ESTIMATION OF CONTINUOUS FUNDAMENTAL FREQUENCY OF SPEECH SIGNALS

Lunji Qiu, Haiyun Yang and Soo Ngee Koh

School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore 2263

ABSTRACT - Most of the well known and widely used speech analysis algorithms are frame-based with the assumption that a speech signal is locally stationary over the analysis frame. A different approach to estimate the continuous fundamental frequency of speech signals is considered in this paper. The approach can detect the true non-stationarity of the speech signals as it provides a continuous (sample by sample) fundamental frequency estimation as a function of time. Our algorithm is based on the instantaneous frequency (IF) estimation technique. A bank of two filters are used to remove and attenuate the harmonics of the speech signals. The fundamental frequency of a speech signal is then estimated by the IF technique.

1 INTRODUCTION

The determination of the pitch period of a speech signal is an important issue in speech processing. Accurate estimation of the time-varying pitch period, or fundamental frequency, of speech signals still constitutes one of the most integral and at the same time most problematic topics. Numerous studies have been dedicated to the design and evaluation of the pitch period or fundamental frequency determination algorithms. Time-domain methods (Hess, 1983) (Rabiner and Cheng, 1976) are commonly based on the use of autocorrelation functions to extract frequency tracks in each frame. Frequency-domain methods (Hess, 1983) (Rabiner and Cheng, 1976) use frequency spectra of each frame for examining the harmonic relations between the spectral components to infer the corresponding fundamental frequency. The conventional pitch period or fundamental frequency determination algorithms in either the time domain or frequency domain are frame-based with the assumption that a speech signal is locally stationary within a frame. This kind of approach occasionally experiences difficulties due to (1) the chosen frame length, (2) the position of the frame with respect to the glottal peak in a voiced speech signal, and (3) the assumption of stationarity. As a consequence, the conventional frame-based algorithms have difficulty in representing the truly non-stationarity of speech signals and sometimes fail to correctly estimate the pitch period since speech signals are naturally nonstationary in characteristics.

In this paper, we present a method which provides a continuous fundamental frequency estimation based on the IF technique. It seems to be able to overcome the difficulties faced by most of the conventional methods. Our method can detect the true nonstationarity of the speech signals as it provides a continuous fundamental frequency as a function of time. Also, the new method does not seem to have the problem of occasional incorrect frequency estimation, such as frequency doubling or halving, usually associated with the conventional methods.

The rest of the paper is organised as follows. In Section 2, a speech model is introduced and our algorithm is presented. The details of the algorithm are discussed in Section 3 and the experimental results are illustrated in Section 4. Section 5 concludes with a discussion of the results obtained and avenues for future research.

2 CONTINUOUS FUNDAMENTAL FREQUENCY ESTIMATION

A speech signal can be considered as a signal which consists of N sinusoids as described bellow :

$$x(t) = \sum_{i=1}^{N} A_i(t) \cos[2\pi i f_0(t) t + \phi_i(t)] \tag{1}$$

The sinusoidal parameters $\{A_1, A_2, ..., A_N, f_0(t)\}$, which consist of amplitudes and the time-varying fundamental frequency, are unknown and are to be estimated. The task of estimating the fundamental frequency is to estimate the parameter $f_0(t)$. The fundamental frequency, $f_0(t)$, is a function of time, t, since speech signals are nonstationary in characteristics. To estimate the continuous time-varying fundamental frequency, $f_0(t)$, the following steps are needed in our algorithm.

1) Pre-processing: To attenuate the harmonics of the speech signal by use of wavelet-transforms (WT) so that the maximum peak of the pre-processed signal takes place at the fundamental frequency.

2) IF estimation: The smoothed central finite difference (CFD) discrete IF estimator is used to estimate the frequencies of the pre-proceed speech signals.

3) Post-processing: The voiced/unvoiced (or silent) decision is made and the fundamental frequency is determined.

3 ALGORITHM FOR CONTINUOUS FUNDAMENTAL FREQUENCY ESTIMATION

3.1 Pre-Processing

The speech signals as defined in equation (1) are naturally multicomponent signals and there are $N-1$ harmonics in the spectrum. Unfortunately, the IF estimation technique works only for mono-component signals. Therefore, a speech signal has to be transformed into a mono-component or near mono-component signal before its IF can be estimated. In other words, the harmonics of the speech signals have to be removed or largely attenuated. The attenuation and elimination of the harmonics are carried out in the pre-processing stage.

A bank of two filters, $\psi_1(t)$ and $\psi_2(t)$, implemented by the use of the dyadic WT, is used to achieve this objective (Cooke and Beet, 1989). The filters are constructed from the sigmoidal function, $s(t) = (1+e^{-2t})^{-1}$, as

$$\psi(\frac{t}{2^j}) = s(\frac{t}{2^j}+2) - s(\frac{t}{2^j}-2) \tag{2}$$

and the Fourier transform of equation (2) is given by

$$|\Psi(\omega)|^2 = \frac{4\pi^2 2^j \sin^4(2^j\omega)}{\sinh^2(2^{j-1}\pi\omega)} \tag{3}$$

where $j=3$ and $j=4$ are for filter functions $\psi_1(t)$ and $\psi_2(t)$ respectively.

The dyadic WT is therefore defined (Mallat, 1989) as

$$D_y WT_x(\tau, 2^j) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} x(t)\psi^*(\frac{t-\tau}{2^j})dt \tag{4}$$

where $\psi^*(t)$ is the complex conjugate of the basis wavelet function $\psi(t)$ which must satisfy the condition that $\int_{-\infty}^{\infty} \psi(t)dt = 0$. The details of this wavelet function can be found in (Mallat, 1989) (Pati and Krishnaprasad, 1993).

Fig.1(a) shows a segment of a male speech and Fig.1(b) shows the dyadic wavelet transformed of the speech segment in Fig.1(a). It is clear that the harmonics of the speech signal are attenuated after the WT. It was found through simulations that for a speech signal of which the fundamental frequency is in the range of $100-450Hz$, the wavelet function at scale $j=3$ in Fig.1 is more suitable. The fundamental frequency for most of the female speech signals falls within this band. It is therefore suggested that for female speech, the WT should be taken at scale $j=3$. However, if the fundamental frequency of the speech signal is very low, say less than $100Hz$ or around $100Hz$, the first harmonic of such speech signal is less than or around $200Hz$. In this case, the first harmonic cannot be eliminated nor attenuated by the use of the wavelet function at scale, $j=3$. This will result in estimation errors. To overcome this problem, the wavelet function at scale $j=4$ should be used to attenuate or even eliminate the first harmonic of the speech signal. From Fig.1, we can see that the wavelet function at scale $j=4$ has the pass-band from $10Hz$ to $250Hz$. The fundamental frequency for most of the male speech signals falls within this band. However, when one deals with an actual

human speech, there is a question of which scale ($j = 3$ or $j = 4$) to be used since the apriori knowledge of the fundamental frequency range is not available. To solve this problem, it is necessary to take the WTs in parallel at both scales of $j = 3$ and $j = 4$. Based on the two transformed speech signals, two fundamental frequencies can be estimated separately. The correct fundamental frequency can then be determined from the two estimates.
After the WT, the speech signal defined by equation (1) becomes

$$\begin{aligned} x'(t) &= A_1'(t)\cos[2\pi f_0(t)t + \phi_1(t)] + \sum_{i=2}^{M} A_i'(t)\cos[2\pi i f_0(t)t + \phi_i'(t)] \\ &\approx A_1'(t)\cos[2\pi f_0(t)t + \phi_i'(t)] \end{aligned} \tag{5}$$

where $M \ll N$ and $A_i' \ll A_i$.

3.2 Estimation of Instantaneous Frequency

For monocomponent signals, the IF describes the frequency modulation law of the signal and its energy is concentrated about this law. The IF of a signal can be uniquely defined by using its corresponding analytic signal. For a given real signal $s(t)$, several complex representations may be obtained. A unique representation may be obtained if the Hilbert transform is used to generate the complex signal (Ville, 1948). The IF of a signal s(t) can then be uniquely defined by using an analytic signal. This is formally expressed as

$$f_i(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{6}$$

where $f_i(t)$ is the IF of the signal, $\phi(t)$ is the phase of the analytic signal, $z(t) = a(t)e^{j\phi(t)}$, which is formed from the real signal $s(t)$.
The definition of IF given by Ville (1948) may be extended to the discrete time signal case, with the discrete version of the IF being referred to as DIF. The corresponding estimator, based on the central finite difference (CFD) of the phase, was defined by Claasen and Mechlenbrauker (Classen and Mecklenbrauker, 1980) as

$$f_i(n) = \frac{f_s}{4\pi}\{arg[z(n+1)] - arg[z(n-1)]\}_{mod.2\pi} \tag{7}$$

where f_s is the sampling frequency. The notation $mod.2\pi$ represents a modulo 2π operation and $arg[z(n)]$ is the phase of $z(n)$ which is defined as

$$arg[z(n)] = tan^{-1}\frac{Im[z(n)]}{Re[z(n)]} \tag{8}$$

where $Im[z(n)]$ and $Re[z(n)]$ are respectively the imaginary and real parts of $z(n)$, and $arctan$ is the principal value of inverse tangent (i. e.$-\pi < arctan(x) < \pi$).
Note that if $z(t) = a(t)e^{j\phi(t)}$, $arg[z(t)] = tan^{-1}\frac{a(t)sin\phi(t)}{a(t)cos\phi(t)} = \phi(t)$, then $f_i(n) = \frac{f_s}{4\pi}[\phi(n+1) - \phi(n-1)]_{mod.2\pi}$. This is just a central finite difference approximation of $\frac{d\phi(t)}{dt}$, and hence the name CFD.
A smoothing window or a time averaging operation may be applied to the CFD DIF estimation to reduce the dispersion at the expense of time resolution. Linear convolution with a smoothing function may be used to obtain the smoothed CFD DIF estimator (Lovell and Williamson, 1992). However, since the CFD DIF estimation is a periodic phase signal (it is modulo $\frac{f_s}{2}$, where f_s is the sampling frequency), linear convolution cannot be simply used in the operation, and a modulo convolution with a smoothing window must be used instead (Lovell and Williamson, 1992).
Let $f_i(n)$, which is modulo $\frac{f_s}{2}$, 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:

$$\begin{aligned} f_i^s(n) &= [f_i(n)(*)h(n)]_{mod.\frac{f_s}{2}} \\ &= \frac{f_s}{4\pi}arg[\sum_{p=-\frac{P-1}{2}}^{\frac{P-1}{2}} h(p)e^{j4\pi\frac{f(n-p)}{f_s}}]_{mod.\frac{f_s}{2}} \end{aligned} \tag{9}$$

3.3 Post-Processing

After obtaining the frequencies of the transformed speech signals, we can see that there are very strong irregularities and large variations in the frequencies estimated for the unvoiced or silent speech signal segments, whereas there is little variation in the voiced speech segments. We can therefore set up 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 $1.4Hz$.

2) The frequencies are higher than $500Hz$.

3) The frequencies are less than $50Hz$.

4) The duration of a sustained frequency is less than $20ms$. (Voiced speech signal does not occur for only that short duration of time.)

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

From the two estimated frequencies, we have to decide which one is the actual fundamental frequency. For the fundamental frequency range of $50Hz - 110Hz$, the estimated frequency using scale $j = 4$ are the correct fundamental frequencies whereas the estimated frequency using scale $j = 3$ are the doubles of the actual fundamental frequency. 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 $110Hz - 250Hz$, the estimated frequencies at both scales result in the same correct fundamental frequency. For the range over $250Hz$ the estimated frequency using scale $j = 3$ has the correct result and the estimated frequency 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 algorithm.

1) Retaining the frequencies which is non zero.

2) Taking the smaller frequency in the case of frequency doubling.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the simulation results are presented to illustrate the fundamental frequency detector described in the previous sections. The Sheffield signals (Cooke and Beet, 1993), 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", spoken by a male speaker from the southern USA dialect region and a female utterance, clean.wav, "Fred can go, Susan can't go, and Linda is uncertain", uttered by an English female speaker with a Southern English accent. The fundamental frequency estimations of timit.wav and clean.wav at scale $j = 3$ and $j = 4$ based on the smoothed CFD DIF estimator with window length 21 samples are illustrated Fig.2(a), 2(b) and Fig.3(a), 3(b), respectively.

The fundamental frequency estimations of timit.wav and clean.wav after the voiced or unvoiced and silence decision are plotted in Figs.2(c), 2(d) and Figs.3(c), 3(d) respectively. Consequently, for the timit.wav and clean.wav, the final estimated fundamental frequency plots are given in Fig.2(d) with $j = 4$ and Fig.3(c) with $j = 3$.

5 CONCLUSION

A continuous fundamental frequency determination method for speech signals is presented in this paper. The IF-based fundamental frequency determination method involves WT and IF estimation. The WT is used to eliminate and attenuate the harmonics of the speech signals. The IF, which can correspond to the fundamental frequency of speech signals, can then be estimated from the wavelet transformed speech signals.

The new method can detect the true non-stationarity of the speech signals and leads to accurate estimation of the pitch period. The new method has the computational advantage of low complexity.
For future research, it would be very interesting to make the IF-based fundamental frequency determination method adaptive such that it can make use of the previous fundamental frequency values to adjust the order of the wavelet transform.

REFERENCES

Claasen T. A. C. M. and Mecklenbrauker W. F. G. (1980), The Wigner Distribution-A Tool for Time-Frequency Signal Analysis Part 2: Discrete-Time Signals, *Philips Journal of Research*, Vol.3, 276–300.

Cooke S. M., Beet S. and Crawford M. (eds.) (1993), *Visual Representations of Speech Signals*, John Wiley & Sons Ltd.

Hess W., (1983) *Pitch Determination of Speech Signals Algorithm and Devices*, Springer-Verlag.

Lovell B. C. and Williamson R. C. (1992), The Statistical Performance of Some Instantaneous Frequency Estimator, *IEEE Trans. on SP*, Vol.40, No.7, July, 1708–1723.

Mallat S., (1989) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *IEEE Trans. on Patt. Anal. Machine. Intell.*, Vol.11, No.7, 674–693.

Pati Y. and Krishnaprasad P., (1993) Analysis and Synthesis of Feedforward Neural Networks Using Discrete Affine Wavelet Transformations" *IEEE Trans. on Neural Networks*, Vol.4, No.1, Jan. 73–85.

Rabiner L., Cheng M. J., Rosenberg A.E. and McGonegal C.A. (1976) A Comparative Performance Study of Several Pitch Detection Algorithms, *IEEE Trans. on ASSP*, Vol.24, 399–418.

Ville J., (1948) Theorie et Applications de la Notion de Signal Analytique, *Cables et Transmission*, Vol.2A, 61–74.

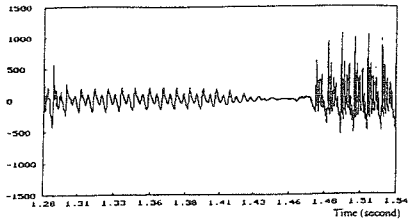

Fig.1(a) A Frame of Female Speech ($f_s = 8000Hz$).

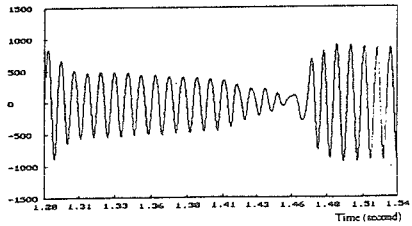

Fig.1(b) The Wavelet Transformed Speech Signal of Fig.1(a) at Scale $j = 3$.

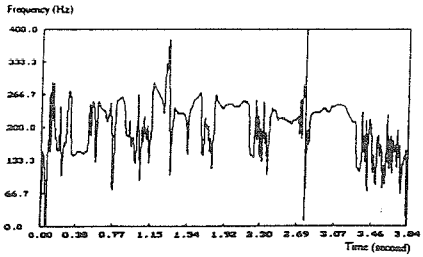

Fig.2(a) DIF of Transformed timit.wav at $j = 3$.

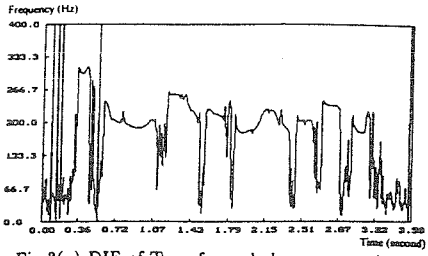

Fig.3(a) DIF of Transformed clean.wav at $j = 3$.

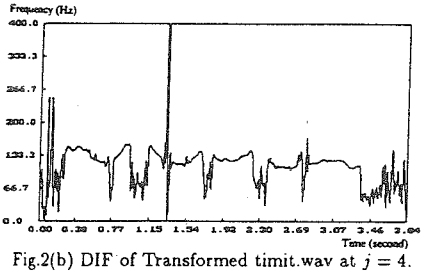

Fig.2(b) DIF of Transformed timit.wav at $j = 4$.

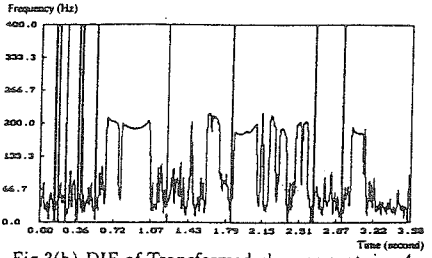

Fig.3(b) DIF of Transformed clean.wav at $j = 4$.

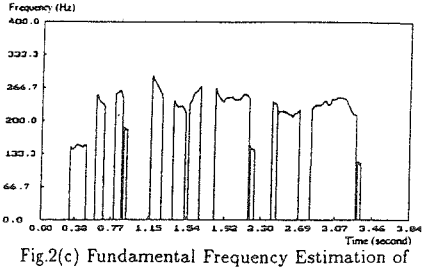

Fig.2(c) Fundamental Frequency Estimation of timit.wav at $j = 3$.

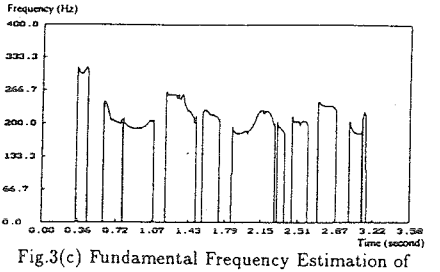

Fig.3(c) Fundamental Frequency Estimation of clean.wav at $j = 3$.

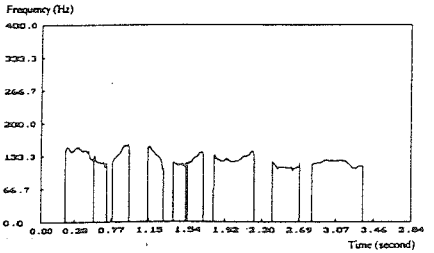

Fig.2(d) Fundamental Frequency Estimation of timit.wav at $j = 4$.

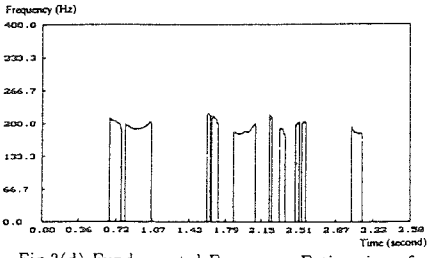

Fig.3(d) Fundamental Frequency Estimation of clean.wav at $j = 4$.