Design and Implementation of a Robust Pitch Detector Based on a Parallel Processing Technique

RAFID A. SUKKAR, JOSEPH L. LoCICERO, MEMBER, IEEE, AND JOSEPH W. PICONE, MEMBER, IEEE

Abstract—**The design and implementation of a parallel processing based pitch detector is presented. Pitch information is extracted by performing pitch detection on four different waveforms derived from the speech signal. A "pitch voter" is then used to combine pitch information from the four pitch detection processes to determine a final pitch estimate. The performance of this pitch detector is evaluated on a large database and compared to other well-known pitch detection algorithms. This parallel processing pitch detector was implemented in real time on a TMS32020 fixed-point digital signal processor as part of a 2.4 kbit/s vocoder. A performance comparison of the real-time fixed-point implementation and a computer simulation are also given. Our results show that the pitch detector performance is maintained in the real-time implementation. This can mainly be attributed to the fact that the majority of the algorithm computations are integer arithmetic and logic-type operations.**

I. INTRODUCTION

WHILE many pitch detection methods have been proposed [1]–[12], accurate and consistent pitch detection remains a very difficult problem. Anomalous or aperiodic behavior of the vocal tract vibrations of the speaker is the main reason why many pitch detectors fail on some occasions in determining the correct pitch. This kind of anomalous behavior often occurs at the transition regions between voiced and unvoiced segments of speech where pitch periods are irregular or pitch pulses vary considerably in amplitude. One might even argue that the vocoder model of the excitation signal as being either periodic or random noise is inadequate and a more realistic model is needed.

In this paper, a pitch detector based on a parallel processing approach is presented. Four individual pitch detectors are used to extract pitch information from four different signals. The outputs of the four pitch detectors are then combined using a pitch voter to determine a final pitch estimate. Performance analysis, on a large database that has been carefully cross balanced for the sex and age of the speaker, indicates reliable and accurate pitch detection. In addition, there is no performance degradation when the pitch detector is implemented in real time on a fixed-point processor. This can be attributed to the fact that the majority of the algorithm computations are integer arithmetic and logic-type operations.

The organization of this paper is as follows. In the next section, a detailed description of the pitch detector algorithm is given. In Section III, the pitch detector parameter design is presented. Performance analysis and experimental results are given in Section IV. Finally, in Section V, we give a general overview of a real-time implementation of the algorithm using a single TMS32020 digital signal processor.

II. THE PARALLEL PROCESSING PITCH DETECTOR ALGORITHM

An overall block diagram of the pitch detector is given in Fig. 1, where low-pass filtered speech signal is sampled at 8 kHz and quantized using a 16 bit linear quantizer. The digitized speech $x(n)$ is processed as 20 ms frames, and a tenth-order linear predictive coding (LPC) analyzer is used to generate the LPC error signal $e(n)$. Since pitch information is present in both the original speech signal and the LPC error signal, pitch detection is performed separately on each one of these signals. Typically, periodicity that appears in the original speech waveform also occurs in the LPC residual, implying some redundancy during definitive voiced frames. However, several cases demonstrate the need for pitch detection with both types of waveforms.

An example is shown in Fig. 2 where a particular formant structure causes the periodicity in the voiced speech waveform to be obscured and hard to detect [13]. Since the LPC residual represents the speech waveform with the formant structure removed, voiced speech periodicity can be detected with little difficulty in the residual. A different case arises when the residual waveform fails to show clear periodicity in voiced frames, as shown in Fig. 3 [13]. This occurs when the fundamental frequency of the excitation falls under a voiced speech formant. As a result, the excitation information normally found in the residual

Manuscript received May 31, 1987; revised October 19, 1987. This work was supported by the Advanced Technology and Data Switching Laboratory, AT&T Bell Laboratories, Naperville, IL. A portion of this paper was presented at the National Communication Forum, Chicago, IL, September 1986. A portion of this work was submitted by R. A. Sukkar in partial fulfillment of the requirements for the M.S. degree in the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616.

R. A. Sukkar and J. L. LoCicero are with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616.

J. W. Picone is with Texas Instruments, Inc., Dallas, TX 75266.

IEEE Log Number 8718534.

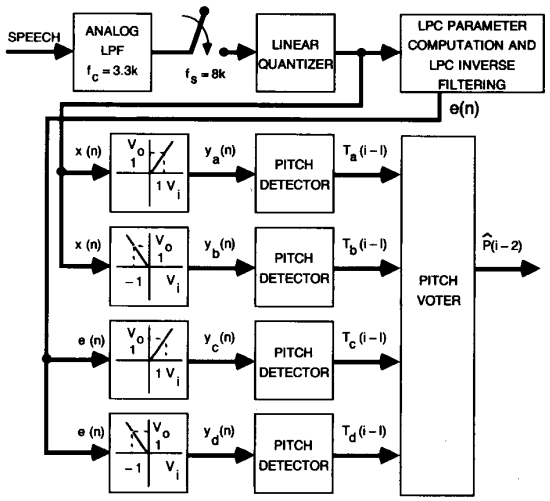


Fig. 1. Block diagram of the parallel processing pitch detector (PPPD).

Fig. 2. An example where the pitch period is easily found in the residual, but not in the speech waveform. The upper trace is the input speech and the lower trace is the LPC residual waveform.

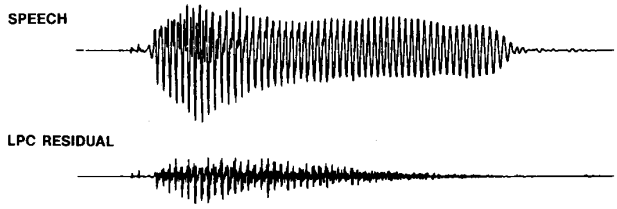


Fig. 3. An example where the pitch period is easily found in the speech waveform, but not in the residual. The upper trace is the input speech and the lower trace is the LPC residual waveform.

is removed by LPC inverse filtering, causing the residual to look noisy while the original speech waveform appears to be clearly periodic.

Before performing pitch detection on $x(n)$ and $e(n)$, each signal is split into its positive-going portion and its negative-going portion. The reason for clipping the waveform in this fashion is that the composite waveform might not show clear periodicity during a voiced frame while one of the clipped waveforms exhibits periodicity that can be easily detected. Let us denote the positive-going and the negative-going speech waveforms by $y_a(n)$ and $y_b(n)$, respectively. Further, let us denote the positive-going and the negative-going residual waveforms by $y_c(n)$ and $y_d(n)$, respectively.

Since the speech is processed as frames of 20 ms in duration $y_a(n)$, $y_b(n)$, $y_c(n)$, and $y_d(n)$ consist of 160 sample points each. Four pitch detectors operate, in parallel, on $y_a(n)$, $y_b(n)$, $y_c(n)$, and $y_d(n)$. These individual pitch detectors have identical structure, and differ only in the values of the control parameters (described in Sections III and IV). The pitch voter combines the four pitch distance estimates $T_a(i-1)$, $T_b(i-1)$, $T_c(i-1)$, and $T_d(i-1)$ to produce a final pitch distance estimate $\hat{P}(i-2)$ where two frames of delay are needed to ensure smooth pitch tracking.

The Individual Pitch Detector

The operation of the individual pitch detector starts by identifying a set of samples (or pulses) over a frame on which the periodicity check is to be performed. To define this set of pulses, the pitch detector first finds the global maximum amplitude M_0 and its location D_0 in the frame. Any pulse selected from this point on must satisfy three conditions.

First, the next pulse selected must be a local maxima, excluding all pulses that have already been picked or eliminated. This condition is applied because pitch pulses usually have higher amplitudes than other samples in a frame. Let us denote the local maxima by M_j.

Second, any pulse satisfying the first condition should have an amplitude greater than or equal to a certain percentage of the global maximum. That is,

$$M_j \geq gM_0, \tag{1}$$

where g is a threshold amplitude percentage and will be discussed in detail in the next section.

Third, any pulse satisfying both of the above conditions must be separated by at least 2.25 ms (18 sample periods) from all the pulses that have already been located. This condition is included because we assume that the largest pitch frequency usually encountered in human speech is 400 Hz, corresponding to a pitch period of 2.55 ms (20 sample periods), and allow a 10 percent tolerance.

The pitch detector can identify at most nine pulses over one frame. These pulses are called candidate pulses. The amplitudes $\{M_j\}$ and locations $\{D_j\}$ of these candidate pulses are used to define a distance that represents the smallest distance for which a subset of candidate pulses is periodic. This distance is determined recursively by considering the distance from the frame global maximum M_0 to the closest adjacent candidate pulse. This distance is called a candidate distance d_c and is given by

$$d_c = |D_0 - D_j|. \tag{2}$$

If such a subset of maxima in the frame is not separated by this distance plus or minus a breathing threshold B (to be discussed in the next section), then this candidate distance is discarded, and the process begins again with the next closest adjacent candidate pulse.

If a subset of candidate pulses separated by d_c plus or minus B is found, then the amplitude of the candidate pulse that is adjacent to M_0 must also pass an interpolation amplitude test to ensure a smooth amplitude transition. A smooth amplitude transition is desirable since the enve-

lope of most voiced speech segments exhibits no sudden jumps.

This amplitude test performs linear interpolation between M_0 and each of the other candidate pulses M_j, $j > 0$, and requires that the amplitude of the candidate pulses in between M_0 and M_j be at least q percent of these interpolated values. The interpolation amplitude threshold q percent will also be discussed in the next section. To clarify by example, consider the candidate pulses shown in Fig. 4. For d_c to be a valid candidate distance,

$$M_1 > q\left[M_2 + \frac{M_0 - M_2}{|D_0 - D_2|}\,|D_1 - D_2|\right], \quad (3)$$

$$M_3 > q\left[M_4 + \frac{M_0 - M_4}{|D_0 - D_4|}\,|D_3 - D_4|\right], \quad (4)$$

$$M_3 > q\left[M_5 + \frac{M_0 - M_5}{|D_0 - D_5|}\,|D_3 - D_5|\right], \quad (5)$$

and

$$M_4 > q\left[M_5 + \frac{M_0 - M_5}{|D_0 - D_5|}\,|D_4 - D_5|\right], \quad (6)$$

where $d_c = |D_0 - D_1| > 2.25$ ms. As noted previously, we must also guarantee that

$$M_j > gM_0, \quad \text{for } j = 1, 2, 3, 4, 5. \quad (7)$$

If any of the above relations is not met, then the subset of candidates is discarded and the process of defining a new d_c begins again with the next adjacent candidate pulse. If a valid d_c is found, a pitch distance is computed as the average distance between adjacent pulses in the set of periodic candidate pulses (i.e., the set of candidate pulses corresponding to the valid d_c). This pitch distance is denoted as $T(i)$, and is then compared to $T(i-1)$ via a pitch consistency test. If no valid d_c is found, $T(i)$ is set to zero, indicating an unvoiced frame.

The pitch consistency test, as the name indicates, ensures pitch consistency over two adjacent voiced frames. It also checks and corrects pitch doubling errors made when defining a candidate distance.

This test starts by checking if $T(i)$ and $T(i-1)$ are close to within a pitch threshold A (to be discussed in the next section). Consequently, if

$$|T(i-1) - T(i)| \le A, \quad (8)$$

then $T(i)$ is a good estimate of the pitch distance and need not be modified. However, if

$$|T(i-1) - T(i)| > A, \quad (9)$$

then we must perform the pitch doubling test.

The pitch doubling test checks if $T(i-1)$ and twice $T(i)$ are close to within the pitch threshold A. Therefore, if

$$|T(i-1) - 2T(i)| \le A, \quad (10)$$

then we set $T(i)$ to be equal to $T(i-1)$. This, in effect, corrects any pitch doubling error that might occur when defining a candidate distance. However, if both of the above tests fail, that is, if

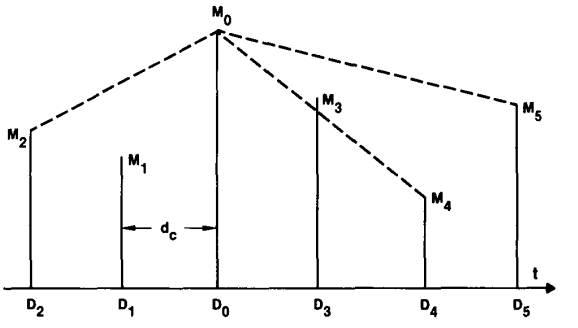


Fig. 4. An example of a subset of candidate pulses when the amplitude interpolation test is applied.

$$|T(i-1) - T(i)| > A \quad (11)$$

and

$$|T(i-1) - 2T(i)| > A, \quad (12)$$

then we set $T(i)$ to zero, implying that frame i cannot be voiced since its initial estimated pitch distance is not consistent with the pitch distance estimate for frame $i-1$.

Since human speech usually does not produce more than one transition between voiced and unvoiced segments within 40 ms, the individual pitch detector checks and eliminates all voiced–unvoiced–voiced (VUV) and unvoiced–voiced–unvoiced (UVU) sequences. In the former case, $T(i-1)$ is set to the arithmetic average of $T(i)$ and $T(i-2)$. In the latter case, the pitch distance estimate is set to zero.

The Pitch Voter

A block diagram of the pitch voter is shown in Fig. 5 and a complete description is found in [13]. In this paper, a brief overview is given. The pitch voter consists of two main building blocks: the discriminant analyzer to determine voicing, and the final pitch value estimator to determine a final pitch value. The voicing classification is obtained using a discriminant analysis approach where a weighted sum of several parameters is computed. These parameters are chosen based on their ability to discriminate between voiced and unvoiced frames. They include three parameters representing the number of nonzero pitch estimates determined by the four individual detectors for the present and the two adjacent frames. They also include the first four reflection coefficients, the log of the speech power, and the log of the LPC gain defined as the speech power divided by the residual power. The optimum weights are determined using a training set of speech where the correct voicing is known.

The voiced/unvoiced classification, along with the pitch estimates from the four individual detectors, are used by the final pitch value estimator. If the discriminant analysis determines the frame is unvoiced, $\hat{P}(i-2)$ is set to zero. Otherwise, the frame is voiced and the final pitch estimate is set to the median value of 13 pitch estimates. These 13

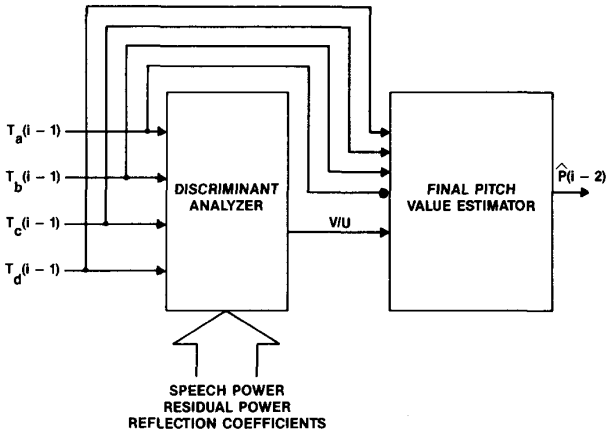


Fig. 5. A functional block diagram of the pitch voter.

pitch estimates come from the four individual detectors for the present and both adjacent frames and the final pitch estimate for the most recent voiced frame.

III. Parallel Processing Pitch Detector Parameter Design

The Amplitude Test Threshold "g"

The purpose of this threshold is to prevent any nonpitch pulses from taking part in the periodicity check performed by the distance detector. In most voiced frames, the amplitudes of the pitch pulses show a small variance. (This fact will be verified shortly when we examine a pitch pulse amplitude histogram.) Thus, it would be reasonable to constrain any possible pitch pulse to have an amplitude greater than or equal to a certain percentage of the maximum pitch pulse amplitude in a frame. We have used the amplitude threshold test

$$M_j \geq gM_0, \tag{13}$$

where g is the amplitude threshold percentage, M_j is the amplitude of any possible pitch pulse chosen by the pitch detector, and M_0 is the maximum pulse amplitude in the frame.

To complete this part of the pitch detector design, we specify an upper and lower bound for a region of acceptable values of g. Statistics have been collected for a normalized amplitude variable g' defined as

$$g' = \frac{M_j^v}{M_0}, \qquad j \neq 0, \tag{14}$$

where M_j^v is a valid candidate pulse amplitude in a voiced frame whose maximum amplitude candidate pulse is M_0. Statistics were collected from 60 sentences spoken by six different speakers, three males and three females. Each sentence is about 2–3 s in duration. The statistics only include data belonging to frames with a nonzero final pitch value where more than two individual pitch detectors determined a nonzero pitch estimate for that frame. In other words, the statistics include data that have a high probability of belonging to voiced frames.

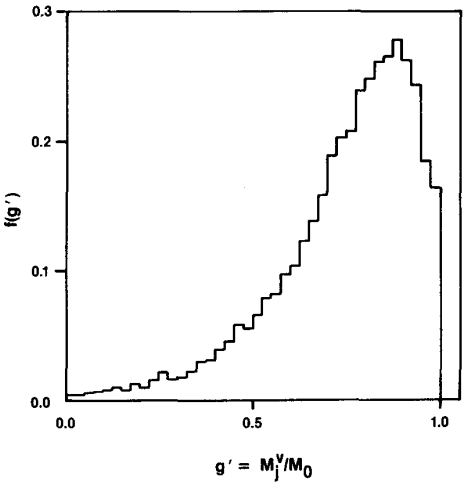


Fig. 6. Probability density function, in histogram form, of the normalized amplitude variable g'.

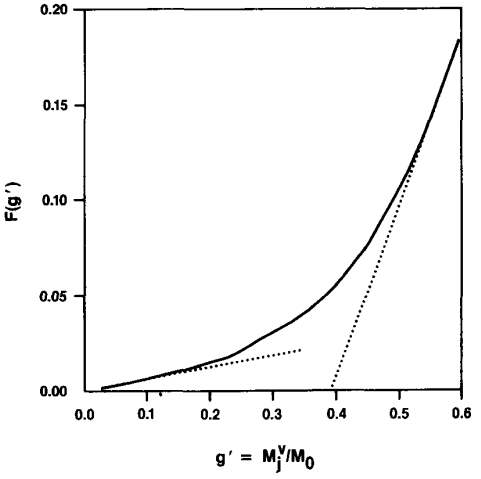


Fig. 7. Cumulative probability distribution function of the normalized amplitude variable g'.

The probability density function (in histogram form) and the resulting cumulative distribution function for g' are shown in Figs. 6 and 7, respectively, where the latter curve is plotted only for $0 \leq g' \leq 0.6$. Fig. 6 clearly shows that most of the pitch pulses are close in amplitude to M_0. For example, 69 percent of all the pitch pulses are at least 70 percent of M_0. However, there are some pitch pulse amplitudes that are not close to M_0. These pulses are either true pitch pulses, for example, at the end of voiced words or they are nonpitch pulses, incorrectly included by the individual pitch detector in the set of valid candidate pulses.

Before drawing any further conclusions, we note that the shape of the plot in Fig. 6 is similar to a translated Rayleigh probability density function curve. A good fit for the probability density plot of Fig. 6 is the Rayleigh-like probability density function shown in Fig. 8. The

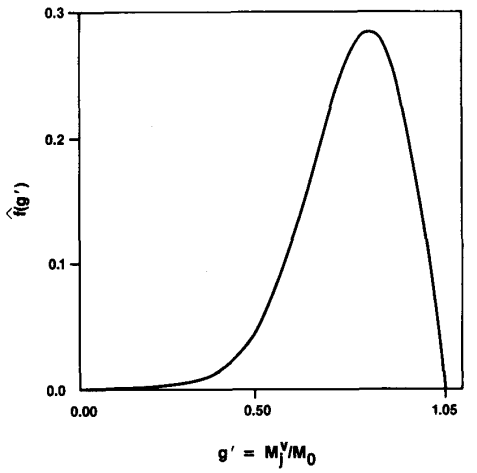


Fig. 8. A Rayleigh-like probability density function approximating the histogram of the normalized amplitude variable g'.

equation of this curve is

$$\hat{f}(g') = \frac{(1.05 - g')}{\alpha^2} \exp\left\{-\left[(1.05 - g')^2/2\alpha^2\right]\right\} \cdot u(1.05 - g'), \quad (15)$$

where $u(x)$ is the unit step function and α is directly proportional to the standard deviation of g'. It was found that $\alpha = 0.2125$ for this speech database. The expected value of the above probability density function [14], [15] was found to be 0.72 and the standard deviation [14], [15] computed to be 0.14.

If we set an upper bound on the amplitude threshold g to be three standard deviations away from the expected value of g', then $g < 0.3$. The Rayleigh-like probability density function curve then yields $P[g' < 0.30] = 0.002$. This probability includes a few valid pitch pulses, but it accounts mostly for nonpitch pulses that passed the periodicity check.

A lower bound should also be determined because if g is set too low, nonpitch pulses might pass the amplitude test, resulting in nonreliable operation of the pitch detector. If we set a lower bound on g to be four standard deviations away from the expected value of g', then $g > 0.16$. The Rayleigh-like probability density function curve then yields $P[g < 0.16] = 0.001$ that accounts for nonpitch pulses that passed the periodicity check.

Considering the cumulative distribution function of g' shown in Fig. 7, we can observe three distinct regions defined by two linear asymptotes (the dotted lines in this figure). One region is $0.525 \leq g' \leq 1.0$ where the curve rises sharply, indicating that candidate pulses with an amplitude greater than $0.525M_0$ are much more likely to be pitch pulses than nonpitch pulses. Another region is $0 \leq g' \leq 0.175$ where we would expect just a few pulses with amplitudes less than $0.175M_0$ to be true pitch pulses. The last region is $0.175 < g' < 0.525$ which represents a transition region and indicates a range of values that g' can possibly take. This experimental result is consistent with the upper and lower bounds given above, implying that g should be set between 0.175 and 0.525 to ensure a reliable pitch detector operation. The performance analysis given in the next section supports these results and gives an optimal value for g.

The Amplitude Interpolation Threshold "q"

A range of values for the amplitude interpolation threshold q can be obtained using a statistical approach also. The derivation has been omitted, but a complete discussion of the design technique for q is given in [14]. A good range for the amplitude interpolation threshold was found to be $0.72 < q < 0.78$.

The Breathing Threshold "B"

It was noted in the previous section that if the pitch detector determines a nonzero value for the pitch distance, then there exists a subset of candidate pulses that are separated by a distance d_c plus or minus a breathing threshold B. Intuitively, we can say that if a voiced frame has a large pitch distance, we would expect a larger breathing allowance than in the case when a voiced frame has a small pitch distance. That is, the breathing allowance should be directly proportional to the pitch distance.

Statistical results for the breathing parameter B are given in Table I. The values in the fourth column of Table I are plotted as a function of the pitch distance estimate in Fig. 9. The equation of the dashed line in Fig. 9, representing a least squares linear fit to $E(B) + 4\sigma_B$ data points, is

$$b(T) = 0.345 + 0.084T, \quad (16)$$

where T is the pitch distance. Note that if (16) is used to compute the breathing threshold B, then the frame's pitch distance is required. Obviously, this is an unrealistic requirement. Recall, however, that the breathing threshold B is used to test if a set of candidate pulses is periodic with the period equal to d_c from (2). Therefore, d_c can be used instead of the frame's final pitch distance in (16) to compute B. Experimental results and performance analysis show that the breathing threshold calculated using (16) results in good pitch detector performance. The only shortcoming arises when T is very large, resulting in candidate pulses of unvoiced frames being declared periodic, thereby making an unvoiced-to-voiced error. To remedy this, we clip (16) and compute the breathing threshold via

$$B = b(T), \quad T < 10.77,$$
$$= 1.25, \quad T \geq 10.77. \quad (17)$$

The Pitch Threshold "A"

The pitch threshold A is used to check if two pitch distance estimates belonging to two adjacent frames are close to one other. If the absolute value of the difference between the two pitch distance estimates is less than A, it is decided that these two estimates are close to one other. Otherwise, it is decided that these two estimates are not

TABLE I
STATISTICAL RESULTS FOR THE BREATHING PARAMETER B (ALL PARAMETERS IN UNITS OF MS)

Pitch distance, T	E(B)	σ_B	E(B) + $4\sigma_B$
$2.50 < T \leq 3.75$	0.1000	0.1319	0.6276
$3.75 < T \leq 5.00$	0.1192	0.1489	0.7148
$5.00 < T \leq 6.25$	0.1390	0.1585	0.7730
$6.25 < T \leq 7.50$	0.1784	0.1889	0.9340
$7.50 < T \leq 8.75$	0.2072	0.2060	1.0312
$8.75 < T \leq 10.0$	0.2450	0.2241	1.1414

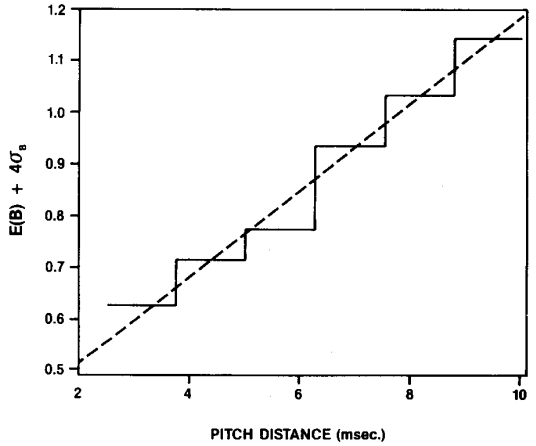


Fig. 9. Experimental values and a linear approximation of the statistics of the breathing parameter B, that is, $E(B) + 4\sigma_B$ plotted versus the pitch distance as measured in ms.

consistent and one of them has to be modified. Our task here is to determine a reasonable value for A.

In the previous section, it was found that the maximum allowable pitch pulse breathing in a voiced frame is 1.25 ms. This implies that if the pitch distance estimate changed by more than 1.25 ms, over two adjacent voiced frames, the distance detector must have made an error in determining one or both of the pitch distance estimates. Obviously, the logical value for the pitch threshold A is 1.25 ms. This indicates that the pitch distance estimates for two adjacent frames are comparable only if the absolute value of the difference between the two estimates is less than or equal to 1.25 ms.

IV. PERFORMANCE ANALYSIS

The parallel processing pitch detector design has been tested using the 58-speaker Texas Instruments pitch detection database [16]. This database has been carefully cross-balanced for the sex and age of the speaker and contains speakers ranging in age from 3 to 86 years old, thereby ensuring a reasonable distribution of fundamental frequency. The speech material consists of Harvard phonetically balanced sentences. It is useful to note that the parallel processing pitch detector performance, shown below, is invariant to the specific database used.

In this section, we first examine the performance of the parallel processing algorithm versus several of its key parameters. Second, we measure the performance of each individual pitch detector, and contrast it to the composite system performance. Third, we compare performance of a computer simulation and a hardware implementation of the parallel processing algorithm to other state of the art algorithms. Finally, the performance of the parallel processing algorithm is evaluated under noisy conditions.

A perceptually weighted objective measure is used to compare pitch and voicing estimates from the pitch detector to reference pitch contours that have been constructed for the database [16], [17]. These reference pitch tracks were constructed under a criterion of optimum synthetic speech quality for an LPC vocoder operating with a 10 ms frame duration and with 14th-order LPC analysis. This objective measure is known to have a high correlation with subjective listening tests [2], [16]–[18] for several speech databases under a variety of recording conditions.

The objective measure tabulates errors in three classes: gross pitch errors (GPE), voiced-to-unvoiced errors (V–U), and unvoiced-to-voiced errors (U–V). A gross pitch error represents a correctly classified voiced frame where the reference pitch track and the candidate pitch track differ in fundamental frequency. The GPE score is computed by summing the per-frame gross pitch errors over all frames in the database. The per-frame gross pitch error is given by [16]

$$\mathrm{GPE}_m = E_m/E_{\max}\left(\left|(F_m - FR_m)/FR_m\right|^2\right) FR_m/500 \tag{18}$$

where E_m is the rms energy in the mth frame, $E_{\max}$ is the maximum rms energy in the sentence, F_m is the pitch frequency in the mth frame for the test contour, FR_m is the reference pitch frequency, and 500 Hz is the maximum pitch frequency considered.

A voiced-to-unvoiced error represents a voiced frame detected as unvoiced by the pitch tracker. Similar to the GPE score, the V–U score is computed by summing the per-frame voiced-to-unvoiced errors over all frames in the database. The per-frame voiced-to-unvoiced error is given by [16]

$$\begin{aligned}(\mathrm{V\text{–}U})_m &= (\mathrm{V\text{–}U})_m^I \quad \text{for } m \text{ in the interior of a reference voiced segment} \\ &= (\mathrm{V\text{–}U})_m^E \quad \text{for } m \text{ at the ends of a reference voiced segment}\end{aligned} \tag{19}$$

where

$$(\mathrm{V\text{–}U})_m^I = E_m/E_{\max}(1 + FR_m/500) \tag{20}$$

and

$$(\mathrm{V\text{–}U})_m^E = E_m/E_{\max}(FR_m/500). \tag{21}$$

An unvoiced-to-voiced error represents an unvoiced frame classified as voiced. Similar to the GPE and V–U scores, the U–V score is computed by summing the per-frame unvoiced-to-voiced errors over all frames in the da-

tabase. The per-frame U–V is given by [16]

$$(\text{U-V})_m = (\text{U-V})_m^I \quad \text{for } m \text{ in the interior of a test voiced segment}$$
$$= (\text{U-V})_m^E \quad \text{for } m \text{ at the ends of a test voiced segment} \qquad (22)$$

where

$$(\text{U-V})_m^I = E_m/E_{\max}(1 + FR_m/500) \qquad (23)$$

and

$$(\text{U-V})_m^E = E_m/E_{\max}(FR_m/500). \qquad (24)$$

The total objective score (TOS) is the sum of the three measures given above, that is,

$$\text{TOS} = \text{GPE} + (\text{V-U}) + (\text{U-V}). \qquad (25)$$

Note that a TOS of zero represents a perfect score, indicating that the test contour is identical to the reference contour. A high TOS score indicates a large difference between the reference and test contours. For the objective evaluations, the parallel processing algorithm used a 20 ms frame period, typical of a 2400 bit/s pitch-excited LPC vocoder.

In the pitch detector design described in the previous section, there are three parameters of importance. The first is referred to as the amplitude threshold g. This threshold requires the amplitude of a candidate pulse to be a certain percentage of the maximum amplitude in the frame. In Table II, we compare the performance of the individual arms of the pitch detector for several values of this amplitude threshold. The performance is optimal when $g = 0.5$ for the speech waveform-based arms of the pitch detector, while performance is optimal when $g = 0.25$ for the LPC residual-based arms of the pitch detector.

In general, the ability to perform accurate voicing decisions diminishes in each arm of the pitch detector as the amplitude threshold deviates from its optimal value. When the threshold is below its optimum value, extraneous candidate pulses are included in the search process, increasing the probability that an unvoiced frame will satisfy the periodicity check and be declared as a voiced frame. Thus, more unvoiced frames will be classified as voiced (U–V errors). Alternately, when the threshold is above its optimum value, valid candidate pulses are omitted from the search procedure, increasing the probability that a valid voiced frame will be declared unvoiced (V–U errors). These experimental results are consistent with the theoretical results presented in the previous section.

The second parameter of importance is the amplitude interpolation threshold q. This threshold is used to ensure that the amplitude envelope of the candidate pulses matches the generally smooth envelope of voiced speech. Table III shows performance versus the interpolation threshold. Tables II and III exhibit similar performance behavior. Optimum performance is obtained when q is 0.75 for the speech waveform-based detectors and when q is 0.55 for the LPC residual-based detectors. The results

TABLE II
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE INDIVIDUAL PITCH DETECTORS FOR DIFFERENT VALUES OF THE AMPLITUDE THRESHOLD g (THE PERCENT VOICING ERRORS WITH RESPECT TO ALL FRAMES ARE GIVEN IN PARENTHESES)

Signal	Amplitude Threshold "g"			
	0.125	0.25	0.5	0.75
-speech	3.63 (17.83)	3.47 (17.27)	3.29 (16.73)	4.15 (20.14)
+speech	2.80 (15.94)	2.60 (15.23)	2.47 (14.88)	3.15 (18.06)
-residual	10.60 (37.95)	10.53 (37.77)	10.98 (38.93)	16.98 (50.06)
+residual	7.17 (29.71)	7.00 (29.36)	7.59 (30.85)	14.24 (44.33)

TABLE III
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE INDIVIDUAL PITCH DETECTORS FOR DIFFERENT VALUES OF THE AMPLITUDE INTERPOLATION THRESHOLD q (THE PERCENT VOICING ERRORS WITH RESPECT TO ALL FRAMES ARE GIVEN IN PARENTHESES)

Signal	Interpolation Amplitude Threshold "q"				
	0.45	0.55	0.65	0.75	0.85
-speech	4.09 (18.95)	3.83 (18.34)	3.68 (17.72)	3.47 (17.27)	3.74 (17.51)
+speech	3.23 (16.71)	2.87 (15.93)	2.71 (15.64)	2.60 (15.23)	2.81 (15.82)
-residual	8.58 (34.61)	8.43 (34.20)	8.71 (34.59)	10.53 (37.77)	12.97 (41.79)
+residual	6.56 (25.09)	6.28 (24.66)	6.65 (25.36)	8.00 (29.36)	11.68 (34.20)

presented in Tables II and III indicate that it is important to prescreen candidate pulses to ensure successful and robust operation of the periodicity checking procedure.

The optimal values of the parameters g and q are definitely signal dependent. This dependency is due to the fact that the LPC residual is a spectrally flattened version of the speech signal. The envelope of the residual is therefore less smooth, resulting in the need for a smaller value of the two amplitude thresholds. Thus, the values of g and q must be smaller for the arms of the pitch detector operating on the LPC residual, as is shown in Tables II and III.

The third important threshold that influences performance of the individual pitch detectors is the breathing threshold B. In Table IV, performance is evaluated as a function of the breathing threshold. The breathing threshold is varied as a percentage of the value determined by (16). When the breathing threshold falls below b, the periodicity check is less reliable, and the V–U errors increase significantly. Similarly, when the breathing threshold is larger than b, the U–V errors increase.

While one can readily argue that all the information required to perform pitch detection is contained in either the speech waveform or the LPC residual, we find that significant improvements in performance can be obtained by the parallel processing strategy. This stems from the fact that there is no definitive way to prescreen candidate

TABLE IV
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE INDIVIDUAL PITCH DETECTORS FOR DIFFERENT VALUES OF THE BREATHING THRESHOLD *B* (THE PERCENT VOICING ERRORS WITH RESPECT TO ALL FRAMES ARE GIVEN IN PARENTHESES)

Signal	Breathing Threshold "B"				
	B = 0.70b	B = 0.80b	B = b	B = 1.20b	B = 1.30b
-speech	4.16 (18.68)	3.80 (18.05)	3.42 (17.32)	3.62 (17.61)	4.05 (18.06)
+speech	3.14 (16.44)	2.81 (15.64)	2.45 (15.21)	2.54 (15.62)	2.98 (15.94)
-residual	11.63 (39.86)	10.92 (38.30)	9.88 (36.92)	10.06 (36.96)	10.83 (37.04)
+residual	7.39 (29.98)	7.01 (29.42)	6.78 (29.34)	6.93 (29.57)	7.16 (29.81)

pulses such that only valid pitch pulses remain in the periodicity checking procedure. Even when all candidate pulses have been correctly identified in a particular waveform, natural variations in the fundamental frequency will sometimes create failures in the periodicity checking procedure. In Table V, the performance of each individual pitch detector is compared to the composite system. In this experiment, the pitch detector parameters are set to their optimal values based on the data in Tables II–IV. In general, the LPC residual gives poorer pitch estimates than the speech signal.

It is important in any pitch detection evaluation to compare the performance of the parallel processing algorithm to state of the art in pitch detection [18]. The performance of several common pitch detectors is compared to two versions of the parallel processing algorithm in Table VI. With the 58-speaker pitch detection database, audible differences between pitch detectors that give a composite performance below 1.0 are generally not significant. Performance at this level has saturated, and a broader database is required to reliably discriminate differences. As can be seen from Table VI, the parallel processing algorithm does significantly better in determining both voicing and pitch frequency information.

Finally, in Tables VII and VIII, the performance is evaluated for various signal-to-noise ratio (SNR) conditions. The data in Table VII represent the pitch detector performance when noise from a long-distance telephone connection is added to the 58-speaker database. A time-averaged FFT of this telephone noise is shown in Fig. 10. The test conditions for Table VIII were created by adding actual helicopter noise, as heard inside the cockpit, to the 58-speaker database. A time-averaged FFT of the helicopter noise is shown in Fig. 11. The level of the additive noise in both cases was adjusted such that a particular SNR is maintained on a per-file basis. It is clear from Tables VII and VIII that good performance is maintained at SNR levels down to 10 dB. Below 10 dB, the SNR is so low that the speech waveform itself becomes obscured, and pitch pulses are difficult to identify from the noisy waveform. This causes a significant increase in the V–U errors, and as Tables VII and VIII indicate, this type of

TABLE V
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE INDIVIDUAL PITCH DETECTORS AND THE FINAL OUTPUT OF THE COMPOSITE SYSTEM (THE PERCENT VOICING ERRORS WITH RESPECT TO ALL FRAMES ARE GIVEN IN PARENTHESES). GPE REPRESENTS GROSS PITCH ERRORS; V–U REPRESENTS VOICED-TO-UNVOICED ERRORS; U–V REPRESENTS UNVOICED-TO-VOICED ERRORS; AND TOTAL IS THE SUM OF ALL THE ERRORS

Signal	GPE	V-U	U-V	Total
-speech	0.07	3.05 (16.2)	0.04 (1.3)	3.16 (17.5)
+speech	0.03	2.24 (11.5)	0.06 (1.3)	2.33 (12.8)
-residual	0.10	8.13 (42.5)	0.01 (0.2)	8.24 (42.7)
+residual	0.05	6.15 (31.6)	0.01 (0.2)	6.21 (31.8)
Composite	0.10	0.08 (1.2)	0.11 (3.2)	0.29 (4.3)

TABLE VI
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE SEVERAL PITCH DETECTORS EVALUATED BY SECREST AND DODDINGTON [17] COMPARED TO THE PARALLEL PROCESSING PITCH DETECTOR (PPPD) COMPUTER SIMULATION AND FIRMWARE IMPLEMENTATION. GPE REPRESENTS GROSS PITCH ERRORS; V–U REPRESENTS VOICED-TO-UNVOICED ERRORS; U–V REPRESENTS UNVOICED-TO-VOICED ERRORS; AND TOTAL IS THE SUM OF ALL THE ERRORS

Pitch Detector	GPE	V-U	U-V	Total
Gold-Rabiner	0.25	4.08	0.56	4.90
Cepstral	0.39	1.62	1.85	3.86
Integrated Correlation	0.23	0.38	0.65	1.29
PPPD Simulation	**0.10**	**0.08**	**0.11**	**0.29**
PPPD Firmware	**0.10**	**0.08**	**0.11**	**0.29**

TABLE VII
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE PARALLEL PROCESSING PITCH DETECTOR FOR DIFFERENT SIGNAL-TO-NOISE RATIOS (SNR) WHERE THE ADDITIVE NOISE IS FROM A LONG-DISTANCE TELEPHONE CONNECTION. GPE REPRESENTS GROSS PITCH ERRORS; V–U REPRESENTS VOICED-TO-UNVOICED ERRORS; U–V REPRESENTS UNVOICED-TO-VOICED ERRORS; AND TOTAL IS THE SUM OF ALL THE ERRORS

SNR	GPE	V-U	U-V	Total
-10	0.18	20.63	0.08	20.89
-5	0.30	12.24	0.02	12.56
0	0.16	4.62	0.01	4.79
5	0.07	1.65	0.02	1.73
10	0.16	0.68	0.03	0.87
20	0.15	0.21	0.05	0.42
30	0.10	0.11	0.08	0.29
40	0.10	0.09	0.10	0.29

error is the dominant factor in the total objective measure when the SNR is below 10 dB. Tables VII and VIII also show the robustness of the pitch detector to different noise characteristics.

Since the majority of the waveform-based parallel processing algorithm consists of integer arithmetic and logic-type operations, there should be no great obstacle in maintaining the performance of the computer simulation in hardware. The only portion of the algorithm that re-

TABLE VIII
PERCEPTUALLY WEIGHTED PERFORMANCE OF THE PARALLEL PROCESSING PITCH DETECTOR FOR DIFFERENT SIGNAL-TO-NOISE RATIOS (SNR) WHERE THE ADDITIVE NOISE IS HELICOPTER NOISE AS HEARD INSIDE THE COCKPIT. GPE REPRESENTS GROSS PITCH ERRORS; V–U REPRESENTS VOICED-TO-UNVOICED ERRORS; U–V REPRESENTS UNVOICED-TO-VOICED ERRORS; AND TOTAL IS THE SUM OF ALL THE ERRORS

SNR	GPE	V-U	U-V	Total
-10	0.15	21.30	0.11	21.56
-5	0.21	11.91	0.11	12.23
0	0.21	5.01	0.08	5.30
5	0.11	1.89	0.03	2.04
10	0.10	0.77	0.03	0.89
20	0.10	0.25	0.05	0.40
30	0.10	0.12	0.07	0.29
40	0.10	0.09	0.10	0.29

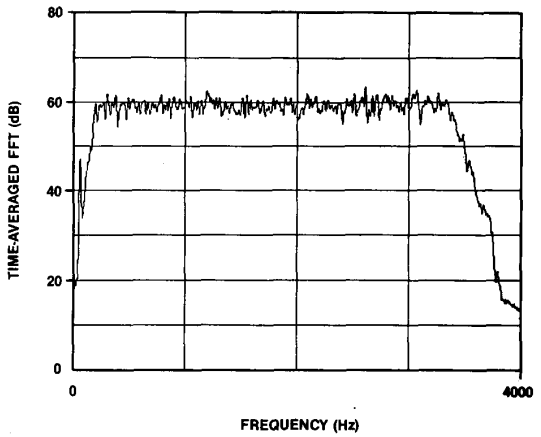


Fig. 10. Time-averaged fast Fourier transform in dB of noise from a long-distance telephone connection, measured over a 4 kHz frequency band.

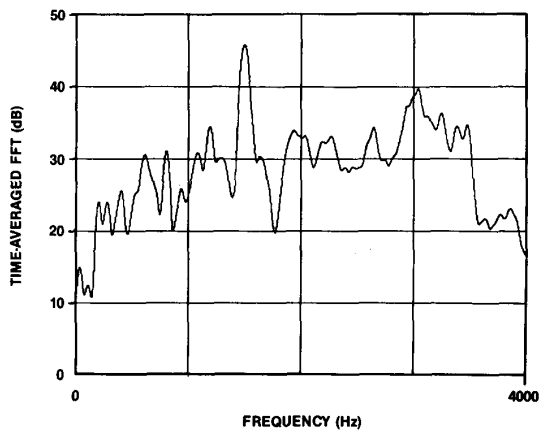


Fig. 11. Time-averaged fast Fourier transform in dB of helicopter noise as heard inside the cockpit, measured over a 4 kHz frequency band.

quires notable arithmetic precision is the LPC analysis and the LPC residual generation. Fixed-point implementation of LPC analysis is a relatively mature process. The performance of the firmware implementation described in the next section was shown in Table VI, and is identical to the floating-point computer simulation. The parallel processing pitch detection algorithm is attractive not only because of its robust performance, but also because of its insensitivity to fixed-point arithmetic.

V. FIRMWARE IMPLEMENTATION AND OPERATION

The parallel processing pitch detector was implemented in real time as part of a 2.4 kbit/s vocoder system. The complete full-duplex system (LPC parameter computation, pitch detection, and LPC synthesis) was implemented on a single TMS32020 digital signal processor with a 200 ns instruction cycle time. The performance analysis given in the previous section indicated that the real-time fixed-point implementation of the pitch detector has identical performance to that of a computer simulation.

The TMS32020 architecture is well suited for efficient implementation of the pitch detector. It is clear why efficient implementation is desirable. We need only to consider the real time required for the LPC parameter computation, the LPC synthesis process, and execution of the pitch detector routine four times per frame. One especially useful feature of the TMS32020 when defining candidate pulses is the 544 words of on-chip data memory. Finding the candidate pulses involves frequent addressing of the speech and residual sample points and consumes a large portion of the total execution time for an individual pitch detector. It is therefore desirable to minimize this addressing time. This can be easily done by storing all of the frame's 320 speech and residual sample points on-chip, thereby reducing the addressing time to a minimum of one instruction cycle time compared to two cycles if the data were to be stored off-chip.

Table IX shows the major operations of the individual pitch detector and their corresponding average execution time computed as a percentage of the total execution time for an individual pitch detector. It must be emphasized that these percentage execution times are per-frame averages, implying some variability from frame to frame. For instance, if the frame is voiced, the probability of finding periodicity in the first or second pass of the periodicity checking procedure is much larger than that if the frame is unvoiced. In fact, for the unvoiced case, the periodicity checking procedure must consider all possible subsets of candidate pulses for periodicity before declaring the frame unvoiced. Therefore, a longer execution time for the periodicity checking procedure is needed in the case of unvoiced frames than in the case of voiced frames. Another example showing the variability of the above percentage execution times occurs when comparing the execution times of a high-pitched speaker versus a low-pitched speaker. If the speaker is high pitched, then there is a larger number of pitch pulses in a voiced frame. Consequently, a larger number of candidate pulses is defined, implying longer execution times for defining candidate pulses and performing the amplitude interpolation test. It was determined that the four arms of the parallel processing pitch detector combined consume 25–35 percent of the total available real time and the pitch voter consumes 4 percent of the total real time.

TABLE IX
AVERAGE EXECUTION TIMES PER FRAME COMPUTED AS A PERCENTAGE OF THE TOTAL EXECUTION TIME FOR AN INDIVIDUAL PITCH DETECTOR

Operation	Percentage Excution Time
Candidate Pulse Definition and Amplitude Test	55%
Interpolation Test	15%
Periodicity Check	30%

VI. CONCLUSIONS

A pitch detection algorithm based on a parallel processing technique has been presented. Accurate pitch detection is achieved by suitably combining pitch estimates from four individual pitch detectors operating on four different signals. The performance of this parallel processing pitch detector was evaluated on a large database and compared to other well-known pitch detection algorithms. Our results show that this pitch detector has a superior performance and gives a very reliable and accurate pitch estimate. Furthermore, this performance is maintained in a real-time fixed-point implementation, implying that this algorithm is insensitive to the limitations of fixed-point arithmetic. This is attributed to the fact that the majority of the algorithm computations are integer arithmetic and logic-type operations.

It is clear from the performance analysis that the pitch detector performance on studio quality speech is excellent. Under noisy conditions, our results show that very good pitch detector performance is maintained at SNR down to 10 dB. The results also show the robustness of the pitch detector operating on noisy speech with different noise characteristics. Future efforts are directed toward maintaining good performance under severe noisy conditions, particularly at SNR below 10 dB.

REFERENCES

[1] W. Hess, *Pitch Determination of Speech Signals.* New York: Springer-Verlag, 1983.

[2] J. Picone, G. R. Doddington, and B. G. Secrest, "Robust pitch detection in a noisy telephone environment," in *Proc. 1987 Int. Conf. Acoust., Speech, Signal Processing*, Dallas, TX, Apr. 1987, pp. 1442-1445.

[3] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 24-33, 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, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353-362, Oct. 1974.

[6] C. K. Un and S. C. Yang, "A pitch extraction algorithm based on LPC inverse filtering and AMDF," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 567-572, Dec. 1977.

[7] S. Seneff, "Real-time harmonic pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 358-365, Aug. 1978.

[8] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309, Feb. 1967.

[9] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1973.

[10] 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.

[11] 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, Signal Processing*, vol. ASSP-24, pp. 399-417, Oct. 1976.

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[13] D. P. Prezas, J. Picone, and D. L. Thomson, "Fast and accurate pitch detection using pattern recognition and adaptive time-domain analysis," in *Proc. 1986 Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, Apr. 1986, pp. 109-112.

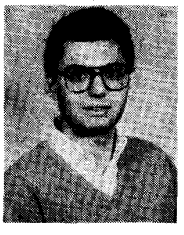
[14] R. A. Sukkar, "A parallel processing pitch detector for LPC," M.S. thesis, Dep. Elec. Comput. Eng., Illinois Inst. Technol., Chicago, 1984.

[15] J. L. LoCicero and R. A. Sukkar, "Pitch detection: A parallel processing technique," in *Proc. 1986 Nat. Commun. Forum*, Chicago, IL, Sept. 1986, pp. 1237-1242.

[16] B. G. Secrest and G. Doddington, "Postprocessing techniques for voice pitch trackers," in *Proc. 1982 Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 172-175.

[17] ——, "An integrated pitch tracking algorithm for speech systems," in *Proc. 1983 Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, Apr. 1983, pp. 1352-1355.

[18] V. R. Viswanathan and W. H. Russell, "New objective measures for the evaluation of pitch extractors," in *Proc. 1985 Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, Mar. 1985, pp. 411-414.



Rafid A. Sukkar was born in Baghdad, Iraq, on February 2, 1962. He received the B.S.E.E. and M.S.E.E. degrees from the Illinois Institute of Technology, Chicago, in 1982 and 1984, respectively.

He is currently a doctoral candidate in the Department of Electrical and Computer Engineering at I.I.T. During 1983 he was a Teaching Assistant, responsible for teaching the Digital Systems and Microprocessor Laboratories. Since 1984 he has been a Research Assistant, conducting research funded by AT&T Bell Laboratories. He conducted his M.S. and Ph.D. research using the speech processing facilities of the Exploratory Voice Capabilities Department at AT&T Bell Laboratories, Naperville, IL. His research interests include digital processing of speech signals and spectral estimation.

Mr. Sukkar is a member of Tau Beta Pi.

Joseph L. LoCicero (S'75-M'76) was born in New York, NY, on September 18, 1947. He received the B.E.E. and M.E.E. degrees from the City College of New York, New York, NY, in 1970 and 1971, respectively, and the Ph.D. degree in electrical engineering from the City University of New York in 1976.

From 1976 to 1982 he was an Assistant Professor of Electrical Engineering at the Illinois Institute of Technology, Chicago. From 1982 to 1987 he held the rank of Associate Professor, and in the Fall 1987 was promoted to Full Professor. He served as Assistant Chairman of the Department of Electrical and Computer Engineering from 1982 to 1986. Since Fall 1986 he has held the position of Acting Chairman of the ECE Department at I.I.T. He is technically active, authoring over 50 professional articles, co-editing a ComSoc book, *Tutorials in Modern Communications*, and being an Associate Editor of the IEEE Press book *Television Technology Today.* He holds four patents in the high-definition television and video communications areas. His research interests include

digital processing of speech and video signals, bandwidth compression, linear predictive coding, high-definition television, and computer-generated holography for photonic switching; he also served as a consultant in isolated word speech recognition.

Dr. LoCicero is very active in the IEEE Communications Society. From 1976 to 1978 he was Assistant to the Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, and from 1978 to January 1988 served as the Publications Editor. He is also quite active on the Communication Theory Committee, chairing or organizing sessions at many of the ComSoc conferences. In addition, he served as Assistant Technical Program Committee Chairman at ICC '85, held in Chicago, IL. He is a member of Eta Kappa Nu, Tau Beta Pi, Sigma Xi, the New York Academy of Sciences, and the Association of Engineering Educators. He has served as Faculty Advisor to the Student Chapters of the IEEE and Eta Kappa Nu in the Department of Electrical and Computer Engineering at I.I.T.

Joseph W. Picone (S'79-M'83) received the B.S.E.E. degree in 1979, the M.S.E.E. degree in 1980, and the Ph.D. degree in 1983 in electrical engineering, all from the Illinois Institute of Technology, Chicago.

During 1981-1983 he was involved in medium-rate speech coding research at Bell Laboratories, Naperville, IL. From 1983 to 1985 he was engaged in low-bit-rate speech coding research at Texas Instruments, Inc., Dallas, TX. In 1985 he returned to AT&T Bell Laboratories, Naperville, to do research in low-bit-rate speech coding and small-vocabulary speaker-independent speech recognition. He is currently with the Speech and Image Understanding Laboratories, Texas Instruments, Inc., Dallas, conducting research into speaker-independent speech recognition.