Real Time Fundamental Frequency Analysis Using the Spectral Comb Method

P.J. Martin
Toronto, Canada

1. Introduction

Although reasonably good real-time fundamental frequency visualizers are now commercially available, it is difficult to conduct a phonetic analysis of prosodic parameters when speech material is recorded in a noisy environment. Indeed, most Fo analyzers are designed to work with a relatively high signal to noise ratio, which rules out their use in certain cases such as the analysis of spontaneous speech recorded in a noisy street. This is due to the fact that these instruments rely on methods of analysis which take into account fundamental frequency as the only parameter.

More reliable methods based on the analysis of the short-time spectrum of the speech signal are now available due to the development of inexpensive highspeed hardware.

2. The Spectral Comb Method

Most methods of pitch extraction from the short- time spectrum are based on detecting the periodicity of F_0 harmonics in the spectrum. The cepstrum (Noll, 1969), for instance, computes the Fourier transform of the logarithm of the power spectrum. Other methods are based on a more direct search for periodicity in the spectrum. Schroeder (1968) uses a histogram of subharmonics derived from spectral peaks, and F_0 is considered as being the smallest common multiple of the periods of its harmonic components. Harris and Weiss (1963) examine a high resolution Fourier spectrum and retain as fundamental frequency the most numerous equal spacings of adjacent peaks. Sreenivas and Rao (1972) use only selected high quality peaks, and compute their approximate highest common factor to obtain the pitch value. Sluyter, Kotmans and Leuwaarden (1980), in order to account for possible phase distortion of the peak harmonics in the spectrum, use a minimum distance criterion to recognize harmonic patterns, and derive pitch value from these patterns.

The spectral comb method (Martin, 1981) is also based on short-time spectral analysis. Although it is similar to other spectral compression techniques, reliability is due essentially to the fact that both the frequency and the amplitude of each harmonic component are considered in the evaluation of

F_0. More precisely, the comb method uses the cross correlation between the short-time spectrum $/F(\omega)/$ and a spectral comb function $P(\omega_p,\omega)$, with teeth of decreasing amplitude and variable intervals. The maximum of this crosscorrelation function is obtained when the comb's teeth coincide with the harmonic peaks of the spectrum, i.e. to the harmonic pattern which has the largest sum of its harmonic amplitude.

Although presenting some similarities with the method based on Goldstein's work (Goldstein, 1973), by Duifhuis et al. (1978), the spectral comb method differs from the harmonic sieve approach by using information pertaining to the amplitudes of the harmonics rather than using a minimum distance criterion in order to select the appropriate Fo.

3. Hardware Implementation

After analog-to-digital conversion, (4 kHz sampling frequency), the speech signal is split up into overlapping segments of 32 ms duration (128 samples). The speech samples of each segment are multiplied by 64×128 complex Fourier coefficients stored in a 16 k byte PROM, and accumulated into a 125 ns cycle time 12×12 multiplier-accumulator chip. (Gaussian windowing is performed on the trigonometric functions themselves, rather than on the speech signal). The module of the real and imaginary frequency values is computed from a table, and the logarithm of the module stored in the memory of a small microprocessor system (8085). These values range from 0 to 1024 Hz, with a 16 Hz resolution. An algorithm determines the spectral peaks above a threshold situated 40 dB below the maximum amplitude of the spectrum. These spectral peaks are 'rewritten' as parabolas whose peaks coincide exactly with the interpolated spectral peaks in order to obtain a modified power spectrum $/F'(\omega)/$ with reduced energy between the harmonic components. The modified spectrum is then crosscorrelated with a comb function with teeth of decreasing amplitude (law $n^{-1/8}$)

$$C(\omega_p,\omega) = \Sigma\, n^{-1/8}\,(n\omega_p - \omega)$$

the crosscorrelation function

$$I(\omega_p) = \Sigma\, n^{-1/8}\, /F'(\omega_p)/$$

is thus the sum of equally sampled values of the modified power spectrum $/F'(\omega)/$ which are weighted according to the amplitude of the comb's teeth.

The computation of $I(\omega_p)$ is performed by the microprocessor using a large table of coefficients containing all the possible parabola sampled values of 128 possible amplitude levels. Fast computation can thus be realized performing only sums and no multiplications. The final F_0 value is obtained by looking at the maximum of the crosscorrelation function $I(\omega_p)$.

The voiced-unvoiced distinction conducted by using an index compares

the maximum value of I (ω_p) to the value obtained with a comb function shifted by $\omega_p/2$. This ratio indicates the relative energy of the detected harmonic structure compared to the weighted sum of noise components.

4. Visualization

The F_0 values are displayed on a color monitor, together woth the intensity curve (which is measured independently). The display converter generates a color video signal containing both curves and alphanumeric information pertaining to display duration and frequency scale. Two cursors allow alphanumeric readout of frequency, intensity and duration values at any point along the curve.

Pitch visualization would be of even greater usefulness in phonetic research if the user were able to segment the speech signal on the screen and identify each segment. Although the speech signal can sometimes be segmented by simply looking at the information given by the intensity and fundamental frequency curves, one needs still more information in order to perform an accurate segmentation.

Usually, this information is provided by the speech wave. Since the resolution of the display (256 dots horizontally) does not allow for displaying the speech wave itself, tne envelope of the signal is displayed, instead. F_0 analysis is pertormed in real time, and any length of speech (up to 9 s.) may be displayed on the screen. Two cursors can be moved along the time axis in order to segment the speech signal according to the cues provided by the envelope. Numerical information pertaining to the segment defined by the two cursors includes the amplitude of F_0 (in Hz), and intensity (in dB) variations, the time duration of the segment, as well as the average and the standard deviation of both the pitch and the intensity between the cursors. It is also possible to obtain absolute values of F_0 and intensity at any point of the curves. With this equipment, an accurate analysis of speech data can be realized quite rapidly, and can even be performed interactively in the presence of an informant if desired.

5. Conclusion

A new reliable real-time pitch-visualizer has been presented, using a noise resistant F_0 tracking algorithn. The instrument allows real-time display of both intensity and fundamental frequency, together with the envelope of the speech signal.. Numerical readout of F_0 and intensity is provided, which makes the instrument easy to use for practical phonetic analysis.

References

Duifhuis, H., Willems, L.F. and Sluyter, R.J. (1978). Measuring Pitch in Speech. *Inst. Perceptie Onderzoek,* Annual Progress Report no **13**, 24-30.

Goldstein, J.L. (1973). An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones. *JASA,* Vol. **54**, 1496-1516.

Harris, C.M. and Weiss, M.R. (1963). Pitch Extraction by Computer Processing of High Resolution Fourier Analysis Data. *JASA,* Vol. **35**, 339-343.

Martin, P. (1981). Comparison of Pitch Detection by Cepstrum and Spectral Comb Analysis. *Proceedings of the Int. Conf. on Acoustic, Speech and Signal Processing,* Vol. **1**, 180-183.

Noll A.M. (1969). Short-time Pitch Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection. *JASA,* Vol. **36**, 296-302.

Schroeder, M.R. (1968). Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement. *JASA,* Vol. **43**, 829-834.

Sluyter, R.J., Kotmans, H.J. and Leuwaarden, A.V. (1980). A Novel Method for Pitch Extraction from Speech and a Hardware Model Applicable to Vocoder Systems. *Proceedings of the Int. Conf. on Acoustic, Speech and Signal Processing* 80, Vol. **1**, 45-48.

Sreenivas, T.V. and Rao, P.V.S. (1979). Pitch Extraction from corrupted Harmonics of the Power Spectrum. *JASA,* Vol. **65**, 223-228.