NEW AND IMPROVED PITCH DETERMINATION FOR THE IMBE VOCODER

T. S. Lim and M. S. Scordilis

Department of Electrical and Electronics Engineering
The University of Melbourne

ABSTRACT - A robust and accurate pitch determination algorithm of infinite resolution is presented in this paper. This method makes use of a hybrid of time domain and frequency domain pitch estimation techniques. For frame to frame analysis, this method was found to provide high accuracy in extracting the pitch period best representing the average pitch within a speech frame. It is a computationally efficient technique, particularly when used as part of the IMBE vocoder.

INTRODUCTION

In many speech synthesis and vocoder applications, one of the most important parameters used are the presence of voicing and the pitch period of the voiced parts of the signal. The human ear is very sensitive to pitch degradation and therefore accurate pitch evaluation is of paramount importance. Moreover, pitch determination is considered to be one of the most difficult tasks in speech processing. Since Dudley's attempt in 1939 to extract pitch by lowpass filtering (Schroeder, 1966), many techniques of automatic pitch extraction have been developed both for the time and for frequency domain (Hess, 1983).

Despite the large number of pitch estimation algorithms, the problem of a robust, reliable, and accurate pitch extraction still remains open. The complexity of pitch determination is due to the variability and irregularity in the nature of speech. In order to overcome the non-stationarity of the speech signal, short time analysis is often used. However, the wide range of values for the pitch period, as well as the changes in the state of voicing of the signal within the analysis frame (i.e., a mixture of voiced and unvoiced segments), lead to a crude average value or even wrong pitch estimation. In addition, a pitch estimate expressed as an integer multiple of the sampling interval, contains time quantization errors which may lead to audible distortion in speech coding application (Hess & Indefrey, 1984).

A pitch detection algorithm using a new similarity model was introduced by Medan, et al (1991) to overcome most of these problems in pitch processing. A more efficient, high resolution pitch estimation was also proposed, based on an IIR second order interpolator (Medan, 1991). After detailed simulation and observations it was found that these techniques do provide a robust and reliable pitch estimation. However, for high resolution (i.e., non-integer) pitch, required in speech coding applications, such as in the Multi-Band Excitation (MBE) Vocoder (Griffin & Lim, 1988), the accuracy of these methods is insufficient to overcome the problems introduced by the severe non-stationarity of the speech signals within an analysis frame.

In an effort to resolve this problem a hybrid method is proposed here and it combines time domain and frequency domain pitch refinement strategies. This hybrid method was found to be robust and reliable technique. It was used to accurately and efficiently estimate pitch in frame by frame analysis/synthesis of speech for the MBE vocoder.

PITCH DETERMINATION USING A SIMILARITY MODEL

Integer Pitch Determination

For short speech segments, voiced speech is of quasi-periodic nature, and although adjacent periods are rarely identical, they tend to be very similar. It has been shown (Medan, et al, 1991) that the normalized cross-correlation of adjacent speech segments provides an optimum criterion for determining their degree of similarity. The normalized cross-correlation over time interval τ is defined as:

$$\rho_\tau(x, y) = \frac{(x, y)_\tau}{|x|_\tau |y|_\tau},$$

where $x_\tau = x(t)\, w_\tau(t)$ and $y_\tau = y(t)\, w_\tau(t)$, with $w_\tau(t)$ denoting a rectangular window of length τ.

An initial estimate of the pitch period T_0 is the value of τ that maximizes the normalized cross-correlation function. In other words,

$$T_0 = \underset{\tau}{\text{argmax}}\ \rho_\tau(x, y).$$

The search interval τ, is limited to the allowable T_0 range.

Speech was lowpass filtered and sampled at 8000 KHz. The adjacent windowed discrete signal segments were $x_N = x(n)\, w_N(n)$ and $y_N = y(n)\, w_N(n)$, with $w_N(n)$, being a rectangular window of length N. The discrete estimate of the pitch is N_0 and it is defined as

$$N_0 = \underset{N}{\text{argmax}}\ \rho_N(x, y)_N\,,$$

where $(x, y)_N$ is the inner product of two finite sequences x_N and y_N and it is given by

$$(x, y)_N = \sum_{n=1}^{N} x(n)\, w(n)\,.$$

The value of N was in the range (N_{min}, N_{max}). For these simulations the values for N_{min} and N_{max} were 20 and 200 respectively, which correspond to 2.5 ms and 25 ms intervals.

Pitch refinement in the frequency domain

The integer pitch N_0 estimated with finite resolution contains sampling rate-dependent "rounding" errors. Also, time-domain integer pitch estimators account for the non-stationary properties of the speech signal within a chosen frame quite poorly. These problems were overcome by using a frequency domain pitch refinement method with the integer N_0 being the estimate of the initial value for pitch before refining.

The MBE model provided the basis for this work. In the MBE vocoder the signal parameters corresponding to a windowed speech segment are estimated and used for synthesis. The FFT is applied to every Hamming-widowed speech segment . In this work, the analysis frames consisted of 512 points (64 ms) and the shift was 128 points (16 ms) resulting in 75% overlap. Figure 1 shows the spectrum of a voiced speech segment used as an illustration.

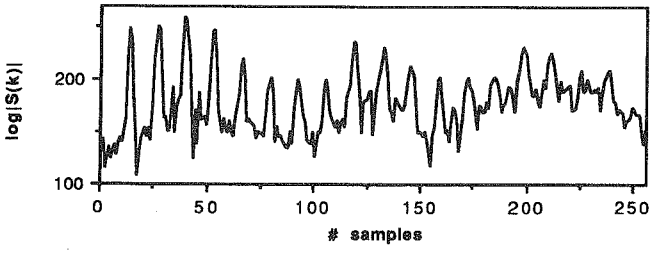

Figure 1. FFT spectrum of the original signal.

If the terminal analog model is used to describe the speech production process, then speech can be modeled as the convolution of the impulse response of a linear system with an appropriate excitation signal which is a mixture of a periodic impulse sequence and gaussian noise. Successful analysis/resynthesis depends on the accuracy of the model in justifying the properties of the produced signal.

For all-voiced speech, let $|S(k)|$ denote the NW-point FFT speech spectrum of the windowed original signal. The corresponding synthetic speech spectrum $|\hat{S}(k, \omega_0)|$ is given as the product of the spectral envelope samples, $|A(k)|$, with the periodic spectrum of the excitation, $|\hat{P}(k, \omega_0)|$, i.e.,

$$|\hat{S}(k, \omega_0)| = |A(k)|\,|\hat{P}(k, \omega_0)|; \qquad 1 \le k \le NW.$$

$|\hat{P}(k, \omega_0)|$ is generated by either computing the FFT of a windowed impulse sequence, or it is synthesized as the Fourier transform of a windowed periodic signal. The speech spectrum can be viewed as the amplitude modulation of the spectral envelope with $|\hat{P}(k, \omega_0)|$. The error between the actual and synthetic spectra is expressed as:

$$e(\omega) = |S(k)| - |\hat{S}(k, \omega_0)| = |S(k)| - |A(k)|\,|\hat{P}(k, \omega_0)|; \qquad 1 \le k \le NW.$$

The error, $e(\omega)$ is minimized when the synthetic spectrum matches the original both in the envelope and in the harmonic (or random) composition. The search for matching is restricted in the range $(\omega_0 - \Delta\omega_0, \omega_0 + \Delta\omega_0)$. The fundamental radian frequency value, ω_0, is given by $\omega_0 = \frac{2\pi}{N_0 T}$, where N_0 is the initial pitch estimate and T is the sampling period. After an initial pitch estimation has been produced, a refined and accurate ω_0 is given by

$$\omega_0 = \underset{\omega}{\text{argmax}}\ \rho_\omega(S, P).$$

The inner product $(S, P) = (S,P)_\omega$, and over one analysis frame it is given as

$$(S, P)_\omega = \sum_{k=1}^{NW} |S(k)|\,|P(k, \omega_0)|.$$

The refined pitch is obtained by evaluating $\rho_\omega(S,P)$ over the fundamental radian frequency range corresponding to $(N_0 - \Delta N_0, N_0 + \Delta N_0)$. The pitch deviation, ΔN_0 was introduced to account for pitch variations due to the non-stationarity of the signal and it was found to provide the best results for a value of 5 samples.

Computation of High Resolution Pitch Refinement

Pitch refinement uses the FFT spectrum computed for windowed speech in the analysis section of the MBE vocoder. The integer pitch estimation is provided by the cross-correlation of adjacent signal segments. The major computational load of this hybrid method comes from the computation of the periodic spectrum of the excitation $|\hat{P}(k, \omega_0)|$, which must be calculated or synthesized for any value of ω_0 corresponding to the nominated range of the hypothesized fundamental frequency. The number of computations for estimating the pitch period depends on the initial pitch value, the fundamental frequency deviation and the desired resolution.

In an effort to reduce the large number of computations required for deriving $\hat{P}(k, \omega_0)|$, an alternative but effective technique was sought. A simplified harmonic spectrum consisting of shifted main lobes of the sinc function was chosen as a practical alternative. The side lobes were removed as being irrelevant to the peak matching procedure. Given a hypothesized ω_0, the harmonic test-spectrum is given by:

$$|P(k, \omega_0)| = \begin{cases} \frac{2A}{\pi} \sin\left(\frac{\pi(k - mF_0)}{2}\right) & \text{if } k \le |mF_0 \pm 2| \\ 0 & \text{otherwise} \end{cases}$$

where A is a constant, $F_0 = (\omega_0 NWT)/2\pi$, k = 0, 1, ..., NW, and m = 0,1, ..., N_0.

Figure 2 shows the synthetic harmonic spectrum for a given pitch value.

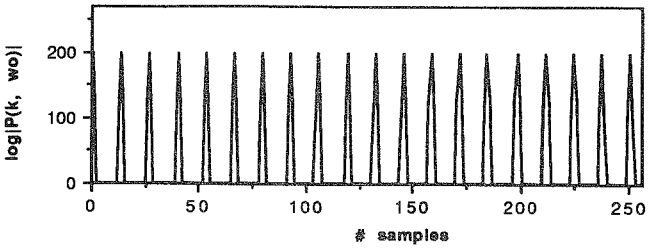

Figure 2. Synthetic periodic spectrum

When very high pitch resolution is desired the computation of the frequency domain pitch refinement can be further reduced by evaluating the target pitch in stages of increased accuracy. This can be achieved by reducing the number of possible pitch candidates within a given pitch range (N_0 - ΔN_0, N_0 + ΔN_0). For each stage the pitch deviation, ΔN_0, is decreased and the resolution is increased until the required resolution is achieved.

SIMULATIONS AND RESULTS

The performance of the proposed hybrid method was tested for several utterances produced by male speakers. The utterances were recorder at a sampling rate of 8 kHz using the Proport signal acquisition peripheral on a Sun Sparc2 workstation. The signal was first lowpass filtered at 2 kHz and used for the initial integer pitch determination. The short time Fourier transform of the original signal was calculated every 16 ms by computing 512-point FFT. The synthetic harmonic spectrum was also derived for pitch periods in the vicinity of the integer pitch estimates. Figures 3 and 4 show the original speech spectrum and the synthesized harmonic spectum for underestimated and overestimated pitch values respectively. It is evident that even errors within one sample can introduce large mismatches in the higher frequency regions. As shown on Figure 5, the exact pitch value produces a harmonic spectrum whose peaks exactly match the original spectrum. The exact pitch is the pitch value that maximizes the cross-correlation between the two sequences.

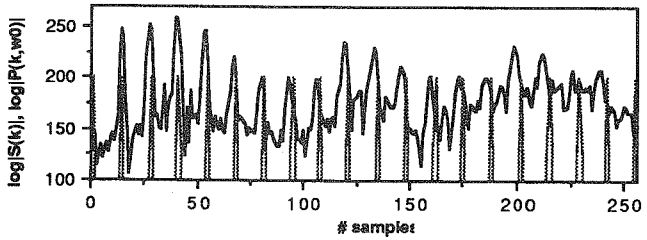

Figure 3. Natural spectrum and harmonic spectrum of underestimated pitch.

The performance of this method is illustrated in the example utterance "smile", shown on Figure 6. The integer estimate, and the high resolution pitch values for the proposed method are shown on Figure 7. The difference between the high resolution pitch algorithm developed by the authors and that by Medan et al (1991) is that this technique does not restrict the pitch deviation form the intial integer estimate but it allows for the final pitch value to vary for ranges more than 2 sample above or below the initial integer estimate. Visual comparisons verified that such a relaxation of the pitch variations was well founded and provides better pitchestimates. Figure 8 shows the resulting

differences between the initial, the refined and the pitch determined by the method proposed by Medan et al (1991. Note that the refined pitch deviation reached about 4.75 samples.

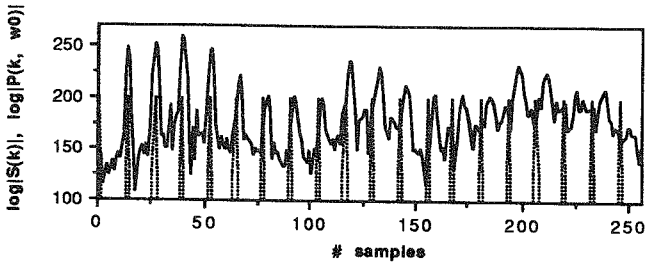

Figure 4. Natural spectrum and harmonic spectrum of overestimated pitch.

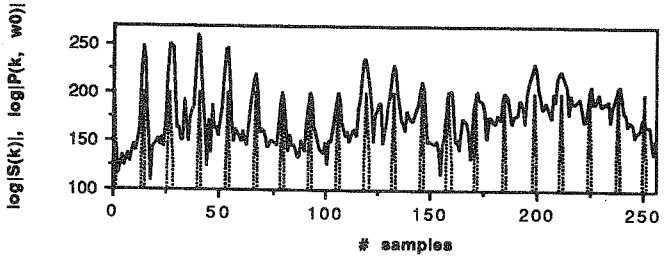

Figure 5. Natural spectrum and harmonic spectrum of exact pitch.

CONCLUSION

Algorithms for extracting integer estimates of the pitch period of speech are widely used in speech processing. For certain analysis/resynthesis applications, such as speech coding, pitch resolutions greater than the available sampling interval are necessary. A high resolution pitch detection scheme was developed in this paper. This method is a hybrid of time-domain and frequency-domain techniques and it has been shown to be a robust and accurate pitch determination. The target application of this new technique is speech vocoders that involve the computation of the Fourier transform of windowed speech, and in particular the multiband excitation vocoder (MBE). This relatively simple and reliable hybrid method makes it possible to extract the pitch period at infinitely high resolution. Real-time implementation of this algorithm as part of a speech coding scheme is well within the capabilities of current DSP technology.

REFERENCES

Dubnowski J.J, Schafer R.W. and Rabiner L.R., "Real-Time Digital Hardware Pitch Detector", "IEEE Trans. ASSP", pp. 2-8, Vol. ASSP-24, Feb. 1976.

Griffin D.W. and Lim J.S., "Multiband Excitation Vocoder", IEEE ICASSP, pp. 1223-1235, Vol. 36, 1988.

Hess, W., "Pitch Determination of Speech Signals", New York Springer, 1983.

Hess, W. and Indefrey, H., "Accurate Pitch Determination of Speech Signals by means of a Laryngograph", Proc. ICASSP-84", pp.1797-1800, Apr. 1984.

Medan, Y., Yair, E. and Dan Chazan, D., "Super Resolution Pitch Determination of Speech Signals", IEEE Trans. on Signal Processing, pp. 40-48, Vol. 39, Jan 1991.

Medan, Y., "Using Super Resolution Pitch in Waveform Speech Coders", IEEE ICASSP, pp. 633-636, 1991.

Schroeder, M.R., "Vocoders: Analysis and Synthesis of Speech", Proc. IEEE, pp. 720-734, Vol. 54, May, 1966.

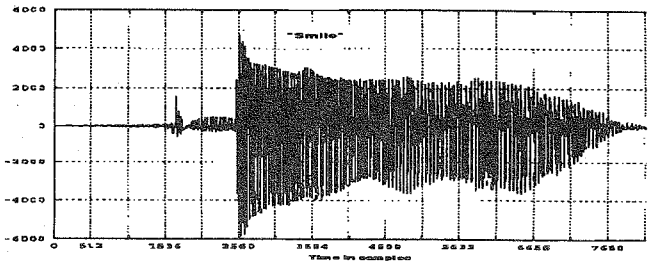

Figure 6. Test utterance "smile".

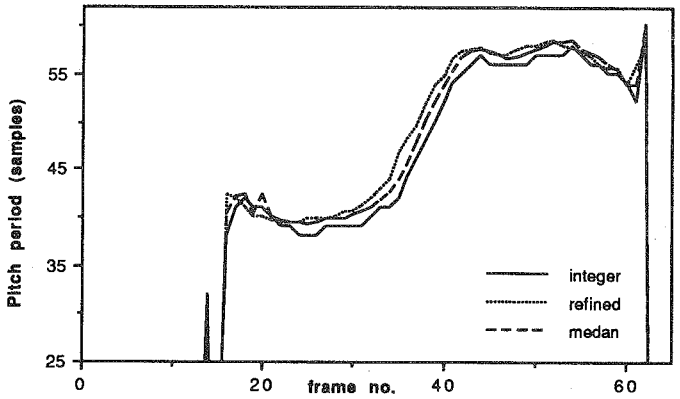

Figure 7. Integer estimate and the high resolution of pitch for test utterance "smile".

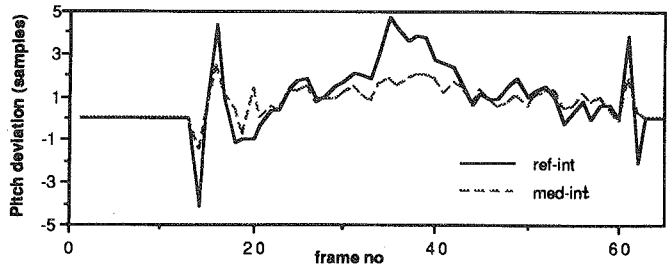

Figure 8. The resulting difference between initial and refined pitch for test utterance "smile".