An Investigation of Estimating Pitch Periods using a Non-linear Differential Operator

Robert K Whitman and Delores M. Etter

Department of Electrical and Computer Engineering
University of Colorado - Boulder

ABSTRACT

Numerous techniques have been developed to estimate the pitch periods in speech signals. These techniques use a variety of methods in both the time and frequency domains to estimate the pitch periods. In this paper, an investigation is made into a time domain algorithm which uses a non-linear differential operator developed by Teager [1,2], called an energy operator, to estimate the pitch period of the speech data. When the energy operator is applied to a voiced speech signal, it produced a signal with enhanced glottal pulses. From this signal, the pitch period is estimated and the resulting pitch track is smoothed. All operations in this algorithm are done in the time domain.

Introduction

The estimation of the pitch frequency in speech signals is an important process. Making an accurate estimate is a difficult task, since the pitch periods vary from one pitch pulse to the next. However, an accurate estimate is essential in performing various speech signal processing.

A variety of techniques have been developed to estimate the pitch in voiced speech data. These techniques involve operations in time, or frequency, or a combination of time and frequency domains. In most of these techniques, regularly-spaced data windows are used, with typical window lengths in the five to 40 ms range. These windows are usually overlapped by some number of samples and parameters are extracted at each window. The processing of these parameters then produces the pitch information.

In this investigation, a time domain technique is developed to extract the pitch frequencies from voiced speech signals. The speech data were modified by low-pass filtering and an energy operation to obtain a new signal, called the energy data or energy signal, which is used in the pitch estimate. The application of the energy operator on the speech signal resulted in a signal with enhanced glottal pulses. The pitch periods were extracted from this new signal. After the pitch frequencies are calculated, the pitch track is smoothed.

The Teager Energy Operator

The energy operator is a non-linear differential operator derived from analysis of the total energy in a second-order harmonic oscillation system [3]. The operator is used to calculate a running estimate of the energy required by a source to produce a signal. Since the energy is obtained by a computation involving three consecutive samples (1), this estimate is almost instantaneous. Previous papers [3,4] derived the operator for both the discrete (1) and continuous (2) signal cases.

$$\Psi_d[x(n)] = x^2(n) - x(n-1)x(n+1) \qquad (1)$$

$$\Psi_c[x(t)] = \left[\frac{dx(t)}{dt}\right]^2 - x(t)\left[\frac{d^2x(t)}{dt^2}\right] \qquad (2)$$

In this investigation, the discrete energy operator (1) is applied to the low-pass filtered speech data.

Pre-processing the Data

The input voiced speech data, s(t), (Fig. 1a) is digitized, s(n), at some sampling frequency, F_s. The data used in this investigation were sampled at 10.4kHz. If the energy operator is applied directly to the unfiltered speech data, one gets enhanced glottal

pulses and pulses resulting from upper harmonics (Fig. 1b), or from consonants.

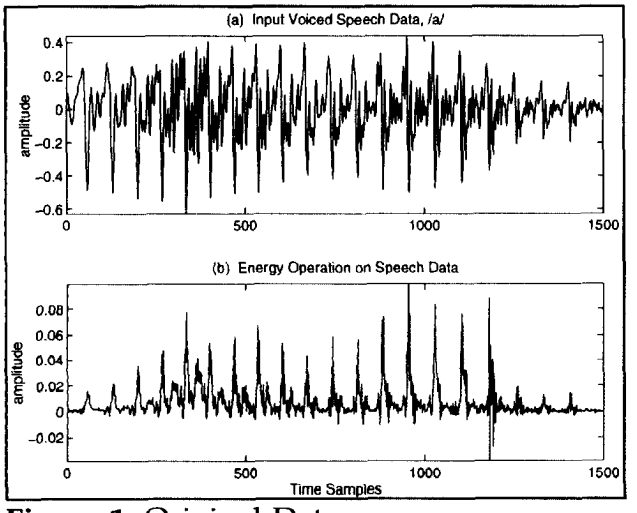

Figure 1: Original Data

The speech data is low-pass filtered, $s_{LP}(n)$, (Fig. 2a) with an FIR filter with a cut-off of 700 Hz. The discrete energy operator is applied to this data to produce the energy signal, $\Psi[s_{LP}(n)]$, (Fig. 2b). Bandlimiting the input data before the application of the energy operator results in data which primarily has glottal pulses. There will be some residual pulses remaining in the energy signal. These extraneous pulses will be processed out in the pitch estimation algorithm.

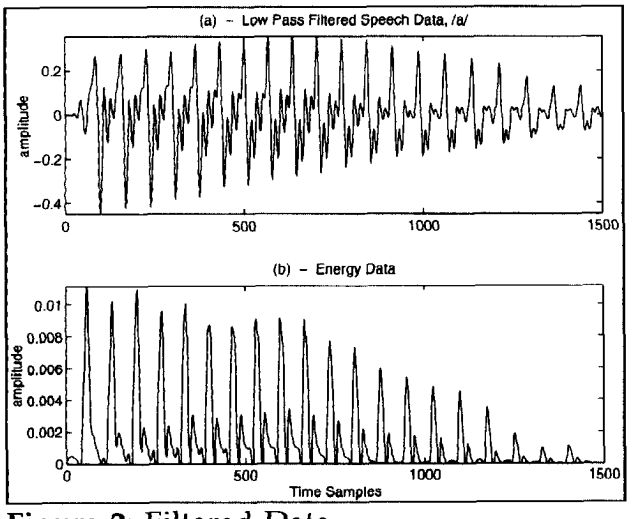

Figure 2: Filtered Data

Pitch Estimation Algorithm

The energy data, $\Psi[s_{LP}(n)]$, is center-clipped at 15% of the maximum (Fig. 3a). This eliminates low-level noise-like pulses. A peak-detection algorithm is employed which will identify the peaks and their locations. The data may also include spurious peaks between the glottal pulses. These extraneous peaks or pulses will be eliminated in order to obtain the location of the glottal pulses.

The purpose of the peak detection algorithm is to locate the position of the peaks of all the pulses, glottal and extraneous. The first step is to eliminate the positive sloped portion of the center-clipped energy signal. This is done to help in finding the location of pulse peaks. The signal that results consist of all the peaks and negative-sloped portions of the signal (Fig. 3b).

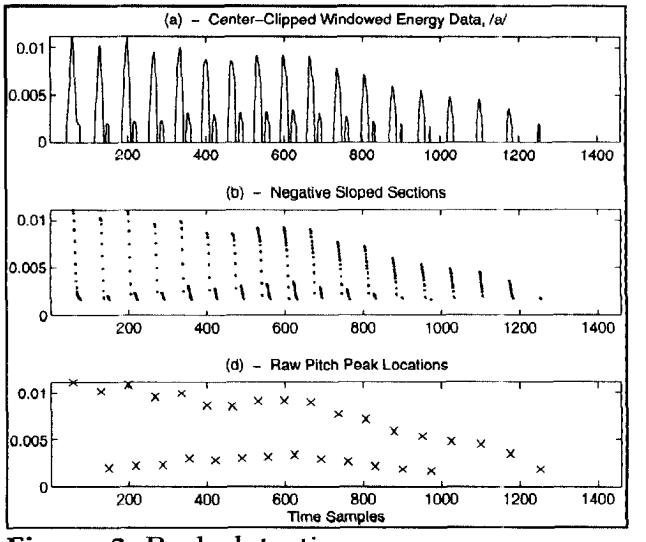

Figure 3: Peak detection process

This negative sloped data is scanned, alternately, for non-zero points, then for sequences of zero entries. With the center-clipping and elimination of the positive sloped segments, the data will initially be zeros when the scan is started. The first non-zero point encountered will be registered as a pulse peak location. After this, the data is scanned for the next sequence of at least five consecutive zeros following the non-zero points of a pulse. This five consecutive zeros criteria is used to eliminate subpulses or slope changes on the negative sloped side of glottal pulses. Experiments indicate that using the five consecutive zeros was effective in eliminating these subpulses.

The result of this process is the location of all the

peaks in the data (Fig. 3c). As previously stated, there could still be some extraneous pulses in this data, which are shown in Fig. 3c as a contour of x's at the bottom part of the plot. These will be eliminated in a later process.

An average pitch period duration is calculated from this peak location data. This parameter will be used as a point of comparison in the calculation of the individual pitch periods. The average pitch period is obtained by first finding the adjacent differences of center-clipped peaks, found in the previous step, at a threshold of 60% of maximum. Then an average is made of durations which were within 1 ms of one another.

Each individual pitch period was calculated by adjacent differences of the peak locations or indices. The calculated period is compared to the average pitch period duration. If the calculated period is smaller than the 80% of the average pitch period and the current pulse height is less than 40% of the preceding pulse height, then the current peak location is eliminated. A new period is calculated using the next pulse location. If the calculated duration is longer than 120% of the average pitch period, then the calculated duration is assigned a value that is 120% of the average.

When all the pitch periods are calculated and the extraneous pulses are eliminated, the pitch frequencies are calculated (Fig. 4a). Some of the estimated pitch frequencies might stray from the dominant pitch track. This could be due to a low sampling frequency, the type of vowel, or some noise induced artifact. To account for these points, the data are smoothed using two non-linear smoothing techniques. The pitch data is first smoothed using an order three median filter (Fig. 4b). This is followed by a three-point Hanning smoother (Fig. 4c). Figure 5 shows the result of running this algorithm on the vowel /a/.

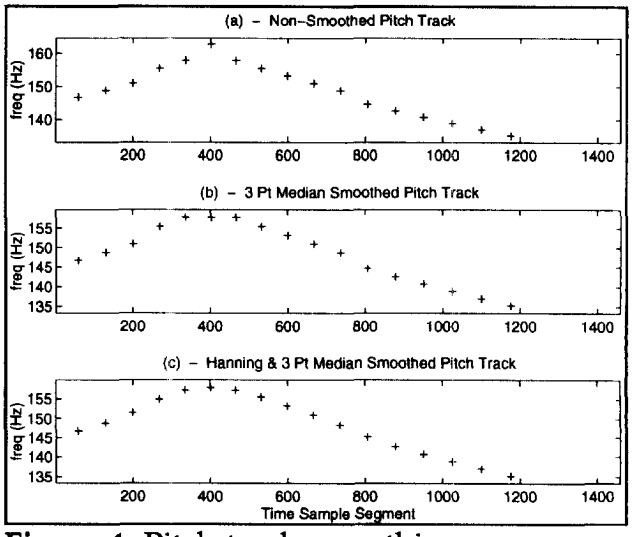

Figure 4: Pitch track smoothing

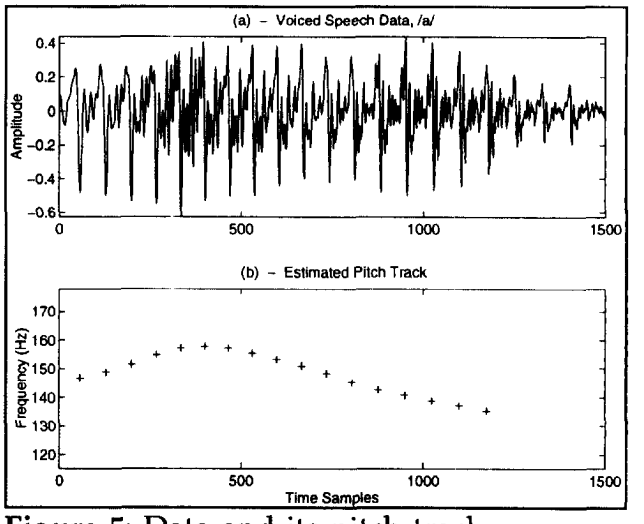

Figure 5: Data and its pitch track

Robustness to Added Noise

To test the robustness of this pitch detection algorithm to additive noise, voiced speech signals with signal-to-noise ratios (SNR) of 2 and 20 were used. The voiced data are from vowels with different points of articulation. The vowels used were: a reduced /i/ front vowel, /æ/ midvowel, and /u/ back vowel. The three vowels were chosen to test the effectiveness of this pitch on different vowels at various levels of SNR. The original speech data were acquired in a lab with no special damping of the background noise.

The signal power for the voiced speech was calculated using the average power within a window of data. Since the power for a speech signal is not constant, an average power was calculated over the voiced segment. From this average power, the scale factor was determined for the speech data to obtain the desired SNR. The noise was generated by MATLAB® Gaussian noise generator. This produces a signal that approximates Gaussian noise with zero mean and variance of one.

For the three vowels tested, this pitch detection

algorithm worked well for high SNR figures, as shown in Figures 6b and 7b. The pitch estimator seemed to perform well with noisy mid and back vowels (Fig. 6d and 7d). Some of the glottal pulses were missed in Figure 7d, which results in a lower pitch frequency. This algorithm had problems with a noisy reduced vowel /i/ (Fig. 8). With added noise, the magnitude of the noise pulses became as large or larger than the glottal pulses, so it was difficult for the peak detector to locate the correct pulses to calculate the pitch periods. The peak detection algorithm needs to be refined to ignore more numerous extraneous pulses in cases with significant noise.

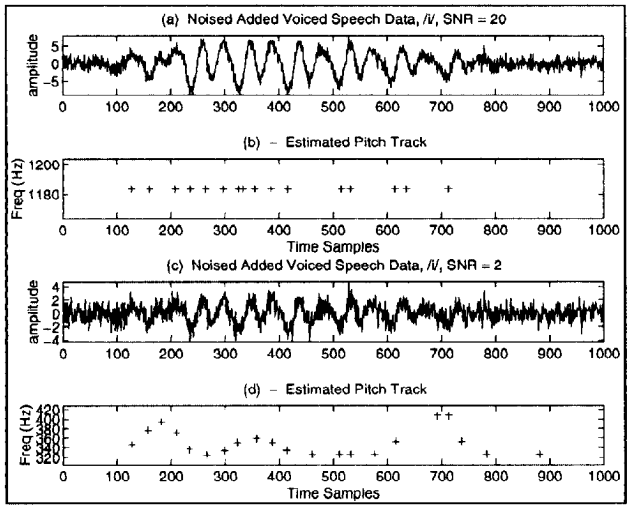

Figure 6: Noise added to /i/ data

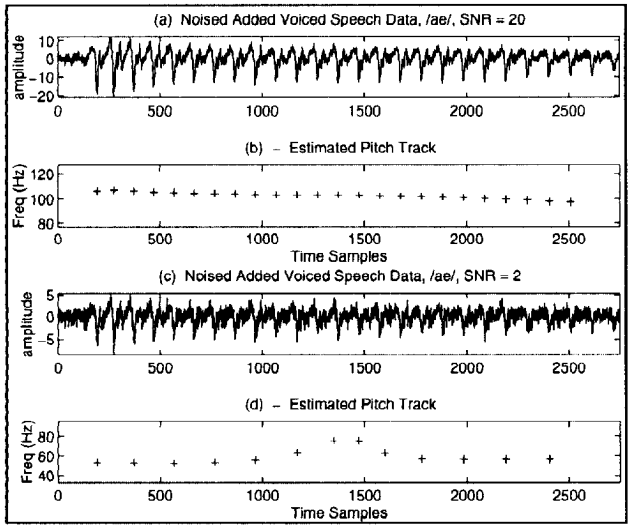

Figure 7: Noise added to /æ/ data

Complexity of the pitch algorithm

This method of estimating the pitch frequencies of voiced speech data did not require transforming the data, i.e. using FFT's or cepstrums. The data was filtered by an order 81 FIR filter. The operations involved in the energy operations are two multiplications and one subtraction for each data point. The amount of computation in the pitch estimation algorithm was low; this is primarily performing adjacent differences.

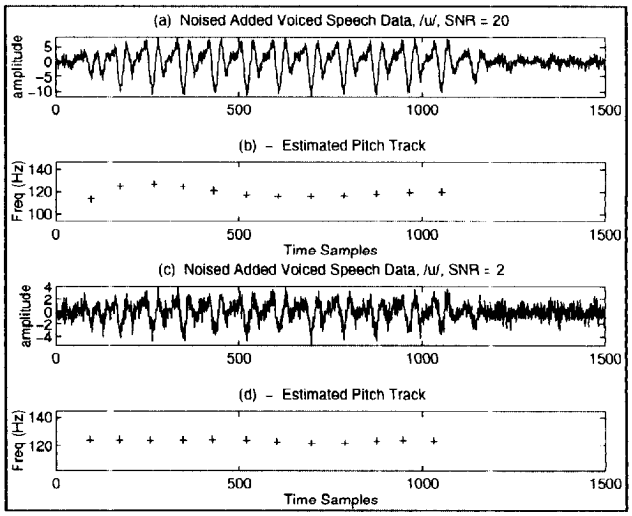

Figure 8: Noise added to /u/ data

Problems with this Method of Estimation

There were several problems encountered in using this method of pitch detection. The fixed thresholds for setting up the data for various calculations (e.g. center-clipping energy data or the extraneous pulse elimination) were speaker dependent. The variation between different speakers cause problems in peak detection algorithm. If the glottal pulses in the voice onset is below a given threshold, then the detection algorithm cut these out. Similarly, in the trailing part of voicing, the glottal pulses are significantly attenuated, therefore usually fall below the threshold. The number of missed pulses are not many, one or two pulses at the onset and at the end of voicing. Some of the vibrations at the end of voicing may not be the glottal excitation but rather the latent vibrations within the vocal tract. If this is the case, then the fixed threshold level will exclude this pulses.

This method of pitch detection and estimation does not resolve the problem of having a formant frequency in the vicinity of the pitch frequency. A low first formant frequency, F_1, will not be distinguished from the pitch frequency. Example of vowels with low first formant frequencies are /i/ and /u/.

One of the problems in using the Teager energy operator on the filtered data is that when the data are analyzed, segments of it could come close to containing a single component sinusoid. Since the Teager energy operator returns a constant value for single frequency component sinusoid, the problem is that the energy data peaks broaden. So, in searching for the peak, the result may not be an accurate determination of the glottal pulse peak. The peaks are still present, but it becomes difficult to estimate the pitch pulses. The energy pulses are no longer vividly defined, therefore the estimate returned may not be accurate.

Conclusion

In this paper, a time domain method of estimating the pitch frequencies of voiced speech data was described. The method is straightforward, involving no transformations of the data. The analysis was done using low-pass filtering, a non-linear energy operation, and non-linear smoothing of the pitch estimates. The results of the study show that it is a viable estimator of pitch periods. The advantage of using this technique is that it simplifies implementation by eliminating the need for regularly-spaced windowing and is not computationally intensive. The post-filtering or data-smoothing stages require calculation only on the points representing the pitch frequencies. The algorithm is still useful for high SNR speech signals.

Acknowledgement

This work was supported by NSF Grant MIP-9106126.

References

[1] B. Gold, "Computer Program for Pitch Extraction," J. Acoust. Soc. of America, Vol. 34, No. 7, pp.916-921, 1962.

[2] B. Gold and L.R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. of America, Vol. 46, No. 2, pt. 2, pp. 442-448, August 1969.

[3] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal," in Proceedings IEEE ICASSP-90, Albuquerque, NM, April 1990, pp. 381-384.

[4] J.F. Kaiser, "Some Useful Properties of Teager's Energy Operators," in Proceedings of IEEE ICASSP-93, Minneapolis, MN, April 1993, Vol. III, pp. 149-152.

[5] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "On Amplitude and Frequency Demodulation Using Energy Operators," IEEE Trans. Signal Processing, vol.41, pp.1532-1550, April 1993.

[6] 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, and Signal Proc., Vol. ASSP-24, No. 5, pp. 399-418, October 1976.

[7] M.J. Ross, H.L.Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans., Acoust., Speech and Signal Proc., Vol. ASSP-22, pp. 353-362, October 1974.

[8] R.K Whitman and D.M. Etter, "Initial Investigation in using an Energy Operator for Pitch Estimation," abstract in Proceedings of 127th meeting of Acoust. Society of America, June 1994.