PITCH DETECTION
USING THE SHORT-TERM PHASE SPECTRUM

F.J. CHARPENTIER

Centre National d'Etudes des Telecommunications
22301 LANNION FRANCE

ABSTRACT

A new frequency domain method for determining the fundamental frequency of speech is presented in this paper. This method uses the information contained in the short-term phase spectrum whereas the previous methods were limited to the amplitude spectrum. The short-term spectrum is computed by DFT and it is interpreted as the output of a bank of band-pass overlapping filters. Harmonic components are detected by searching for sets of three contiguous filters having the same instantaneous frequency. The frequency of a detected harmonic is given by the instantaneous frequency itself. A conventional harmonic numbering algorithm is used to convert the set of detected harmonics to a value of the fundamental frequency. Preliminary results show the validity of the method.

INTRODUCTION

The importance of the phase information is commonly aknowledged in short-time Fourier analysis-synthesis of speech [1,2]. The short-time Fourier transform (STFT) as a function of time is conveniently described by the time-varying amplitude and phase spectra. A useful alternative representation of the phase spectrum consists of its first-order time derivative, called the instantaneous frequency distribution (IFD). In the case of narrow-band spectral analysis, it is well-known that the IFD carries information about the vocal-tract excitation [1], and more specifically about the fundamental frequency itself [3,4]. In many applications, where appropriate modifications of the STFT provide ways of manipulating some of the speech parameters, the IFD must be manipulated with much care. For instance, in order to change the rate of natural speech, the amplitudes and instantaneous frequencies must be preserved, but along a different time-scale [1,5]. The pitch also can be modified by a constant factor if the instantaneous frequency is multiplied by the same factor [6]. Since the IFD conveys much of the vocal-tract excitation information, its proper manipulation is crucial for preserving the naturalness of speech. Inversely, artificial sounding speech can easily be obtained by assigning arbitrary values to the IFD. In this paper, it is shown that the IFD is sufficient for an exact determination of the pitch.

The usefulness of the IFD for wide-band spectral analysis of speech was recently demonstrated by Friedman [7]. By using a graphical representation of the IFD information, accurate and smooth tracks of the speech formants could be obtained. The method for evaluating the IFD necessitates the computation of two STFTs, using the analysis window and its first-order derivative. The method is equivalent to an evaluation of the phase shifts due to an infinitesimal displacement of the time window. The same approach is taken in this paper, from a digital point of view. But in the specific conditions of a narrow-band analysis using a raised-cosine (Hanning) window, the computation of one STFT will prove to be sufficient.

After the IFD has been computed, the problem arises of how to use this information in an proper manner. The method proposed here is similar to a method proposed by Ghitza in the case of an auditory model [8]. In this method, a spectral envelope is estimated from an auditory model filter bank, by determining the "in-synchrony" regions of the spectrum, defined as the groups of contiguous filters exhibiting a same dominant frequency component. The algorithm proposed in this paper concerns the detection of pitch harmonics but it can be seen as an application of the in-synchrony concept to a DFT filter bank in the case of a narrow-band spectral analysis.

ESTIMATING THE INSTANTANEOUS FREQUENCIES

When the STFT is computed by DFT, it is usefully interpreted as the output of an uniform filter bank, with the input being the speech signal itself. The coefficients of the STFT are given by:

$$X_k(n) = \sum_{m=0}^{N-1} w^{-km}\, x(n+m)\, h(m)$$

where w denotes the N-th complex root of unity, x(n) the speech signal and h(n) an analysis window. Each coefficient Xk is interpreted as the output of a bandpass filter located around the center frequency:

$$f_k = (k/N)\, F_s$$

where Fs is the sampling frequency of the signal.

ICASSP 86, TOKYO CH2243-4/86/0000-0113 $1.00 © 1986 IEEE

The bandwidth of each "filter/coefficient" Xk is determined by the analysis window h(n) itself. In the case of a Hanning window, the frequency response of Xk spreads over three contiguous frequency points. Consequently, the response of a given Xk coefficient overlaps the responses of the two adjacent coefficients.

We now adopt a slightly modified definition of the instantaneous frequency, in order to obtain the true dimension of a frequency. Therefore, the instantaneous frequency f'k of a coefficient Xk will be defined as the time derivative of its phase, divided by 2π . In a practical implementation, the determination of f'k implies the computation of the STFT at two successive instants. This corresponds to a small displacement of the analysis window by one sample in time, introducing a phase shift $\Delta\varphi_k$ for each Xk coefficient. The instantaneous frequency can then be approximated by the formula:

$$f'_k = \frac{\Delta\varphi_k}{2\pi} F_s$$

A complete computation of two successive phase spectra can be avoided by using the dependency relations existing between two successive STFTs. The algorithm is organized as follows. First, a DFT is computed without using a window:

$$Y_k = \sum_{m=0}^{N-1} w^{-km}\, x(n+m)$$

Then, the coefficients Xk are obtained from the Yk by performing the Hanning windowing in the frequency domain:

$$X_k(n) = Y_k - (Y_{k-1} + Y_{k+1})/2$$

If the analysis is performed one sample earlier, the STFT coefficient can be obtained by the following formula:

$$X_k(n-1) = w^{-k}\left[Y_k - (wY_{k-1} + w^{-1}Y_{k+1})/2 \right]$$

In fact, this formula can be interpreted in two steps: first, it performs the convolution of the DFT by the response of the Hanning window shifted backwards by one sample, and then it moves the time origin one sample back in time. Finally, the phase shift for the filter Xk is given by:

$$\Delta\varphi_k = 2\pi(k/N) - \mathrm{Arg}\left[\frac{2Y_k - wY_{k-1} - w^{-1}Y_{k+1}}{2Y_k - Y_{k-1} - Y_{k+1}} \right]$$

Thus, the computational load for evaluating the IFD consists mainly of one DFT, N complex divisions and N phase extractions. In a pitch detection application, the computational load can be further reduced by limiting the analysis to a low frequency range (typically from 0 to 1.6 kHz), and by simultaneously using the information of the amplitude spectrum to restrict the calculations to the neighbourhood of the spectral peaks.

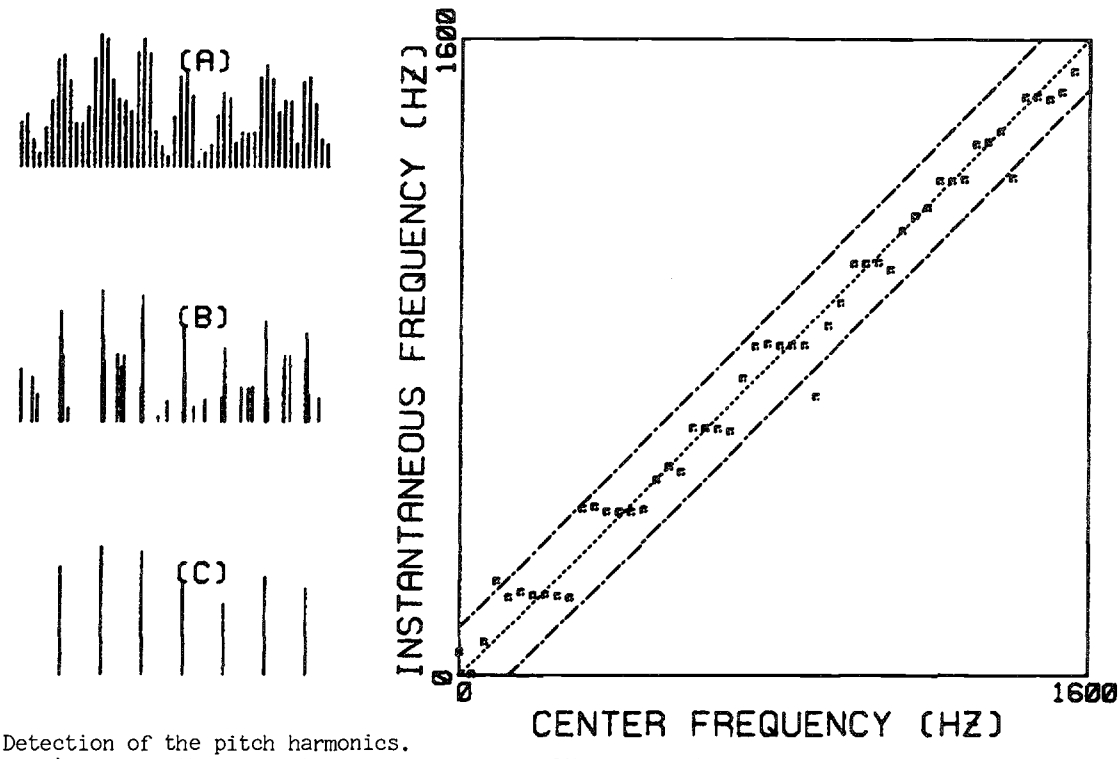

Fig.1 Detection of the pitch harmonics.
(right) plot of the instantaneous frequencies f'k versus the center frequencies fk
(left) amplitude spectrum using (A) the usual frequency scale
(B) a frequency scale warped by the f'k = f'(fk) mapping
(C) the same as (B) with elimination of the non-harmonic components

3. 9. 2

DETECTION OF THE FUNDAMENTAL FREQUENCY

Harmonic detection

The harmonic detection is based on an appropriate treatment of the IFD information. The length of the analysis window is determined so that the bandwidth of the window response is inferior to the fundamental frequency. Therefore, there will be at most one harmonic passing through a given coefficient Xk. In the case of a strong harmonic component, the instantaneous frequency f'k will reflect accurately the harmonic frequency. Moreover, the harmonic component will also be present in the coefficients Xk-1 and Xk+1 because of their frequency overlap with Xk.

This point of view is illustrated in Fig.1. In the right part of Fig.1, the IFD is displayed as a function of the central frequencies fk of the Xk coefficients. The resulting plot is a distribution of points around the dotted line f'k = fk . The presence of a strong harmonic is revealed by a local accumulation of the points around a particular value of the instantaneous frequency. In the left part of Fig.1, the corresponding amplitude spectrum is given using three different representations. In representation (A) the coefficients Xk are distributed uniformly along the usual frequency scale, according to their center frequencies fk. In representation (B) they are distributed according to their instantaneous frequencies f'k, and they tend to cluster around the harmonics frequencies. Representation (C) is obtained by applying to (B) the harmonic detection algorithm, so that only the true harmonic components are retained.

The harmonic detection algorithm works in two steps:

(1) it determines all the coefficients Xk containing a strong harmonic; the harmonicity criterion consists of selecting a filter Xk if the instantaneous frequencies f'k-1 and f'k+1 of the two adjacent filters are sufficiently close to f'k.

(2) it specifies the values of the corresponding harmonics: these values are simply given by the instantaneous frequencies f'k of the coefficients Xk selected in step (1).

A distribution of harmonics can therefore be obtained, and when the analysis is repeated at a certain rate, a "harmonogram", i.e. a spectrographic representation of the harmonics, can be displayed (Fig.2).

Estimation of the fundamental frequency

The estimation of the pitch value is obtained from the distribution of harmonics previously detected. Many algorithms have been proposed for similar tasks [9]. The algorithm used here was derived from the "sieve" or "comb" algorithms [10,11], and it consists of finding an appropriate numbering for a maximum number of detected harmonics. To perform the voiced/unvoiced decision, a voicing criterion is defined as the sum of the energies of the successfully numbered harmonics.

RESULTS AND DISCUSSION

The algorithm was tested on speech digitized at a 8 kHz sampling rate. A 256-points Hanning window was used, corresponding roughly to a 30 ms duration. The Fig.2 presents the result of our algorithm on a French sentence spoken by a male speaker. Smooth tracks of the harmonics can be observed on the harmonogram. A similar representation can also be obtained by plotting the frequencies of the amplitude peaks of the spectrum. In this case, the harmonic frequencies must be approximated by use of a parabolic interpolation scheme, to obtain smooth tracks comparable to those in Fig.2. The harmonics detected by the harmonicity criterion generally coincide with spectral peaks, but their number is much smaller than the total number of peaks. In other words, the harmonicity criterion selects a small number of spectral maxima, those that are the most likely to reflect a true harmonic. Because of this selectiveness the harmonicity criterion can be viewed as a quality criterion for the spectral peaks. Such criteria have already been proposed from the amplitude information itself [11-13]. Those criteria are designed to retain the most reliable peaks and to eliminate the spurious ones. Terhardt has proposed a psychoacoustic criterion accounting for the masking phenomena between neighbouring harmonics [12]. Duifhuis et Willems have proposed a criterion, in which peaks are eliminated when they do not match with a local parabolic model of the amplitude [11]. In our case, spectral peaks are discarded when the corresponding DFT coefficients do not exhibit a certain phase coherence. The voicing criterion represented in Fig.2 follows roughly a bimodal distribution, providing an easy way of distinghishing between the voiced and unvoiced portions. The pitch contour computed by our method and displayed in Fig.2 is quite satisfactory. By comparing the results of our method and those of the "comb" method, a good correspondence could be observed.

CONCLUDING REMARKS

Most frequency domain methods for determining the pitch are limited to the analysis of the amplitude spectrum. The harmonics are generally located at the spectral peaks. By using the phase information, the method presented in this paper is capable of a more selective detection of the harmonics and it provides a direct estimation of their frequencies, avoiding the need for interpolation schemes. The computational load comprises basically one FFT and the extraction of a phase spectrum, but due to further computational savings, the complexity of the method is comparable to that of the other frequency domain methods. Preliminary results indicate the validity of the method for nondegraded speech signals. An objective comparison with other methods remains necessary in order to evaluate the real benefit of combining the phase and the amplitude information in a pitch detection process.

REFERENCES

[1] J.L. Flanagan, R. Golden, "Phase vocoder", Bell Syst. Tech. J., 45, 1494-1509, Nov.66

[2] A.V. Oppenheim, J.S. Lim, "The importance of phase in signals", Proc. IEEE, 69(5), 529-541, May 81

[3] M.R. Portnoff, "Short-time Fourier analysis of sampled speech", IEEE Trans. ASSP, 29(3), 364-373, Jun.81

[4] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time-scaling of speech signals", IEEE Trans. ASSP, 27(2), 121-133, Apr.79

[5] M.R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis" IEEE Trans. ASSP, 29(3), 374-390, Jun.81

[6] S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", IEEE Trans. ASSP, 30(4), 566-578, Aug.82

[7] D.H. Friedman, "Instantaneous-frequency distribution versus time: an interpretation of the phase structure of speech", Proc. ICASSP, 1121-1124, Mar.85

[8] O. Ghitza, "A measure of in-synchrony regions in the auditory nerve firing patterns as a basis for speech vocoding", Proc. ICASSP, 505-508, Mar.85

[9] W. Hesse, "Pitch determination of speech signals - Algorithms and devices", Springer Verlag, 1983

[10] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis", Proc. ICASSP, 180-183, May 82

[11] H. Duifhuis et al., "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception", J. Acoust. Soc. Am., 71(6), 1568-1580, Jun.82

[12] E. Terhardt, "Calculating virtual pitch", Hearing Research, 1, 155-182, 1979

[13] S. Seneff, "Real-time harmonic pitch detector", IEEE Trans. ASSP, 26(4), 358-365, Aug.78

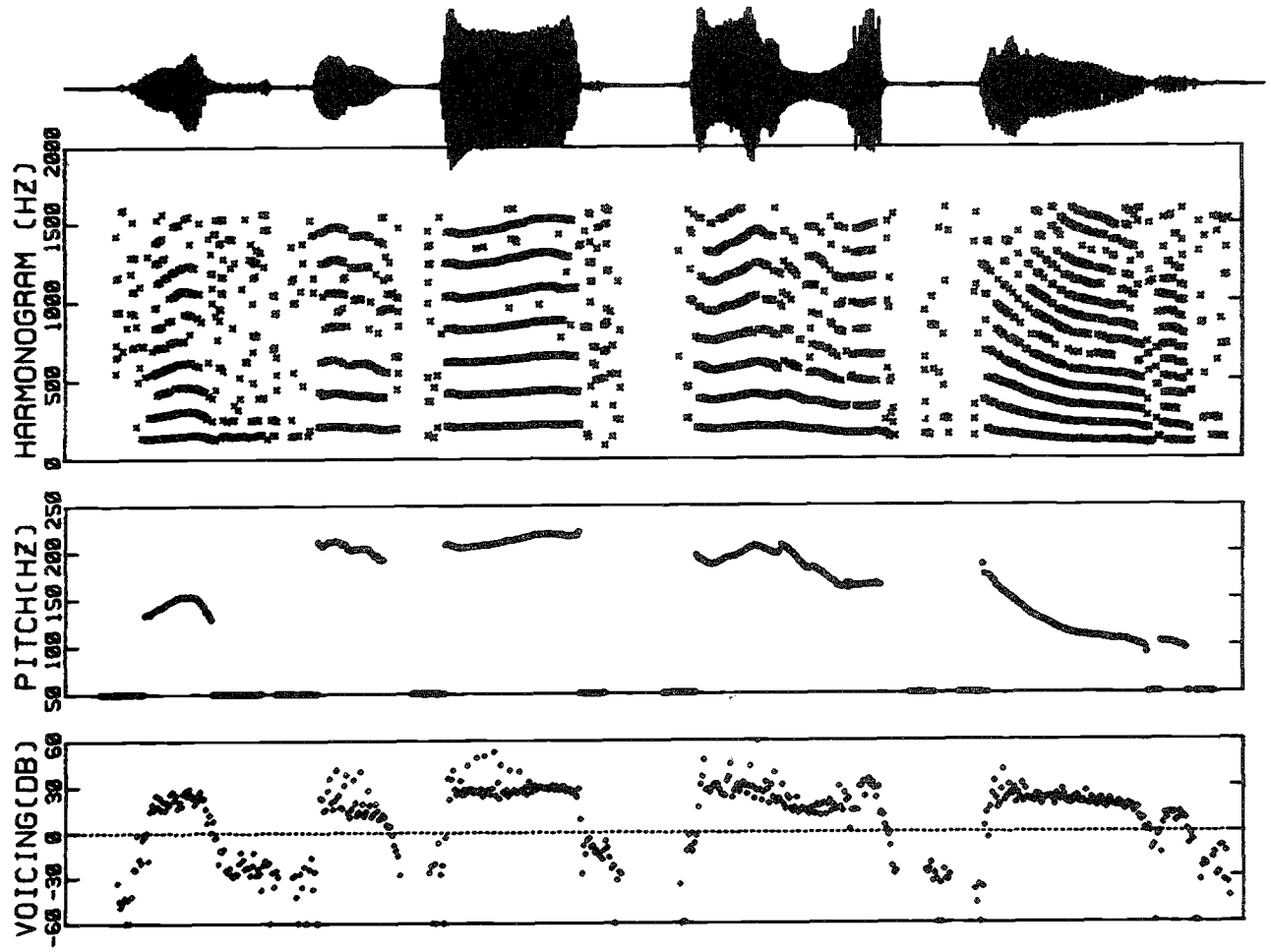

Fig.2 Pitch detection for a sentence pronounced by a male speaker.
In addition to the pitch contour, the harmonogram (display of the harmonic component tracks) and the voicing criterion have also been represented.