Super Resolution Pitch Determination of Speech Signals

Yoav Medan, Eyal Yair, and Dan Chazan

Abstract—**Based on a new similarity model for the voice excitation process, a novel pitch determination procedure is derived. The unique features of the proposed algorithm are infinite (super) resolution, better accuracy than the difference limen for F_0, robustness to noise, reliability, and modest computational complexity. The algorithm is instrumental to speech processing applications which require pitch synchronous spectral analysis.**

I. Introduction

PITCH determination is considered one of the most difficult tasks in speech processing. Many pitch determination algorithms (PDA's) were proposed, both in the time and the frequency domains (e.g., [1]–[7]). The most comprehensive survey of PDA's is presented in [8], where it is claimed that "we do not have a single pitch determination algorithm which operates reliably (and accurately) for all applications. . . " [8, p. 521]. The complexity of pitch determination stems from the variability and irregularity of speech that is characterized by the following parameters:

1) *Spectrum:* The spectral content of the signal changes constantly, depending on the articulation of the different sounds. This reduces the degree of similarity between successive segments of speech, both within the duration of a specific sound as well as during the transition interval between distinct sounds. The situation is further complicated by the existence of *unvoiced* sounds that do not exhibit a periodic structure associated with pitch.

2) *Intensity:* Even with a constant articulation, the glottal pulses are not uniform in terms of their amplitude (volume) and thus, the resulting speech signal is amplitude modulated.

3) *Pitch:* The elapsed time between two successive glottal pulses (i.e., pitch interval) is controlled by the tension of the vocal folds and the buildup of air pressure from the lungs. It may vary, leading to pitch variations in the range of 2–10%, between two successive periods [8], [13].

In order to combat the nonstationarity, short analysis windows must be used. However, due to the wide range of possible pitch values (3–4 octaves), the analysis window may contain several periods, as well as a mixture of voiced (V) and unvoiced (UV) segments, leading to an average or even erroneous indication of pitch.

An additional error source which limits the resolution, and hence the accuracy, of common PDA's is time discretization (quantization) of the pitch estimates, introduced by sampling the speech signal asynchronously with the instantaneous pitch value. A pitch estimate, expressed as an integer multiple of the sampling interval, contains a time quantization error which may lead to audible distortions in speech coding applications [9].

The PDA introduced in this paper overcomes most of these difficulties by introducing a new model for the similarity in the pitch process. This model allows quantifying the degree of similarity between exactly two adjacent and nonoverlapping pitch intervals, with an infinite time resolution. The similarity model takes into account the intensity modulation that may exist between successive periods to yield an instantaneous value of the pitch interval. The resulting algorithm offers a robust, high-resolution, and efficient implementation scheme which is capable of avoiding the audible distortions associated with common pitch based speech coding techniques. Given the high resolution and accuracy of the estimated pitch values afforded by the new approach, it was possible to develop an efficient and accurate pitch synchronous spectral analysis scheme, that was used in various speech processing applications [10], [11].

The paper is organized as follows. The basic PDA kernel is derived in Section II. Pitch tracking procedure is outlined in Section III, and Section IV presents a computational complexity analysis of the proposed PDA. Performance evaluation results are presented in Section V followed by a brief discussion and summary.

II. Pitch Determination

For each time instant t_0 we define two signals $x_\tau(t, t_0)$ and $y_\tau(t, t_0)$ as follows:

$$x_\tau(t, t_0) = s(t)w_\tau(t - t_0)$$
$$y_\tau(t, t_0) = s(t + \tau)w_\tau(t - t_0) \qquad (2.1)$$

where $s(t)$ denotes the speech signal, and $w_\tau(t)$ is a rectangular window of length τ seconds given by $w_\tau(t) = 1$ for $0 \le t < \tau$ and $w_\tau(t) = 0$ otherwise. Note that the above two signals are nonzero only inside the interval $[t_0, t_0 + \tau]$. The definition of (2.1) extracts two adjacent speech segments of duration τ seconds which are aligned on the time axis to reside in $[t_0, t_0 + \tau]$. These segments will be referred to as the x and y segments for time instant t_0.

Now consider a speech frame that starts at $t = t_0$ and consists of exactly two pitch periods of duration $\tau = T_0$, where $x_{T_0}(t, t_0)$ is the first period, $y_{T_0}(t, t_0)$ is the second period, and T_0 denotes the pitch period (in seconds) related to the time instant $t = t_0$. Since it is assumed that the similarity between two successive pitch periods is high, they can be assumed to be amplitude modulated versions of each other, as expressed by the following similarity model:

$$x_{T_0}(t, t_0) = a(t_0)\, y_{T_0}(t, t_0) + e(t, t_0) \qquad (2.2)$$

where $a(t_0)$ is an unknown, positive amplitude modulation factor (gain) at time t_0 which reflects the change in the glottal pulse

Manuscript received February 21, 1989; revised January 8, 1990.

The authors are with IBM Israel Science and Technology Center, Technion City, Haifa 32000, Israel.

IEEE Log Number 9040372.

1053-587X/91/0100-0040$01.00 © 1991 IEEE

volume. The error term $e(t, t_0)$ reflects other dissimilarities between the two periods. To maximize similarity between the two segments $x_\tau(t, t_0)$ and $y_\tau(t, t_0)$, the time interval $\tau = T_0$ for which $e(t, t_0)$ is minimized over the time interval $[t_0, t_0 + \tau]$ (according to some desired norm), is defined as the pitch period at the time instant $t = t_0$. Minimizing the normalized squared error yields the following optimization problem:

$$T_0 = \underset{\tau, a(t_0) > 0}{\operatorname{argmin}} \left\{ J = \frac{\int_{t_0}^{t_0+\tau} \left[x_\tau(t, t_0) - a(t_0)\, y_\tau(t, t_0)\right]^2 dt}{\int_{t_0}^{t_0+\tau} \left[x_\tau(t, t_0)\right]^2 dt} \right\}. \tag{2.3}$$

The normalization term is required to compensate for the variable size of the speech segments involved and uneven energy distribution over the pitch interval. The optimization of (2.3) can therefore be considered as a maximization of the signal-to-noise ratio. For practical reasons, τ may be restricted to the range of feasible pitch periods: $T_{0_{\min}} \le \tau \le T_{0_{\max}}$ according to the range of pitch expected in spoken speech.

In order to determine the optimal value of $a(t_0)$, the derivative of the cost function J with respect to $a(t_0)$ (for an arbitrary τ) is taken and compared to zero. Consequently, the optimal modulation gain is obtained as: $a(t_0) = (x, y)_\tau / |y|_\tau^2$, where $(x, y)_\tau$ is the inner product between $x_\tau(t, t_0)$ and $y_\tau(t, t_0)$ over the time interval $[t_0, t_0 + \tau]$ given by

$$(x, y)_\tau = \int_{t_0}^{t_0+\tau} x_\tau(t, t_0)\, y_\tau(t, t_0)\, dt$$

and $|y|_\tau^2 = (y, y)_\tau$ is the energy of the segment $y_\tau(t, t_0)$. By substituting the optimal value of $a(t_0)$ into (2.3), the cost function can be written as

$$J = 1 - \rho_\tau^2(x, y)$$

where $\rho_\tau(x, y)$ is the cross-correlation coefficient between the x and y segments

$$\rho_\tau(x, y) = \frac{(x, y)_\tau}{|x|_\tau |y|_\tau} \tag{2.4}$$

which is restricted to be positive since $a(t_0)$ was assumed to be positive. The pitch period T_0 at time t_0 can therefore be computed by the following maximization problem:

$$T_0 = \underset{\tau}{\operatorname{argmax}}\ \rho_\tau(x, y) \qquad \text{s.t.} \quad T_{0_{\min}} \le \tau \le T_{0_{\max}}. \tag{2.5}$$

Thus, minimization of the normalized mean square error in (2.3) is equivalent to a maximization of the cross-correlation function $\rho_\tau(x, y)$ (as a function of τ) over the range of feasible pitch values. Any specific value of the quantity τ is called the hypothesized pitch value at time $t = t_0$. Note that the function $\rho_\tau(x, y)$ should not be confused with an autocorrelation function that is a function of the relative time lag and not of the window length.

In light of (2.5), the minimization of (2.3) can be viewed, in a geometrical sense, as a minimization of the angle between the two "vectors" $x_\tau(t, t_0)$ and $y_\tau(t, t_0)$ (regardless of their magnitudes) over their varying "length" τ, where the cosine of that angle is expressed by the cross-correlation coefficient $\rho_\tau(x, y)$.

A. Integer Pitch Determination

A realizable solution of (2.5) can be obtained using digital techniques in which the signal $s(t)$ is sampled uniformly with a sampling interval T. The pitch period is obtained in this case with a finite resolution dictated by the sampling interval and is called hereafter the *integer pitch*.

Denote a vector of speech samples $s[n_1: n_2]$ (where $1 \le n_1 < n_2$) by

$$s[n_1: n_2] \triangleq (s_{n_1}, \cdots, s_{n_1+k}, \cdots, s_{n_2})^T; \qquad k = 0, 1, \cdots, n_2 - n_1 \tag{2.6}$$

where

$$s_i \triangleq s(t + t_0)\big|_{t=(i-1)LT}; \qquad i = 1, 2, \cdots.$$

The constant L is a decimation factor used to expedite the computation (as will be discussed in Section IV). It is assumed that when decimation is used, the signal is appropriately low-passed to enable the decimation. For convenience of reading the reader may assume in this section that $L = 1$.

Since the speech signal is sampled, the x and y segments of (2.1) cannot be obtained for all times t. Instead, they are replaced by two n-dimensional vectors $\boldsymbol{x}_n(i_0) = (x_1, \cdots, x_j, \cdots, x_n)^T$ and $\boldsymbol{y}_n(i_0)$ $(y_1, \cdots, y_j, \cdots, y_n)^T$ given by

$$\boldsymbol{x}_n(i_0) = s[1: n] \quad \text{and} \quad \boldsymbol{y}_n(i_0) = s[n + 1: 2n] \tag{2.7}$$

where the index i_0 indicates the sample index associated with the time instant t_0. The length n of these vectors is the hypothesized value of the integer pitch which is related to the hypothesized pitch τ of the continuous case through

$$(n - 1)LT < \tau \le nLT.$$

Based on (2.3), an optimal integer pitch period N_0 at time $t = t_0$ minimizes the following normalized discrete squared error function:

$$N_0 = \underset{n, a(t_0) > 0}{\operatorname{argmin}} \left\{ J = \frac{\sum_{j=1}^{n} \left[x_j - a(t_0)\, y_j\right]^2}{\sum_{j=1}^{n} x_j^2} \right\} \tag{2.8}$$

for the range $N_{\min} \le n \le N_{\max}$ of feasible integer pitch values, where $N_{\min}$ and $N_{\max}$ correspond to $T_{0_{\min}}$ and $T_{0_{\max}}$, respectively.

The optimization of (2.8) leads to a discrete version of (2.5) where a cross-correlation sequence $\rho_n(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0))$ between the vectors $\boldsymbol{x}_n(i_0)$ and $\boldsymbol{y}_n(i_0)$ has to be maximized for n over a finite range $[N_{\min}, N_{\max}]$

$$N_0 = \underset{n}{\operatorname{argmax}}\ \rho_n(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0)) \qquad \text{s.t.} \quad N_{\min} \le n \le N_{\max} \tag{2.9}$$

where the evaluation of $\rho_n(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0))$ is carried out similarly to (2.4), for which the inner product $(\boldsymbol{x}, \boldsymbol{y})_n$ is given by

$$(\boldsymbol{x}, \boldsymbol{y})_n = \sum_{j=1}^{n} x_j y_j.$$

The minimization of (2.9) can be carried out by evaluating $\rho_n(\boldsymbol{x}, \boldsymbol{y})$ for the full range of $[N_{\min}, N_{\max}]$ and picking its maximum, where the speech is prefiltered by a low-pass filter in order to remove high frequency components which are not necessary for tracking the pitch and may reduce the magnitude of

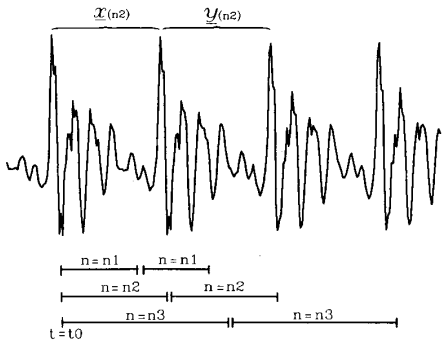


Fig. 1. Determination of the integer pitch. Two adjacent windows of variable length of n samples each, starting at $t = t_0$, are used to form the vectors x_n and y_n. The cross-correlation coefficient $\rho_n(x, y)$ is calculated over the range $N_{\min} \le n \le N_{\max}$. The integer pitch, N_0, corresponds to the value of n for which $\rho_n(x, y)$ attains its maximum. In this example $N_0 = n_2$.

$\rho_n(x, y)$. The procedure to determine the integer pitch is illustrated in Fig. 1. Other considerations and computational simplifications are discussed in Sections III and IV.

B. Exact Pitch Determination

The integer pitch N_0 has been estimated with a finite resolution which contains a sampling rate dependent "rounding" error. The exact pitch, expressed as a fractional number of samples, is denoted by N and is given by $N = T_0/LT$. We also denote the integer part of N by $\underline{N}$, and the pitch truncation error by $\beta = N - \underline{N}$, which can take any value in the range $0 \le \beta < 1$. In this subsection we introduce a procedure to obtain an estimate of the exact pitch with infinite resolution, regardless of the sampling interval, by estimating β and adding it to $\underline{N}$.

Ideally, when sampling the x and y segments of (2.1) to form the two vectors x and y, one would require that for segments of length T_0 the samples of the y segment would reside in the same relative location of the period as in the x segment. Such a sampling of the two periods would result in the highest value for the correlation coefficient between them. This requirement implies that there exists an n such that the vector $y_n(i_0)$ is aligned with the beginning of $y_{T_0}(t, t_0)$, meaning that the first element y_1 of $y_n(i_0)$ satisfies

$$y_1 = y_{T_0}(t_0, t_0) = s(t_0 + T_0). \tag{2.10}$$

Clearly, due to the fixed and arbitrary sampling rate, the exact pitch period N is generally not an integer number. Thus, in view of the definition of $y_n(i_0)$ in (2.7), one cannot find an integer n to comply with the condition of (2.10), unless by coincidence $N = N_0$.

To solve this synchronization problem we note that the first element of the vector $y_{\underline{N}}(i_0 + \beta)$ is aligned with $y_{T_0}(t_0, t_0)$, as required. The notation of $(i_0 + \beta)$ here means that the samples of $y_{\underline{N}}(i_0 + \beta)$ are delayed with respect to those of $y_{\underline{N}}(i_0)$ by βLT seconds. The elements of the vector $y_{\underline{N}}(i_0 + \beta)$, however, are not immediately available from the uniform sampling of $s(t)$, unless $\beta = 0$, but they can be estimated using interpolation techniques. Since a low-passed version of the speech signal is used for the pitch determination, where the bandwidth of the low-passed signal is much smaller than the sampling rate, a linear interpolation should suffice, and $y_{\underline{N}}(i_0 + \beta)$ can be approximated by a linear combination of two available vectors, $y_{\underline{N}}(i_0)$ and $y_{\underline{N}}(i_0 + 1)$, as follows:

$$y_{\underline{N}}(i_0 + \beta) \cong (1 - \beta) y_{\underline{N}}(i_0) + \beta y_{\underline{N}}(i_0 + 1). \tag{2.11}$$

If for $n = \underline{N}$, $y_{\underline{N}}(i_0)$ in (2.8) is substituted with $y_{\underline{N}}(i_0 + \beta)$, where the minimization is performed over $\beta \in [0, 1)$, then an optimal value of β, denoted by β^*, is computed, as in (2.9), by maximizing the cross-correlation coefficient between the vectors $x_{\underline{N}}(i_0)$ and $y_{\underline{N}}(i_0 + \beta)$:

$$\beta^* = \underset{\beta}{\operatorname{argmax}}\, \rho_{\underline{N}}\big(x(i_0), y(i_0 + \beta)\big)$$
$$\text{s.t.} \quad y_{\underline{N}}(i_0 + \beta) = (1 - \beta) y_{\underline{N}}(i_0) + \beta y_{\underline{N}}(i_0 + 1). \tag{2.12}$$

The maximum correlation will be obtained for a value of β^* such that the vectors $x_{\underline{N}}(i_0)$ and $y_{\underline{N}}(i_0 + \beta^*)$ are properly aligned with respect to the period as to satisfy (2.10). As outlined in Appendix A, the optimization in (2.12) can be carried out using the orthogonal projection theorem to yield the optimal value for β as

$$\beta^* = \frac{(x(i_0), y(i_0 + 1))\,|y(i_0)|^2 - (x(i_0), y(i_0))(y(i_0), y(i_0 + 1))}{(x(i_0), y(i_0 + 1))\big[|y(i_0)|^2 - (y(i_0), y(i_0 + 1))\big] + (x(i_0), y(i_0))\big[|y(i_0 + 1)|^2 - (y(i_0), y(i_0 + 1))\big]} \tag{2.13}$$

where the subscript $\underline{N}$ was omitted for simplicity.

Following the definition of β, the estimation of the exact pitch period is given by

$$\hat{T}_0 = (\underline{N} + \beta^*) LT. \tag{2.14}$$

The integer $\underline{N}$ is determined, based on the integer pitch value N_0, as follows:

$$\underline{N} = \min\{N_0, N_1\} \tag{2.15}$$

where N_1 is defined as the integer adjacent to N_0 for which $\rho_{N_0}(x(i_0), y(i_0)) - \rho_{N_1}(x(i_0), y(i_0))$ is minimal. That is

$$N_1 = N_0 + \operatorname{sign}\big[\rho_{N_0+1}(x(i_0), y(i_0)) - \rho_{N_0-1}(x(i_0), y(i_0))\big] \tag{2.16}$$

where $\rho_n(x(i_0), y(i_0))$ is assumed to be zero outside the interval $[N_{\min}, N_{\max}]$.

Finally, the maximum value of the cross-correlation coefficient, denoted by ρ^*, is obtained by

$$\rho^* = \rho_{\underline{N}}\big(x(i_0), y(i_0 + \beta^*)\big) = \frac{(1 - \beta^*)(x(i_0), y(i_0)) + \beta^*(x(i_0), y(i_0 + 1))}{\Big[|x(i_0)|^2\big((1 - \beta^*)^2 |y(i_0)|^2 + 2\beta^*(1 - \beta^*)(y(i_0), y(i_0 + 1)) + \beta^{*2}|y(i_0 + 1)|^2\big)\Big]^{1/2}}. \tag{2.17}$$

Note that to obtain the exact pitch estimate $\hat{T}_0$ from the integer pitch N_0LT, *only* two extra inner products, namely $(x(i_0), y(i_0 + 1))_{\underline{N}}$ and $(y(i_0), y(i_0 + 1))_{\underline{N}}$, have to be calculated in the evaluation of β^* of (2.13). The other inner products in (2.13), $(x(i_0), y(i_0))_{\underline{N}}$ and $|y_{\underline{N}}(i_0)|^2$, have already been computed for $\rho_{\underline{N}}(x(i_0), y(i_0))$ during the evaluation of the integer pitch, and $|y_{\underline{N}}(i_0 + 1)|^2$ can be easily obtained from $|y_{\underline{N}}(i_0)|^2$. Hence, the improvement of the resolution of the estimated pitch and the elimination of the sampling rate "rounding" errors are obtained with a relatively low computational complexity with respect to estimating the integer pitch.

In some cases β^* may fall outside the interval $[0, 1)$. This may happen when the integer pitch N_0 deviates by one sample from the true value. In such cases the integer pitch is incremented by one when $\beta^* \geq 1$, and decremented by one when $\beta^* < 0$. Then β is reestimated using (2.13). It was shown [12], that if the signal is a pure cosine function, the expression of (2.13) is the correct value of the fractional pitch even for the extrapolation case (i.e., for β outside $[0, 1)$) without reestimating β.

III. Pitch Tracking

Ideally, for a perfectly periodic signal $s(t)$, the function $\rho_\tau(x, y)$ may have several identical maxima at $\tau = kT_0$ within the range $[T_{0_{\min}}, T_{0_{\max}}]$. Practically, due to the nonstationarity of the signal, such as in voice onset, the values of certain maxima at $\tau = kT_0$ for $k > 1$ may be more prominent than the maximum at T_0, resulting in a multiple pitch estimate. A reduced pitch estimate is also possible. This may be due to a dominant first formant. In the latter case, the signal may have a decaying periodicity $T_1 \leq T_0/2$ contained within the true period. Thus, $\rho_\tau(x, y)$ may attain additional local maxima at $\tau = kT_1$ as well, due to the existence of some modulation factor $a(t_0)$ that makes these kT_1-length periods similar in the sense of (2.2). Thus, taking the global maximum of $\rho_\tau(x, y)$ does not necessarily provide the correct value of the pitch. Instead, the pitch should be selected from all pitch candidates, defined as the set of all local maxima of $\rho_\tau(x, y)$ in $[T_{0_{\min}}, T_{0_{\max}}]$ which exceed some specified threshold.

Let τ_m, $m = 1, \cdots, M$ be the set of pitch candidates, sorted in increasing order. All the pitch candidates τ_m have the property that two signal segments of length τ_m spaced τ_m apart are similar up to a multiplicative factor. The true pitch period, however, is characterized by a high correlation coefficient between two segments of *arbitrary* length τ spaced pitch period apart, as long as τ is not too long. Multiple pitch estimates also share this property. However, the spurious maxima of $\rho_\tau(x, y)$ resulting from subperiodicities within the pitch period do not have this property, unless they occur at an exact submultiple of the true pitch and are at constant magnitude. It is therefore natural to choose the correlation interval τ for eliminating such spurious pitch values, to be the largest pitch candidate τ_M. In order to eliminate both types of erroneous pitch candidates, multiple and fractional pitch periods, the candidates τ_m are assessed in increasing order of m. For each candidate τ_m, the correlation coefficient between two segments of length τ_M spaced τ_m apart is evaluated. Namely, the correlation between $s(t)w_{\tau_M}(t - t_0)$ and $s(t + \tau_m)w_{\tau_M}(t - t_0)$. The first candidate for which this correlation coefficient exceeds the specified threshold is selected as the integer pitch estimate.

While inside a voiced segment, the deviation of the pitch from one period to the next one is limited. This deviation is generally within $\pm 10\%$ of the pitch value, and almost never exceeds 25%. Therefore, after the onset transients have settled down (e.g., after 3–5 periods of voiced speech) the search for the integer pitch is consentrated only in the neighborhood of the previous pitch value, instead of the full range of possible values $[T_{0_{\min}}, T_{0_{\max}}]$. This reduction of the search interval, while reducing the computational requirements, also results in a smooth pitch contour, free of spurious pitch values (cf. Fig. 7).

Both the voiced/unvoiced decision and the selection of the pitch estimate are based on a comparison of the correlation values with a threshold $T(t)$. The setting of the threshold value should be made low enough to detect the periodicity in the voiced segments, especially at voice onset, and high enough to avoid misclassification during unvoiced segments due to random high correlation values. It is somewhat difficult to determine a fixed value for the threshold which copes with the variability in the correlation values for different sounds, speakers, and background noise. A better strategy is to adapt the threshold at each time instant to the level of correlation between adjacent pitch periods found for the current speech segment at that instant.

Two limit values, $T_{\text{low}}(t)$ for voiced segments, and T_{high} for unvoiced segments, are defined. $T_{\text{low}}(t)$ is updated at each pitch calculation instant of the algorithm to be the maximum of the following two terms:

$$T_{\text{low}}(t) = \max \{T_{\min}; T_{\max}\} \tag{3.1}$$

where $T_{\min}$ is a fixed value, used as a global lower bound for $T_{\text{low}}(t)$, and $T_{\max}$ is a value proportional to the maximum cross-correlation coefficient that was obtained at the present voiced segment. Note that $T_{\text{low}}(t)$ is at least equal to the global threshold $T_{\min}$, but it is generally increased by the high correlation values in the voiced segment such that it tracks the correlation function in that segment (see Fig. 2).

From our experience with floating point arithmetic and 8 kHz sampling rate we have found the following values to be adequate for good performance: $T_{\min} = 0.80$, $T_{\text{high}} = 0.85$, $T_{\max} = 0.87 \max \rho(t)$, where the maximal value of the correlation is taken for the current voiced segment. It should be noted that apart from the type and accuracy of the computation (e.g., floating or fixed point arithmetic) and the sampling rate, these values are also affected by the low-pass filter used to preprocess the speech signal, and the decimation ratio L (see also Sections IV and V).

An example of the behavior of the adaptive threshold in comparison to the cross-correlation curve is depicted in Fig. 2 for the word *somewhat*. The threshold is denoted by $T(t)$ (dotted line), and the cross-correlation curve by $\rho(t)$ (solid line). The waveform of the word is also shown for the same time scale. In the unvoiced region /s/, the correlation values are low and the threshold is set at $T_{\text{high}}(0.85)$. When entering the voiced region, the correlation values increase and when the threshold T_{high} is exceeded a voiced segment is identified, and a transition to $T_{\text{low}}(t)$ occurs for the threshold. $T_{\text{low}}(t)$ starts from the value of $T_{\min}(0.80)$, but is immediately increased by the high cross-correlation values of the voiced speech to be $0.87 \max \rho(t)$.

The magnitude of $\rho_{T_0}(x, y)$ can be used not only to classify V/UV segments but also to segment voiced speech into distinct voiced sounds. During a V/V transition period, $\rho_{T_0}(x, y)$ is slightly reduced with respect to values obtained during the steady portion of the sound, as can be seen in Fig. 2 for the three successive voiced segments /o/, /m/, and /wha/ of the word *somewhat*. This momentary drop is followed by a fast recovery to high correlation values as soon as the new sound be-

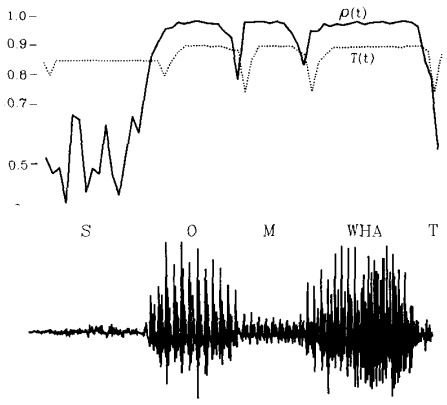


Fig. 2. Behavior of the cross-correlation sequence and the adaptive threshold for the word "somewhat" as a function of time.

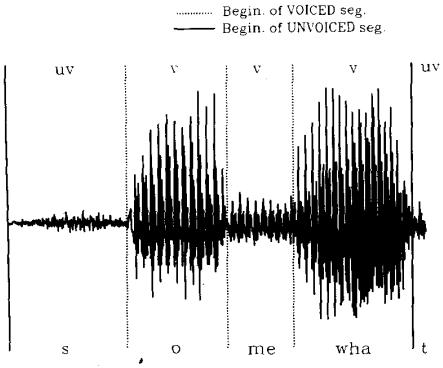


Fig. 3. Segmentation of the word "somewhat." The adaptive threshold is used to automatically segment the word into its phonetic units.

comes stationary. The distinction between successive voiced segments is therefore made by identifying the time instants in which the cross-correlation curve intersects the threshold curve. At each intersection, the threshold is momentarily lowered (to 0.75) to allow smooth segmentation of the voiced portions of the speech. If the transition between the phonemes is indeed V/V, then the correlation will recover to exceed $T_{\text{low}}(t)$, and the threshold will increase too with the increase of the correlation values. Otherwise, the threshold is set to T_{high} and an unvoiced segment is identified as can be seen in the last unvoiced region /t/. The final segmentation of the word *somewhat* into phonetic units using the above technique is demonstrated in Fig. 3.

IV. Computational Analysis

The major computational load of the proposed algorithm is the computation of the integer pitch, whose complexity is $O(N_{\max}^2)$, where $N_{\max}$ corresponds to the lowest candidate fundamental frequency that has to be evaluated. Since $N_{\max}$ is related to the sampling interval T via

$$N_{\max} = \underline{T_{0_{\max}}/LT} \tag{4.1}$$

where the underbar means taking the integer part of the above expression, the computational complexity can be reduced by selecting a decimation factor $L > 1$ (see (2.6) for a formal definition of L). From (4.1) it is evident that, for a given $T_{0_{\max}}$ and a sampling interval T, the complexity is reduced by a factor of L^2. This factor is a result of the decimation of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of (2.7), as well as a reduction by a factor of L of the number of cross-correlation evaluations in the range $[N_{\min}, N_{\max}]$.

While the computational complexity is reduced significantly by such a decimation, the integer pitch N_0 is obtained with a lower resolution, LT seconds instead of L. Nevertheless, the subsequent interpolation technique of Section II-B compensates for the reduced sampling rate and may still provide a high-resolution and accurate estimate of the pitch period interval.

Since, for most PDA's, the sampled signal is low passed before pitch is extracted, decimation is allowed. Depending upon the specific sampling rate and the bandwidth of the low-pass filter, L can be selected in the range $2 \le L \le 8$. However, it might be desirable to obtain the integer pitch estimate with the resolution of the original sampling interval T (instead of LT) and still gain the reduced computational complexity associated with decimation. Fortunately, this can be accomplished with a minor increase in complexity, adding *only* L cross-correlation evaluations.

Rewrite (2.14) as

$$\hat{T}_0 = (\underline{N}_L L + \beta_L^* L)T \tag{4.2}$$

where a subscript L denotes the values obtained with a decimation factor $L > 1$. Based on the definition of β, the quantity $\beta_L L \in [0, L)$ can be expressed as

$$\beta_L L = k + \beta \tag{4.3}$$

where k is an integer in the range $[0, L-1]$, and $0 \le \beta < 1$ is the residual fraction. Substitution of (4.3) in (4.2) yields

$$\hat{T}_0 = (\underline{N}_L L + k^* + \beta^*)T \tag{4.4}$$

as if the pitch is estimated with the original sampling resolution T, where by comparing (4.4) with (2.14), $\underline{N}$ can be expressed as

$$\underline{N} = \underline{N}_L L + k^*. \tag{4.5}$$

From (4.4) and (4.5) it follows that the pitch estimate $\hat{T}_0$ can be computed using decimated vectors in three steps:

1) Evaluation of $\underline{N}_L$ as the integer pitch of the L: 1 decimated signal.

2) Evaluation of k^* from (2.9) as

$$k^* = \underset{k}{\text{argmax}}\, \rho_{\underline{N}_L}(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + k/L))$$
$$k = 0, \cdots, L-1. \tag{4.6}$$

The evaluation of k^* adds only L correlation evaluations while maintaining the resolution of the original sampling interval for the integer pitch. Note that, though decimated, the vectors $\boldsymbol{y}_{\underline{N}_L}(i_0 + k/L)$ are all available using different sets of the original samples. Such vectors are known as the *polyphase* vectors of the signal for decimation L: 1. The construction of those L vectors is illustrated in Fig. 4, for an example of $L = 4$.

3) Computing β^* according to (2.13) as before, with the decimated vectors $\boldsymbol{y}_{\underline{N}_L}(i_0 + k^*/L)$ and $\boldsymbol{y}_{\underline{N}_L}(i_0 + (k^* + 1)/L)$.

The above procedure exploits the fact that the elements of the polyphase vectors have been acquired at the original sampling rate. It enables to determine $\underline{N}$ with full resolution and a reduced complexity of $O(N_{\max}^2/L^2 + \underline{N})$, as compared to $O(N_{\max}^2)$ of

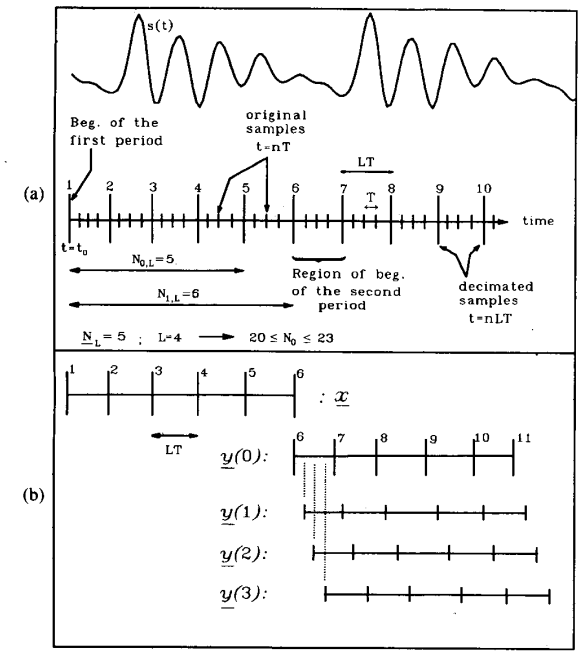


Fig. 4. Construction of the polyphase vectors (example). The L polyphase vectors $y_{\underline{N}_L}(i_0 + k/L)$ for $k = 0, \cdots, L-1$, are shown for a decimation factor of $L = 4$ and an integer pitch of $N_{0,L} = 5$ decimated samples. $s(t)$ at the top of (a) is a low-passed speech segment which corresponds to the time interval illustrated below. First, the integer pitch $N_{0,L}$ is determined as 5 decimated samples that correspond to an integer pitch in the range of $20 \le N_0 \le 23$ original samples. To determine where exactly the second period starts, the decimated vector $x_{\underline{N}_L}(i_0)$ (indicated as $\underline{x}$) is correlated with all the polyphase vectors $y_{\underline{N}_L}(i_0 + k/L)$ (indicated by $\underline{y}(k)$) that are illustrated in (b).

the straightforward approach. This is accomplished by focusing around the candidate decimated integer pitch, using decimated polyphase vectors. From a practical point of view it was found that decimating a low-pass version of the signal to 2–4 kHz (i.e., $L = 2 - 4$ at 8-kHz sampling rate) may suffice to obtain the desired accuracy and resolution level for most applications.

V. Performance Evaluation

The proposed pitch determination algorithm was tested on synthetic signals as well as on real speech data. The speech was prefiltered by a 4-4-5 filter, which is an 11-tap FIR filter comprised of a concatenation of three moving average filters of 4, 4, and 5 tapes, respectively. The filter can thus be implemented without multiplications. It has a smooth cutoff at about $\pi/5$ radians (or 800 Hz in 8-kHz sampling rate). The filtering is used to allow decimation and to remove high frequency components which may reduce the magnitude of the cross-correlation function. A bandwidth of 800 Hz is sufficient for tracking the signal periodicity in a pitch range of 50–600 Hz.

Unfortunately, there is no standard criterion or test for quantifying the performance of pitch determination algorithms. Such a standard test will be proposed here based on the fact that the pitch values represent the time duration between the glottal pulses. In this representation the voice excitation process is modeled as an impulse train in which the impulses appear along the time axis in time instances which correspond to the instances of glottal pulsation. Thus, the impulse train, denoted by $p(t)$, is given as

$$p(t) = \sum_n a(t_n)\delta(t - t_n) \tag{5.1}$$

where t_n are the instants of the glottal pulsation and $a(t_n)$ is the pulse modulation sequence (see (2.2)). The pitch contour is characterized by the sequence p_n, which is the sequence of the time durations between the pulses: $p_n = t_n - t_{n-1}$. If we denote by $h(t)$ a time-invariant impulse response of the vocal tract filter for a given sound, then we can synthesize a speech waveform $s(t)$ of that sound by convolving the excitation $p(t)$ with the impulse response of this filter

$$s(t) = p(t) * h(t) = \sum_n a(t_n)h(t - t_n). \tag{5.2}$$

It should be noted that (5.2) models the time modulation (variations) of the pitch intervals but not the transfer function modulation due to articulation of different sounds.

By sampling the synthesized speech $s(t)$ with a sampling interval T, the sequence $\{s(kT)\}$ is obtained, which is then used as a test sequence for estimating the pitch detector performance. Since the synthetic speech is generated using a reference pitch contour p_n, the pitch estimate $\hat{p}_n$, computed by the PDA under test, can be compared to the known reference. Obviously, such an evaluation is impossible for real speech data. The test can be performed for any desired spectrum of $H(f)$ and $P(f)$. To allow flexibility in choosing the desired vocal tract filter according to the desired sound to be synthesized, $h(t)$ is suggested to be a sum of damped cosine waves

$$h(t) = \sum_{m=1}^{M} a_m g_m(t) \cos(2\pi f_m t) \tag{5.3}$$

where

$$g_m(t) = \begin{cases} e^{-b_m t}; & t \ge 0 \\ 0; & \text{otherwise} \end{cases}$$

and the desired frequency response $H(f)$ can be achieved by properly selecting the parameters a_m, b_m, f_m, for $m = 1, \cdots, M$ in

$$H(f) = \tfrac{1}{2} \sum_{m=1}^{M} a_m [G_m(f - f_m) + G_m(f + f_m)] \tag{5.4}$$

where

$$G_m(f) = \frac{1}{b_m + j2\pi f}.$$

Hence, $H(f)$, the frequency response of the vocal tract filter, is composed of resonators characterized by the modulating frequencies f_m, the bandwidths b_m and the magnitudes a_m.

The voice excitation process $p(t)$ can be modeled based on real speech statistics. The pitch values p_n should be chosen such that *any* real value within some interval $[p_{\min}, p_{\max}]$ will be feasible. That is, if the algorithm from which the pitch values are determined can supply pitch values with only a fixed resolution (for example, only integer number of samples), then these values should be varied within that resolution to allow any real value. Furthermore, if the pitch process is synthesized, then the bandwidth of the pitch variation process should be much narrower than the average pitch frequency, as is the case in real speech.

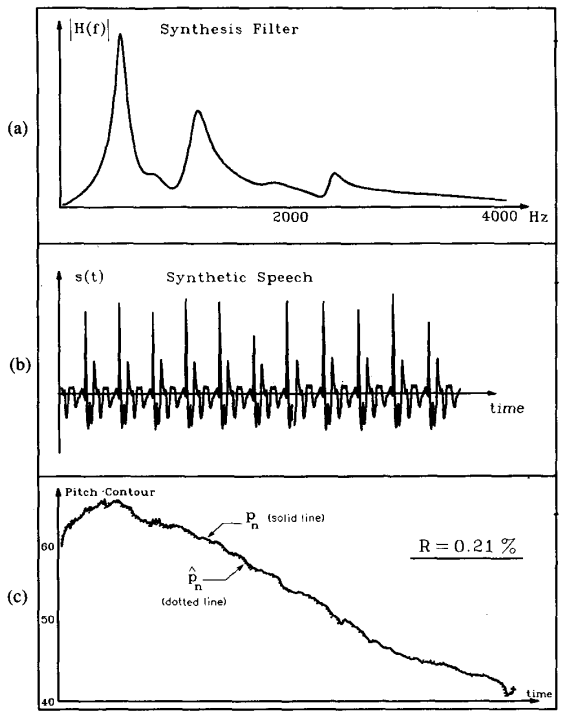


Fig. 5. A test example of the proposed pitch detector using the impulse train model. (a) $H(f)$ is the synthesis filter representing the vocal tract transfer function for the vowel /a/. (b) Segment of the synthesized speech $s(t)$. (c) Comparison of p_n (the input pitch contour) and $\hat{p}_n$ (the pitch detector output). The low value of R indicates the high accuracy of the pitch determination algorithm in determining the *exact* pitch periods while using a finite (8-kHz) sampling rate.

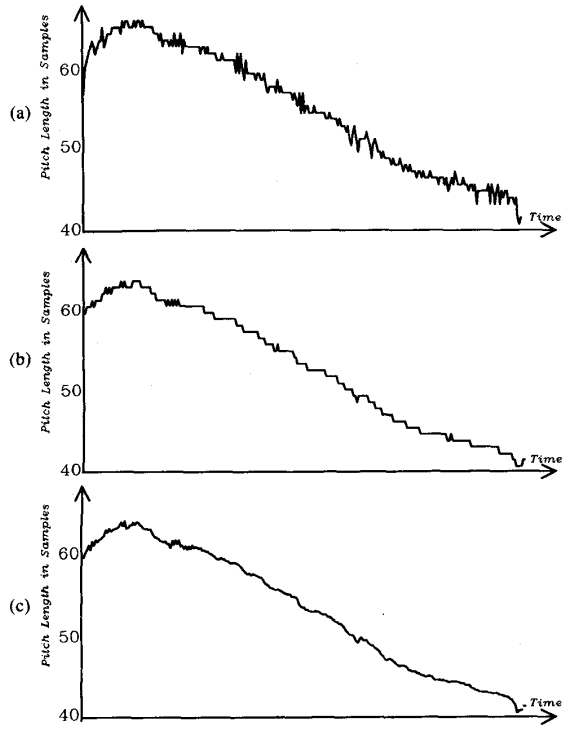


Fig. 6. High resolution pitch contour determination. Comparison of three analysis schemes for the vowel /a/ uttered with descending pitch: (a) The distance between main peaks as a rough estimate of the contour which corresponds to a visually marked pitch. (b) Integer pitch contour N_0 at 8-kHz sampling rate with its associated time quantization noise. (c) High-resolution pitch contour N by the proposed algorithm, having a smooth envelope as expected from a human voice source.

The proposed pitch detector was tested using this approach. Note that such a test examines the ability of a pitch determination algorithm to find the *exact* values of the pitch from a sampled speech, i.e., not restricted to any finite resolution.

The quality criterion R of the algorithm is the average relative deviation of the estimated pitch value from the input one. That is, the average (relative) error in estimating the pitch values is $\pm R$ where

$$R = \frac{\sum_n |p_n - \hat{p}_n|}{\sum_n p_n}. \tag{5.5}$$

Demonstrated in Fig. 5 are the results for the pitch contour p_n of Fig. 6 with a synthesis filter $H(f)$ which corresponds to the transfer function of the vocal tract for the vowel /a/. The parameters used to synthesize the speech were:

$$\{f_m\} = 520\ 800\ 1190\ 1840\ 2390$$

$$\{a_m\} = 250\ 60\ 215\ 25\ 40$$

$$\{b_m\} = 320\ 720\ 520\ 770\ 350$$

for which the formants are obtained in 520, 1190, 2390 Hz (which correspond to the vowel /a/ [15]). Note also that p_n could take *any* value between 40 to 70 samples (and consequently $\hat{p}_n$ too).

In Fig. 5(c), the comparison between the input pitch contour to the synthesis filter p_n and the pitch detector's output $\hat{p}_n$ are compared. The average relative deviation was obtained as $R = 0.21\%$. Such a low error in the pitch value is below human perception. This example demonstrates the capability of the proposed pitch detector to accurately determine the pitch, while analyzing a speech signal sampled at a finite (8-kHz) sampling rate.

The real speech material of our experiments covered a variety of speakers and a full range of pitch frequencies. However, the results cannot be compared objectively since the true pitch value is unknown. Nevertheless, the correlation values inside *voiced* segments were around 0.97–0.99 for the integer pitch values and were improved to about 0.99–1 after interpolation. This observation indicates a detection of an almost perfect periodicity for adjacent periods.

The super resolution feature of the new PDA is illustrated for a long phonation of the vowel /a/ uttered with descending pitch intervals, in Fig. 6. The high-resolution pitch estimate is compared to the a rough estimate of the pitch contour obtained by a peak analysis of the signal. The lack of robustness of such a procedure is obvious. Then, the integer analysis in 8-kHz sampling rate resolution is shown. Although it performs better than peak analysis, the quantization noise is still evident. Only the interpolated pitch estimate has a smooth estimate, free of the quantization noise.

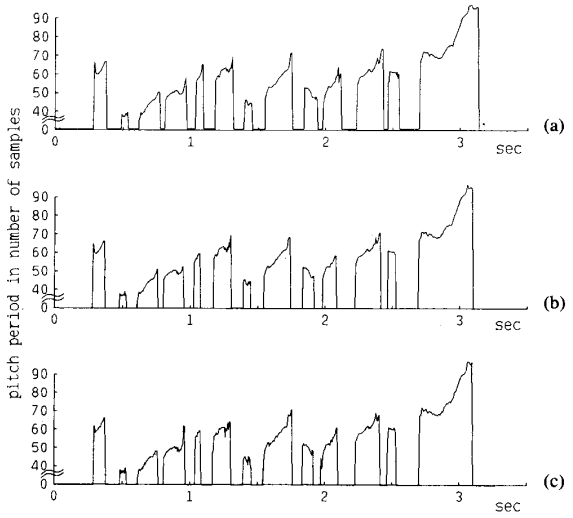


Fig. 7. Pitch contour of fluent speech. The utterance was, "in speech communication systems, the speech signal is transmitted, stored, and processed in many ways." It was uttered by a male speaker and sampled at 8 kHz. (a) Without noise, (b) SNR = 10 dB, (c) SNR = 3 dB.

Pitch contours of long sentences of fluent speech are demonstrated in Fig. 7. A white Gaussian noise was added to the speech signals in two levels of 10 and 3 dB SNR. It is evident that the algorithm performs well even in such a noisy environment. Still, no multiple and half-pitch values were found, and the only effect of the noise was a slight bias in the estimated pitch.

In an independent study, the proposed PDA was evaluated for a potential use in a system that gives a visual feedback of voice parameters (including pitch) in real-time for deaf children [12]. The evaluation results of the PDA for 32254 pitch values extracted from human voice were as follows: V/UV errors—0.16%, UV/V errors—0.36%, half pitch—0.01%, double pitch—0.01%. In addition, in order to validate the accuracy of the algorithm against *a priori* known values, synthetic speech was used. The standard deviation of the estimation error was 0.007 ms at a sampling rate of 9.6 kHz. This is about 0.1% error for an average F_0 of 140 Hz, which is below F_0 difference limen of human pitch perception [14].

VI. Summary

An accurate, robust, and reliable pitch determination scheme was outlined. The algorithm extracts the pitch with a very high resolution (i.e., as a real number) despite the finite (8-kHz) resolution of the sampled speech sequence. This made it possible to successfully analyze the pitch variation process in disarthric patients for medical aims [13] and to develop an accurate pitch synchronous spectral analysis scheme [10], [11]. The computational complexity of the proposed algorithm is well within the capacity of modern DSP technology and therefore can be implemented in real time.

Appendix A

In this Appendix we prove that (2.13) is the solution to the optimization problem of (2.12). Hereafter, the subscript $\underline{N}$ of the quantities involved will be omitted, for simplicity.

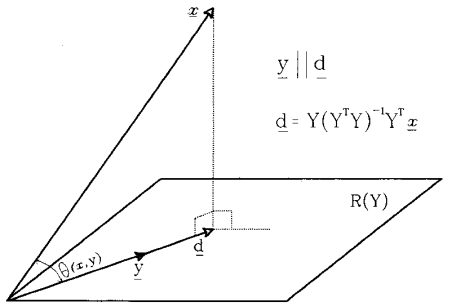


Fig. 8. Projection operation. The vector $\boldsymbol{d}$ is the projection of $\boldsymbol{x}(i_0)$ (marked by $\underline{x}$) into the subspace $\boldsymbol{R}(\boldsymbol{Y})$, spanned by the columns of $\boldsymbol{Y}$. The vector $\boldsymbol{y}(i_0 + \beta)$ (marked by $\underline{y}$) is chosen parallel to $\boldsymbol{d}$ and thus $\theta(\boldsymbol{x}(i_0), \boldsymbol{d}) = \theta(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + \beta))$.

Denote by $\boldsymbol{Y}$ the matrix whose columns are the two vectors $\boldsymbol{y}(i_0)$ and $\boldsymbol{y}(i_0 + 1)$, and by $\boldsymbol{R}(\boldsymbol{Y})$ the subspace spanned by the columns of $\boldsymbol{Y}$. Since $\rho(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + \beta))$ is the cosine of the angle $\theta(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + \beta))$ between the two vectors $\boldsymbol{x}(i_0)$ and $\boldsymbol{y}(i_0 + \beta)$, then, maximizing $\rho(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + \beta))$ is equivalent to minimizing $\theta(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + \beta))$. But since, according to (2.11), $\boldsymbol{y}(i_0 + \beta)$ belongs to the subspace $\boldsymbol{R}(\boldsymbol{Y})$, we first choose the minimal angle between $\boldsymbol{x}(i_0)$ and that space. This is done by projecting $\boldsymbol{x}(i_0)$ into $\boldsymbol{R}(\boldsymbol{Y})$ (yielding the projected vector $\boldsymbol{d}$), and then choosing $\boldsymbol{y}(i_0 + \beta)$ inside the space parallel to $\boldsymbol{d}$, preserving the minimal angle. The above projection is illustrated in Fig. 8. The angle between $\boldsymbol{x}(i_0)$ and $\boldsymbol{d}$ is the minimal value of $\theta(\boldsymbol{x}(i_0), \boldsymbol{y}(i_0 + \beta))$.

From the orthogonal projection theorem it follows that

$$\boldsymbol{d} = \boldsymbol{Yb} = b_1\boldsymbol{y}(i_0) + b_2\boldsymbol{y}(i_0 + 1) \tag{A.1}$$

where $\boldsymbol{b}$ is resolved by

$$\boldsymbol{b} = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\boldsymbol{x}(i_0). \tag{A.2}$$

Now it is desired to choose β so that $\boldsymbol{y}(i_0 + \beta)$ is parallel to $\boldsymbol{d}$. From (2.11) and (A.1) it follows that

$$\frac{b_1}{1 - \beta} = \frac{b_2}{\beta}. \tag{A.3}$$

Substitution of (A.3) into (A.2) and rearrangement yields the final result given in (2.13).

Acknowledgment

The authors wish to acknowledge the anonymous reviewers for their careful examination of the manuscript. Their insight and comments led to a better presentation of the ideas expressed in this paper.

References

[1] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, Aug. 1969.
[2] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 2–8, Feb. 1976.
[3] L. R. Rabiner, "On the use of autocorrelation analysis for pitch determination," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 22–33, Feb. 1977.
[4] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262–266, June 1968.

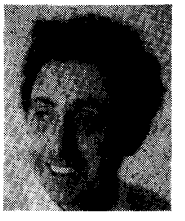
[5] M. J. Ross *et al.*, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353–362, Oct. 1974.
[6] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.
[7] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.
[8] W. Hess, *Pitch Determination of Speech Signals.* New York: Springer, 1983.
[9] W. Hess and H. Indefrey, "Accurate pitch determination of speech signals by means of a laryngograph," presented at the ICASSP-84, Apr. 1984.
[10] Y. Medan and E. Yair, "Discrete spectral analysis of periodic time functions," in *Proc. ICASSP-87*, Apr. 1987, pp. 1797–1800.
[11] Y. Medan and E. Yair, "Pitch synchronous spectral analysis scheme for voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 9, pp. 1321–1328, Sept. 1989.
[12] H. Crepy and G. Rouquie, "YAPP: A fast and precise pitch extraction algorithm," IBM France Paris Scientific Center Rep., June 1986.
[13] E. Yair, "Voiced based quantitative evaluation of Parkinsonism," Ph.D. dissertation, Dep. Elec. Eng., Technion-Israel Institute of Technology, May 1987.
[14] J. L. Flanagan and M. G. Saslow, "Pitch discrimination for synthetic vowels," *J. Acoust. Soc. Amer.*, vol. 30, pp. 435–442.
[15] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

Yoav Medan was born on November 11, 1951 in Haifa, Israel. He received the B.Sc. (*cum laude*) and Ph.D. degrees from the Technion-Israel Institute of Technology, in 1973 and 1983, respectively.

From 1973 to 1978 he was with the Israeli Air Force, working on the design and simulation of digital flight control systems. Since 1983 he has been with the IBM Israel Scientific Center, working on digital processing of speech signals, vector algorithms for the IBM 3090VF, and real-time multitasking operating systems for digital signal processors. Currently he is the Manager of the computer engineering and signal processing research group.

Eyal Yair was born in Haifa, Israel, on February 7, 1956. He received the B.Sc. degree in 1982, and the Ph.D. degree in 1987, both from the Department of Electrical Engineering at the Technion-Israel Institute of Technology.

In 1983 he joined the IBM Israel Scientific Center in Haifa, where he was involved in various projects in speech processing. From 1987 to 1989 he was a visitor in the Department of Electrical and Computer Engineering at the University of California, Santa Barbara, where he was involved in neural network and pattern classification research. Currently, he is engaged in handwritten character recognition research. His areas of interest are pattern recognition, data compression, signal processing, and neural networks.



Dan Chazan was born in Tel-Aviv, Israel, on November 11, 1939. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1961, 1963, and 1965, respectively.

From 1965 to 1969 he was a staff member of the IBM Research Center at Yorktown Heights, NY, where he worked on various problems involving parallel computation, operations research, and numerical analysis. From 1969 to 1972 he worked for the Israel Defense Ministry Research Division. In 1973 he rejoined IBM, this time at the IBM Israel Scientific Center, where his research interests involved a range of fields from ground water modeling to econometrics. Since 1982 his main interest has been in the area of voice signal processing for a variety of applications.