estimator (smoothed CFD DIF with window length 21 samples) with those obtained using conventional pitch period estimation method. The conventional pitch period determination method is based on the auto-correlation function (ACF) estimation with forward and backward tracking [2]. Also the submultiple of the estimated pitch period is considered to reduce the occurrences of wrong estimation of pitch period.

Sixteen male and female utterances of more than 100 s in duration have been tested in the experiment. We found that there are wrong estimates for the conventional pitch period determination method in two utterances which are plotted in the same figure in Figs. 6 and 7. The continuous line is the fundamental frequency estimated using the DIF-based method and the dots are the fundamental frequencies estimated using the conventional method.

In two time periods, the fundamental frequencies estimated by the conventional pitch period estimation method are half of the actual fundamental frequencies. Although the conventional method has taken into account the submultiple problem of the estimated pitch period, the problem cannot be

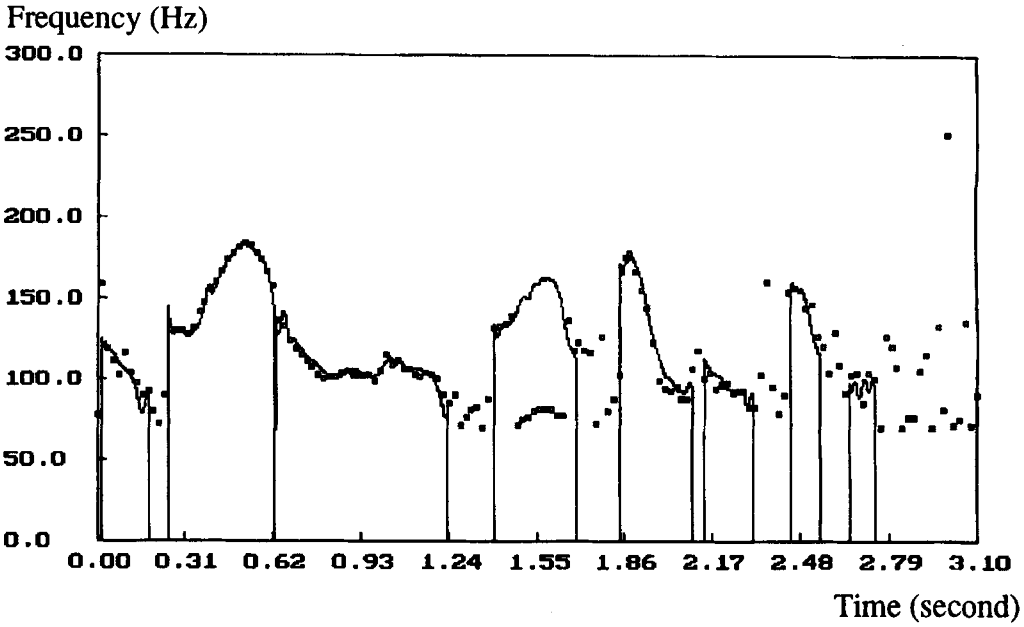

Fig. 6. Fundamental frequency estimations of a male speech "the swan dive was far short of perfect".

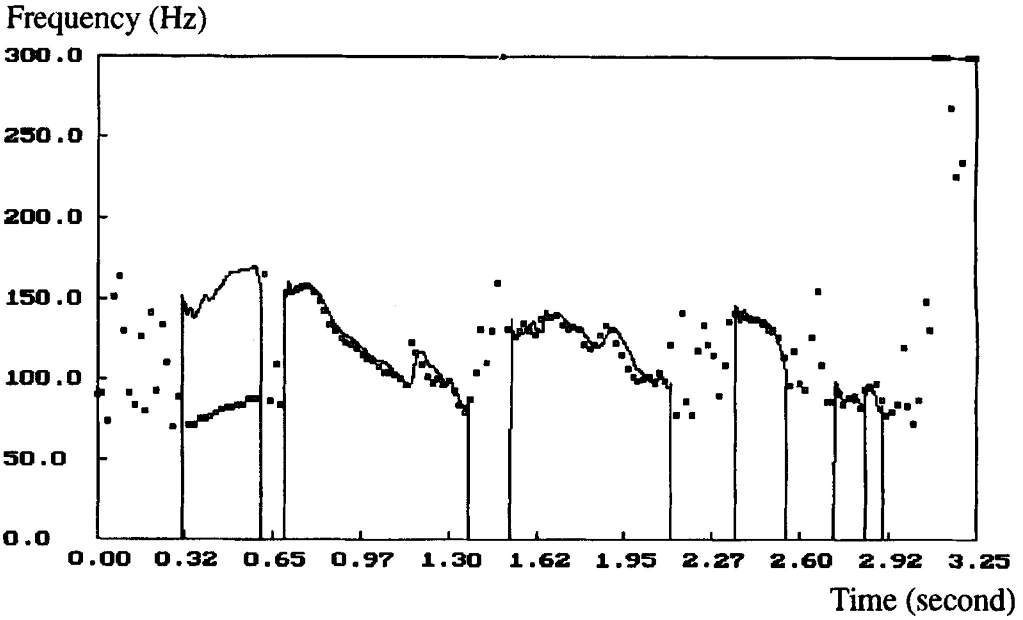

Fig. 7. Fundamental frequency estimations of a male speech "four hours of studying work faced us".

overcome completely. The DIF-based fundamental frequency estimator does not have the problem of the multiples faced by the conventional method in this case.

5. Conclusion

A new fundamental frequency determination method for speech signals is presented in this paper. The DIF-based fundamental frequency determination method involves harmonic elimination and attentation, and DIF estimation. The DIF corresponds to the fundamental frequency of the filtered speech signal. The new method can detect the true nonstationarity of the speech signals. The new method does not seem to have the problem of double pitch period estimation found in the conventional pitch determination algorithm.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments.

References

[1] T.A.C.M. Claasen and W.F.G. Mecklenbrauker, "The Wigner distribution – A tool for time-frequency signal analysis Part 2: Discrete-time signals", *Philips J. Res.*, Vol. 3, 1980, pp. 276–300.

[2] D.W. Griffin and J.S. Lim, "Multiband excitation vocoder", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, No. 8, August 1988, pp. 1223–1235.

[3] W. Hess, *Pitch Determination of Speech Signals Algorithm and Devices*, Springer, Berlin, 1983.

[4] B.C. Lovell and R.C. Williamson, "The statistical performance of some instantaneous frequency estimator", *IEEE Trans. Signal Process.*, Vol. 40, No. 7, July 1992, pp. 1708–1723.

[5] Y. Pati and P. Krishnaprasad, "Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations", *IEEE Trans. Neural Networks*, Vol. 4, No. 1, January 1993, pp. 73–85.

[6] L. 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.*, Vol. 24, 1976, pp. 399–418.

[7] J. Ville, "Theorie et applications de la notion de signal analytique", *Cables Transmission*, Vol. 2A, 1948, pp. 61–74.