A Comparative Performance Study of Several Pitch Detection Algorithms

LAWRENCE R. RABINER, FELLOW, IEEE, MICHAEL J. CHENG, STUDENT MEMBER, IEEE, AARON E. ROSENBERG, MEMBER, IEEE, AND CAROL A. McGONEGAL

Abstract—A comparative performance study of seven pitch detection algorithms was conducted. A speech data base, consisting of eight utterances spoken by three males, three females, and one child was constructed. Telephone, close talking microphone, and wideband recordings were made of each of the utterances. For each of the utterances in the data base, a "standard" pitch contour was semiautomatically measured using a highly sophisticated interactive pitch detection program. The "standard" pitch contour was then compared with the pitch contour that was obtained from each of the seven programmed pitch detectors. The algorithms used in this study were 1) a center clipping, infinite-peak clipping, modified autocorrelation method (AUTOC), 2) the cepstral method (CEP), 3) the simplified inverse filtering technique (SIFT) method, 4) the parallel processing time-domain method (PPROC), 5) the data reduction method (DARD), 6) a spectral flattening linear predictive coding (LPC) method, and 7) the average magnitude difference function (AMDF) method. A set of measurements was made on the pitch contours to quantify the various types of errors which occur in each of the above methods. Included among the error measurements were the average and standard deviation of the error in pitch period during voiced regions, the number of gross errors in the pitch period, and the average number of voiced-unvoiced classification errors. For each of the error measurements, the individual pitch detectors could be rank ordered as a measure of their relative performance as a function of recording condition, and pitch range of the various speakers. Performance scores are presented for each of the seven pitch detectors based on each of the categories of error.

I. INTRODUCTION

A PITCH DETECTOR is an essential component in a variety of speech processing systems. Besides providing valuable insights into the nature of the excitation source for speech production, the pitch contour of an utterance is useful for recognizing speakers [1], [2], for speech instruction to the hearing impaired [3], and is required in almost all speech analysis-synthesis (vocoder) systems [4]. Because of the importance of pitch detection, a wide variety of algorithms for pitch detection have been proposed in the speech processing literature (e.g., [5]-[11]). In spite of the proliferation of pitch detectors, very little formal evaluation and comparison among the different types of pitch detectors has been attempted. There are a wide variety of reasons why such an evaluation has not been made. Among these are the difficulty in selection of a reasonable standard of comparison; collection of a comprehensive data base; choice of pitch detectors for evaluation; and the difficulty in interpreting the results in a meaningful and unbiased way. This paper is a report on an attempt to provide such a performance evaluation of seven pitch detection algorithms. Although a wide variety of alternatives were available in almost every aspect of this investigation, several arbitrary (but hopefully reasonable) decisions were made to limit the scope of the performance evaluation to a reasonable size.

Manuscript received March 3, 1976; revised June 7, 1976.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

In this section we provide an overview of the investigation. We begin with a discussion of the general problems and issues in pitch detection. Then we discuss the various types of pitch detection algorithms which have been proposed and review their general characteristics. We conclude with a discussion of the types of performance measures which would be suitable for various applications.

In Section II the detailed implementations of the seven pitch detectors used in this study are reviewed. Included in this section is a brief discussion of the method of operation of the pitch detector and the method used to make a voiced-unvoiced classification. In Section III we discuss the data base selected for evaluating the seven pitch detectors. In Section IV the method used to measure the standard pitch contour is outlined. Section V presents the results of several error analyses. Section VI provides a discussion of the error analyses and Section VII discusses the computational considerations in the implementation of each of the algorithms.

A. Problems in Pitch Detection

Accurate and reliable measurement of the pitch period of a speech signal from the acoustic pressure waveform alone is often exceedingly difficult for several reasons. One reason is that the glottal excitation waveform is not a perfect train of periodic pulses. Although finding the period of a perfectly periodic waveform is straightforward, measuring the period of a speech waveform, which varies both in period and in the detailed structure of the waveform within a period, can be quite difficult. A second difficulty in measuring pitch period is the interaction between the vocal tract and the glottal excitation. In some instances the formants of the vocal tract can alter significantly the structure of the glottal waveform so that the actual pitch period is difficult to detect. Such interactions generally are most deleterious to pitch detection during rapid movements of the articulators when the formants are also changing rapidly. A third problem in reliably measuring pitch is the inherent difficulty in defining the exact beginning and end of each pitch period during voiced speech segments. The choice of the exact beginning and ending locations of the pitch period is often quite arbitrary. For example, based on the acoustic waveform alone, some candidates for defining the

beginning and end of the period include the maximum value during the period, the zero crossing prior to the maximum, etc. The only requirement on such a measurement is that it be consistent from period-to-period in order to be able to define the "exact" location of the beginning and end of each pitch period. The lack of such consistency can lead to spurious pitch period estimates. Fig. 1 shows two possible choices for defining a pitch marker directly based on waveform measurements. The two waveform measurements shown in Fig. 1 can (and often will) give slightly different values for the pitch period. The pitch period discrepancies are due not only to the quasiperiodicity of the speech waveform, but also the fact that peak measurements are sensitive to the formant structure during the pitch period, whereas zero crossings of a waveform are sensitive to the formants, noise, and any dc level in the waveform. A fourth difficulty in pitch detection is distinguishing between unvoiced speech and low-level voiced speech. In many cases transitions between unvoiced speech segments and low-level voiced speech segments are very subtle and thus are extremely hard to pinpoint.

In addition to the difficulties in measuring pitch period discussed above, additional complications occur when one is faced with the problem of pitch extraction of speech that has been transmitted through the telephone system. Many systems, in which pitch detection is required, must process telephone-quality speech. The effects of the telephone system on speech include linear filtering, nonlinear processing, and the addition of noise to the speech signal. With regard to linear filtering, the telephone system acts like a bandpass filter (low-frequency cutoff of approximately 200 Hz, high-frequency cutoff of approximately 3200 Hz) which can significantly attenuate the fundamental pitch frequency and many of the higher pitch harmonics. The result is that the periodicity of the signal is much harder to detect. Nonlinear contributions of the telephone system to the speech signals can, depending on the particular transmission system used, include the following.

1) Phase distortion.
2) Fading or amplitude modulation of the speech signal.
3) Crosstalk between two or more messages.
4) Clipping or distortion of extremely high-level sounds.

(It should be noted that one would *not* expect all the above listed effects to occur simultaneously.) Thus the overall effect of the telephone line is to obscure the periodic structure of the speech waveform such that the pitch period becomes more difficult to detect.

B. Types of Pitch Detectors

As a result of the numerous difficulties in pitch measurements, a wide variety of sophisticated pitch detection methods have been developed. Basically, a pitch detector is a device which makes a voiced-unvoiced decision, and, during periods of voiced speech, provides a measurement of the pitch period. However, some pitch detection algorithms just determine the pitch during voiced segments of speech and rely on some other technique for the voiced-unvoiced decisions. Pitch detection

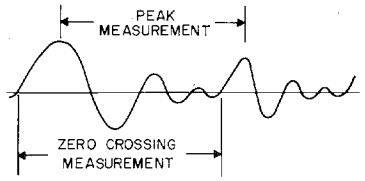


Fig. 1. Two waveform measurements which can be used to define pitch markers.

algorithms can roughly be divided into the following three broad categories.

1) A group which utilizes principally the time-domain properties of speech signals.
2) A group which utilizes principally the frequency-domain properties of speech signals.
3) A group which utilizes both the time- and frequency-domain properties of speech signals.

Time-domain pitch detectors operate directly on the speech waveform to estimate the pitch period. For these pitch detectors the measurements most often made are peak and valley measurements, zero-crossing measurements, and autocorrelation measurements. The basic assumption that is made in all these cases is that if a quasiperiodic signal has been suitably processed to minimize the effects of the formant structure, then simple time-domain measurements will provide good estimates of the period.

The class of frequency-domain pitch detectors use the property that if the signal is periodic in the time domain, then the frequency spectrum of the signal will consist of a series of impulses at the fundamental frequency and its harmonics. Thus simple measurements can be made on the frequency spectrum of the signal (or a nonlinearly transformed version of it as in the cepstral pitch detector [5]) to estimate the period of the signal.

The class of hybrid pitch detectors incorporates features of both the time-domain and the frequency-domain approaches to pitch detection. For example, a hybrid pitch detector might use frequency-domain techniques to provide a spectrally flattened time waveform, and then use autocorrelation measurements to estimate the pitch period.

In this investigation four time-domain, one frequency-domain, and two hybrid pitch detectors were studied. Detailed discussions of the algorithms which were used will be given in Section II.

C. Criteria for Evaluating Pitch Detectors

One of the most difficult problems in comparing and evaluating the performance of pitch detectors is choosing a meaningful objective performance criterion. The basic problem is that a criterion suitable for one application may not be suitable for a different application of pitch detectors.

There are many characteristics of pitch detection algorithms which influence the choice of a set of performance criteria. Among these factors are the following.

1) Accuracy in estimating pitch period.
2) Accuracy in making a voiced-unvoiced decision.
3) Robustness of the measurements, i.e., they must be

modified for different transmission conditions, speakers, etc.
4) Speed of operation.
5) Complexity of the algorithm.
6) Suitability for hardware implementation.
7) Cost of hardware implementation.

Depending on the specific application, various weights must be given to each of the above factors in choosing a single objective performance criterion. In this paper we will present results based on factors 1 and 2 in the above list, i.e., only those factors which are most amenable to numerical tabulations. However, whenever possible, we will try to discuss how factors 3-7 enter into an overall assessment of the pitch detectors discussed in this paper.

There is one major factor which was omitted from the above list and which, for many applications, is the dominant factor in evaluating pitch detectors. This factor is the perceptual accuracy of the pitch detectors, i.e., the question of how faithfully the pitch contour measured by the pitch detector matches the natural excitation pitch contour in terms of synthetic speech quality. We have omitted this factor from the list because it is not an objective performance criterion, but is instead a subjective criterion that can only be assessed through a series of extensive perceptual tests using synthetic speech samples. Such a companion investigation is being undertaken by the authors and will be reported on at a later date.

II. Pitch Detection Algorithms

As stated earlier, seven distinct algorithms for detecting pitch were investigated. These algorithms were the following.

1) Modified autocorrelation method using clipping (AUTOC) (Dubnowski [11]).
2) Cepstrum method (CEP) (Schafer [12]).
3) Simplified inverse filtering technique (SIFT) (Markel [8]).
4) Data reduction method (DARD) (Miller [9]).
5) Parallel processing method (PPROC) (Rabiner [2]).
6) Spectral equalization LPC method using Newton's transformation (LPC) (Atal, unpublished).
7) Average magnitude difference function (AMDF) (NSA version, [10]).

The names in parentheses are the individual (or group) responsible for providing Fortran code for the computational parts of each algorithm, and the code following the name of the method is the abbreviation which will be used to refer to the pitch detector throughout this paper.

The choice of pitch detectors was based on both practical considerations (i.e., availability of reasonably portable Fortran code) as well as the desire to choose a good cross section of recent examples of each of the three types of pitch detectors discussed in Section I. Thus, included in this study were two time-domain (waveform) pitch detectors (4 and 5), two autocorrelation pitch detectors (1 and 7), one frequency-domain pitch detector (2), and two LPC hybrid pitch detectors (3 and 6).

In the following section we provide a summary of the method of operation of each of the seven pitch detectors. Whenever possible, exact values of parameters of the pitch detector (e.g., section length, window, etc., will be given).

A. Modified Autocorrelation Method (AUTOC)

The modified autocorrelation pitch detector is based on the center-clipping method of Sondhi [7]. Fig. 2 shows a block diagram of the pitch detection algorithm. The method requires that the speech be low-passed filtered to 900 Hz. (A 99-point linear phase, finite impulse response (FIR) digital filter is used to low-pass filter the speech. Detailed characteristics of this low-pass filter, which is used for several of the pitch detectors, are given in [13].)

The low-pass filtered speech signal is digitized at a 10-kHz sampling rate and sectioned into overlapping 30-ms (300 samples) sections for processing. Since the pitch period computation for all pitch detectors is performed 100 times/s, i.e., every 10 ms, adjacent sections overlap by 20 ms or 200 samples.

The first stage of processing is the computation of a clipping level c_L for the current 30-ms section of speech. The clipping level is set at a value which is 64 percent of the smaller of the peak absolute sample values in the first and last 10-ms portions of the section. Following the determination of the clipping level, the 30-ms section of speech is center clipped, and then infinite peak clipped, resulting in a signal which assumes one of three possible values—+1 if the sample exceeds the positive clipping level, −1 if the sample falls below the negative clipping level, and 0 otherwise.

Following clipping the autocorrelation function for the 30-ms section is computed over a range of lags from 20 samples to 200 samples (i.e., 2-ms-20-ms period). Additionally, the autocorrelation at 0 delay is computed for appropriate normalization purposes. The autocorrelation function is then searched for its maximum (normalized) value. If the maximum (normalized value) exceeds 0.3, the section is classified as voiced and the location of the maximum is the pitch period. Otherwise, the section is classified as unvoiced.

In addition to the voiced-unvoiced classification based on the autocorrelation function, a preliminary test is carried out on each section of speech to determine if the peak signal amplitude within the section is sufficiently large to warrant the pitch computation. If the peak signal level within the section is below a given threshold,[1] the section is classified as unvoiced (silence) and no pitch computations are made. This method of eliminating low-level speech sections from further processing was also used for pitch detectors 2 (CEP), 3 (SIFT), and 5 (PPROC).

B. Cepstral Method (CEP)

Fig. 3 shows a flow diagram of the cepstral pitch detector described in [12]. Each block of 512 samples (51.2 ms) is weighted by a 512-point Hamming window, and then the cepstrum of that block is computed. The peak cepstral value and its location is determined and if the value of this peak exceeds a fixed threshold, the section is called voiced and the pitch period is the location of the peak. If the peak does not

[1] The threshold chosen is 1/15 of the peak absolute signal value within the utterance.

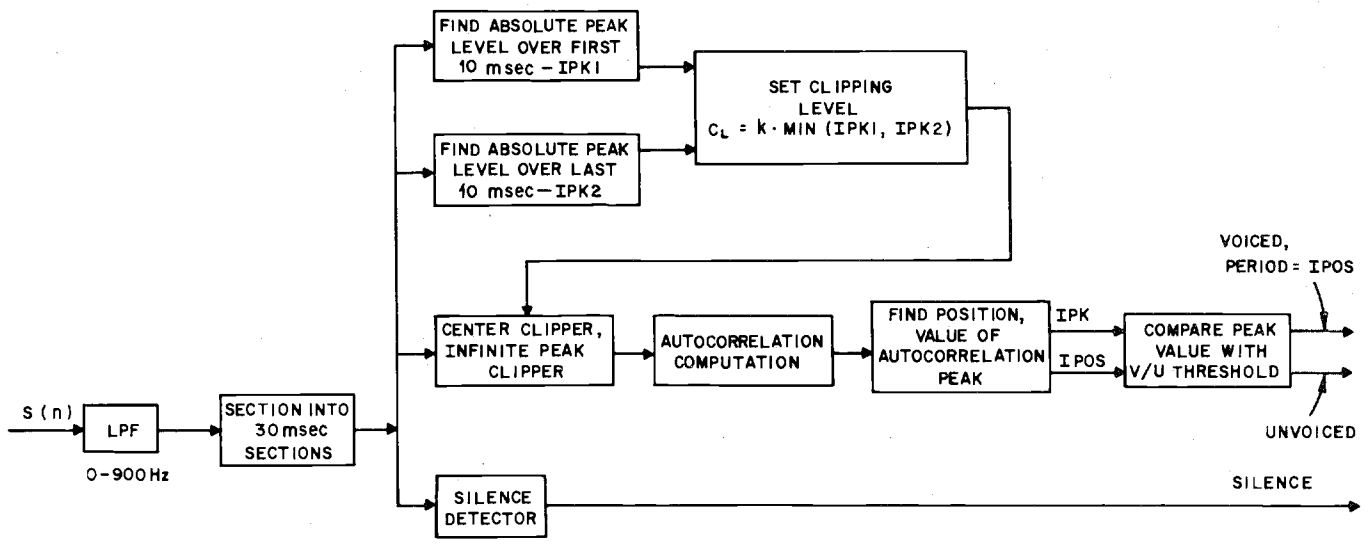


Fig. 2. Block diagram of the AUTOC pitch detector.

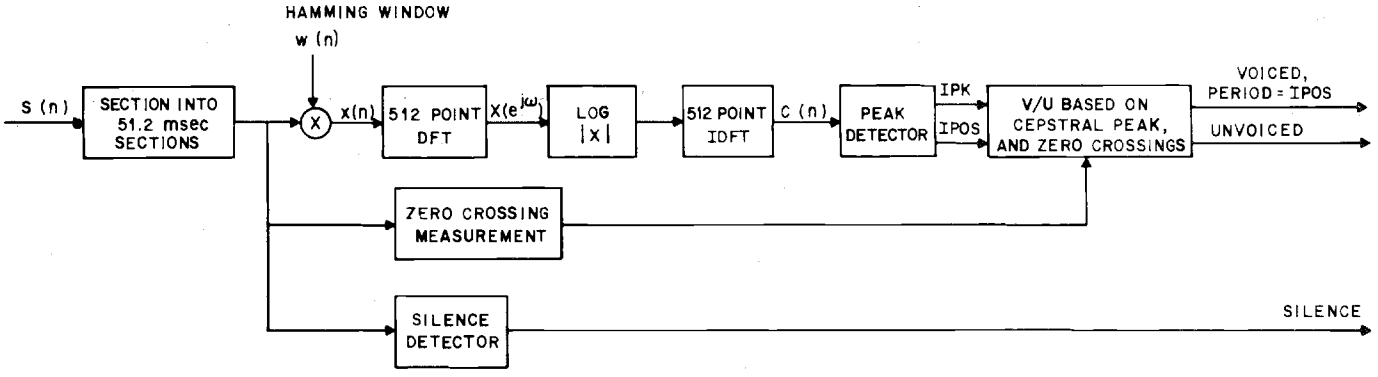


Fig. 3. Block diagram of the CEP pitch detector.

exceed the threshold, a zero-crossing count is made on the block. If the zero-crossing count exceeds a given threshold, the block is called unvoiced. Otherwise it is called voiced and the period is the location of the maximum value of the cepstrum.

As in the modified autocorrelation method, a preliminary silence detector (based on the signal level) is used to classify all low-level blocks as silence (unvoiced speech) prior to the cepstral computation. It should also be noted that the cepstral pitch detector uses the full-band speech signal for processing.

C. Simplified Inverse Filtering Technique (SIFT)

Fig. 4 shows a block diagram of the SIFT method of pitch detection [8]. A block of 400 speech samples (40 ms at a 10-kHz rate) is low-pass filtered to a bandwidth of 900 Hz, and then decimated (down sampled) by a 5 to 1 ratio. The coefficients of a 4th-order inverse filter are obtained using the autocorrelation method of LPC analysis [14]. The 2-kHz speech signal is then inverse filtered to give a spectrally flattened signal which is then autocorrelated. The pitch period is obtained by interpolating the autocorrelation function in the neighborhood of the peak of the autocorrelation function. A voiced-unvoiced decision is made on the basis of the amplitude of the peak of the autocorrelation function. The threshold used for this test is a normalized value of 0.4 for the autocorrelation peak.

As with the previous two pitch detectors, a preliminary silence detector is used to classify low-level sections as silence and eliminate them from further consideration.

D. Data Reduction Method (DARD)

Fig. 5 shows a block diagram of the data reduction pitch detector of Miller [9]. This pitch detector places pitch markers directly on a low-pass filtered (0-900 Hz) speech signal and thus is a pitch synchronous pitch detector.

To obtain the appropriate pitch markers, the data reduction method first detects excursion cycles in the waveform based on intervals between major zero crossings. The remainder of the algorithm tries to isolate and identify the principal excursion cycles, i.e., those which correspond to true pitch periods. This is accomplished through a series of steps using energy measurements and logic based on permissible pitch periods and anticipated syllabic rate changes of pitch. An error correction procedure is used to provide a reasonable measure of continuity in the pitch markers.

Since there is no inherent voiced-unvoiced calculation within this pitch detector, regions of unvoiced speech are identified by the lack of pitch markers.

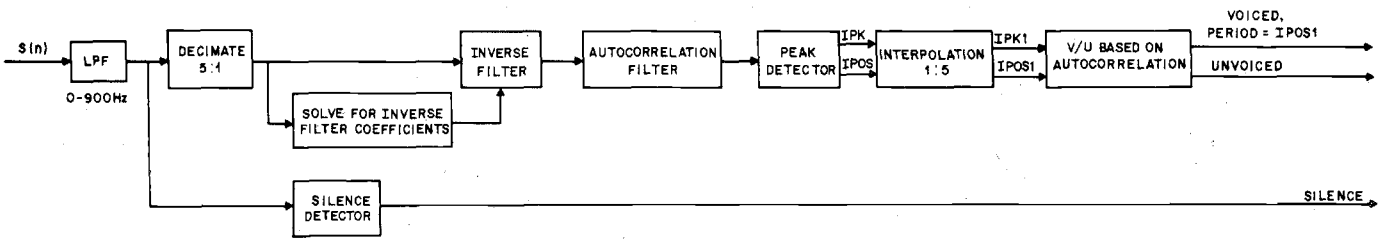


Fig. 4. Block diagram of the SIFT pitch detector.

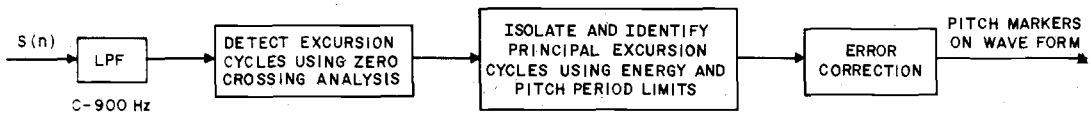


Fig. 5. Block diagram of the DARD pitch detector.

E. Parallel Processing Method (PPROC)

Fig. 6 shows a block diagram of the parallel processing pitch detector [6]. The speech signal is first low-pass filtered to a bandwidth of 900 Hz. Then a series of measurements are made on the peaks and valleys of the low-pass filtered signal to give six separate functions. Each of these six functions is processed by an elementary pitch period estimator (PPE), giving six separate estimates of the pitch period. The six pitch estimates are then combined by a sophisticated decision algorithm which determines the pitch period. A voiced-unvoiced decision is obtained based on the degree of agreement among the six pitch detectors. Additionally, the preliminary silence detector described in Section II-A is used to classify low-level segments as silence.

F. Spectral Equalization LPC Method Using Newton's Transformation (LPC)

Fig. 7 shows a block diagram of an LPC pitch detector proposed by Atal (unpublished). The first step in this pitch detector is a voiced-unvoiced detector which uses a pattern-recognition technique to classify each 10-ms interval of speech as voiced or unvoiced [15]. If the speech section is classified as voiced, the 10-kHz sampled speech is digitally low-pass filtered to a bandwidth of about 900 Hz, and then decimated by 5 to 1 to a 2-kHz sampling rate. A 41-pole LPC analysis is performed on a 40-ms frame of speech to give a good representation of the speech spectrum in terms of the pitch harmonics. A Newton transformation is used to spectrally flatten the speech, i.e., to transform the signal into one which has sharp peaks at the pitch impulses, and is approximately zero everywhere else. A peak picker is used to determine the pitch period at the 2-kHz rate and a simple interpolation network is used to obtain higher resolution in the value of the pitch period.

It should be pointed out that the voiced-unvoiced pattern recognition algorithm uses a training set which provides a statistical description of the measurements used in the algorithm for each of classes. The success of this method of making a voiced-unvoiced decision depends heavily on how well the training set of data characterizes the different speech classes. With careful training, voiced-unvoiced accuracies on the order of 99 percent have been obtained [15].

G. Average Magnitude Different Function (AMDF)

Fig. 8 shows a block diagram of the AMDF pitch detector [10]. (The version used in this study was kindly supplied by M. Malpass of the Massachusetts Institute of Technology Lincoln Laboratory, based on the NSA version of the AMDF method. Details of implementation differ somewhat from those of [10].) The speech signal, initially sampled at 10 kHz, is decimated to a 6.67-kHz rate using a system of the type discussed in [16]. A zero-crossing measurement (NOZ) is made on the full-band speech file, and an energy measurement (ENG) is made on a low-pass filtered version (0-900 Hz) of the signal. The average magnitude difference function is computed on the low-pass filtered speech signal at 48 lags running from 16 to 124 samples. The pitch period is identified as the value of the lag at which the minimum AMDF occurs. Thus a fairly coarse quantization is obtained for the pitch period. Logic is used to check for pitch period doubling, etc., and to check on continuity of pitch periods with previous pitch estimates (a type of nonlinear smoothing). In addition to the pitch estimate, the ratio between the maximum and minimum values of AMDF (MAX/MIN) is obtained. This measurement, along with NOZ and ENG is used to make a voiced-unvoiced decision using logical operations.

III. Data Base for Evaluation

In order to evaluate the performance of these seven pitch detectors, an appropriately chosen data base was required to span the range of pitch, types of utterances, and recording and transmission environments which are normally encountered in speech processing. In this section we describe the data base used in this study.

A. Speakers

The set of seven speakers for this study included the following.

1) Low-pitched male (LM).
2) Male speaker 1 (M1).
3) Male speaker 2 (M2).

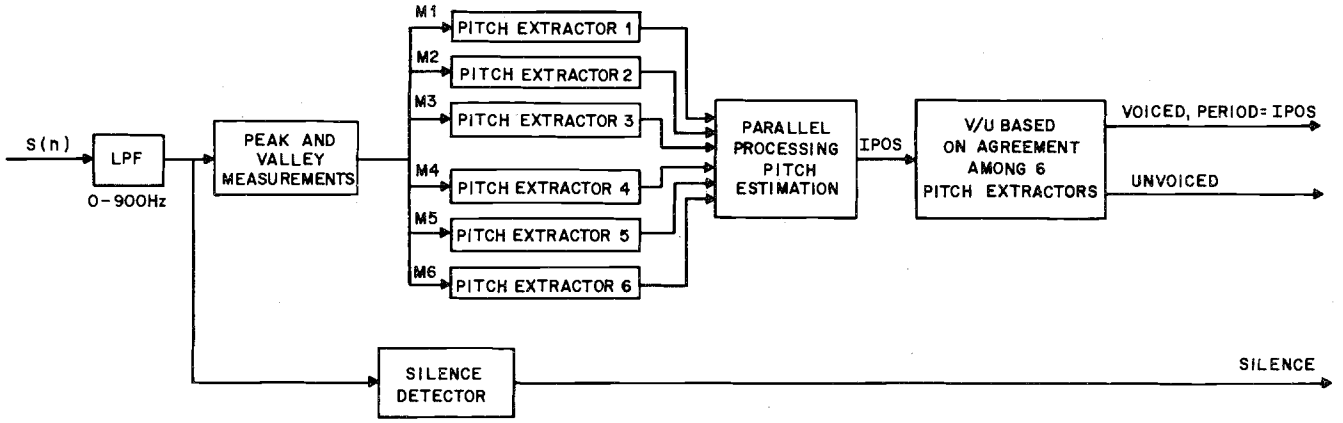


Fig. 6. Block diagram of the PPROC pitch detector.

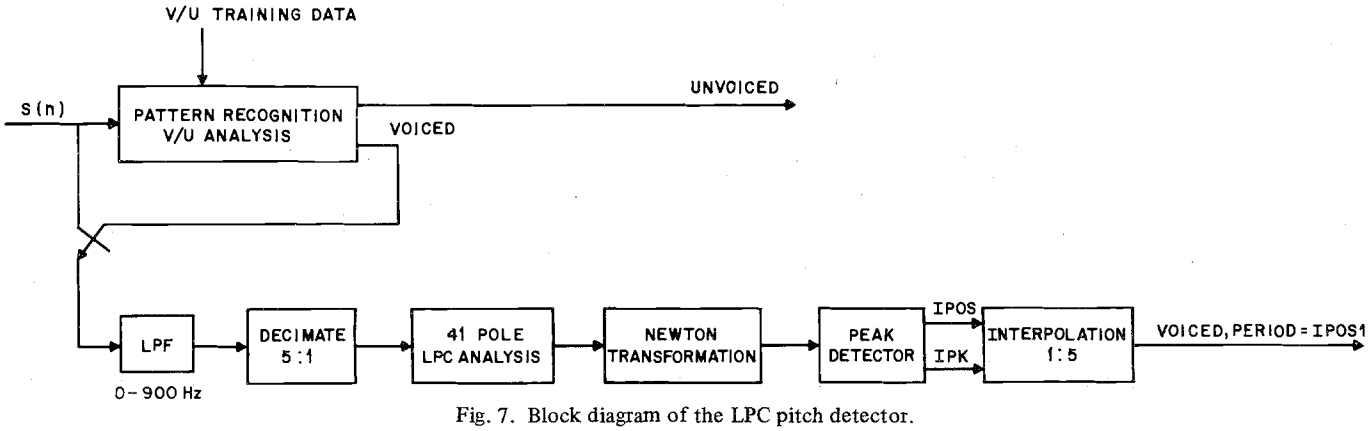


Fig. 7. Block diagram of the LPC pitch detector.

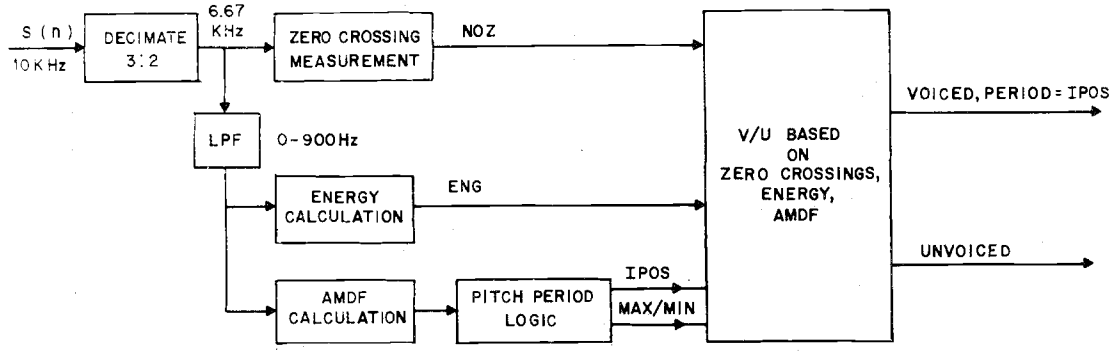


Fig. 8. Block diagram of the AMDF pitch detector.

4) Female speaker 1 (F1).
5) Female speaker 2 (F2)
6) Child (4 year old) (C1).
7) Diplophonic speaker (D1).

Diplophonia is a condition in which a person's alternate glottal pulses are more strongly correlated (both in length and amplitude) than adjacent glottal pulses. Thus, it is extremely difficult to detect the pitch of a diplophonic speaker—even under the best of conditions. Fig. 9 shows a section of waveform from the diplophonic speaker. It is hard to detect, even by eye, the correct pitch periods. For diplophonic speakers, many pitch detectors calculate the pitch period as the distance between major peaks, and not the distance between major and minor peaks. As a result, the pitch contour for a diplophonic speaker often exhibits a large amount of pitch period doubling.

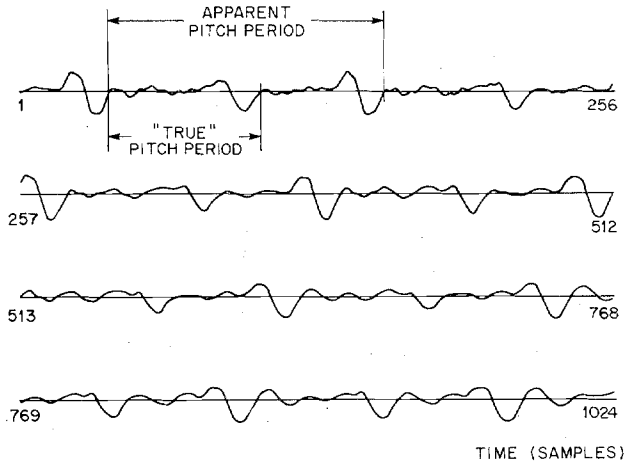


Fig. 9. Section of the waveform from the diplophonic speaker.

To illustrate the range of pitch (both period and frequency) for the speakers in the data base, Fig. 10 shows a plot of the pitch variation for each of the seven speakers for the utterances used in this evaluation (see Section III-B). It can be seen that a wide range of pitch is encompassed by these seven speakers. Additionally, Fig. 11 shows the individual histograms for each of these speakers. It can be seen from this figure that the low-pitched (long period) speakers used in this study (i.e., LM, M1, M2) had a much larger range of pitch period variation than the high-pitched speakers. The histogram for the low-pitched male (LM) shows that on several occasions his pitch period exceeded 200 samples (i.e., the pitch frequency fell below 50 Hz). Since this was outside the anticipated range of pitch variation, all the pitch detectors made errors during these regions.

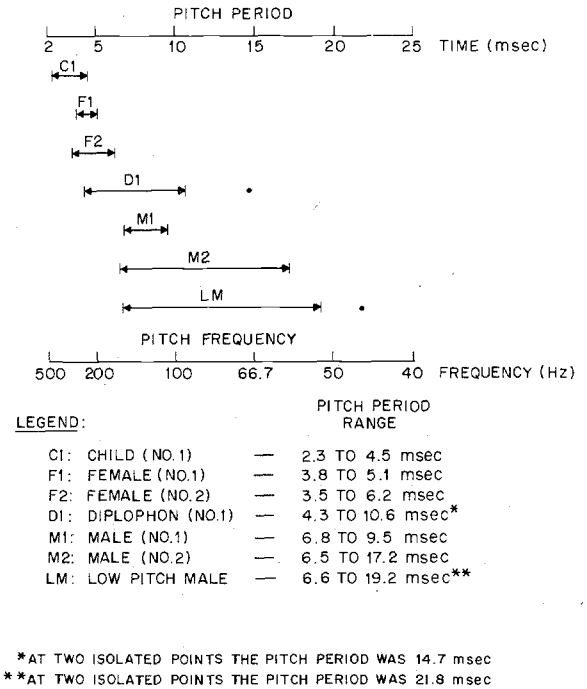


Fig. 10. Pitch variation for each of the speakers used in this study.

B. Recorded Utterances

The utterances used in this study included the four monosyllabic nonsense words:

1) Hayed
2) Heed
3) Hod
4) Hoed

and the four sentences:

5) We were away a year ago.
6) I know when my lawyer is due.
7) Every salt breeze comes from the sea.
8) I was stunned by the beauty of the view.

Sentences 5 and 6 are all voiced (except for the stop gaps) whereas sentences 7 and 8 contain both voiced and unvoiced speech.

C. Recording Conditions

The three types of recording conditions that were used in this study included:

1) Close-talking microphone (M).
2) Standard telephone transmission (T).
3) High-quality microphone (W).

} simultaneous recording

The close-talking microphone recordings were made simultaneously with the telephone recordings since this was the most convenient method of providing a good-quality signal (during voiced regions) for manual pitch detection which could be time aligned with the telephone recordings and which did not interfere with using a standard telephone handset in a natural manner. However, because of its placement close to the mouth of the speaker, the close-talking microphone was quite sensitive to breath noise, plosives, and other unvoiced transients. The telephone recordings were made over the local PBX using an ordinary telephone handset. The close-talking microphone recordings were band-limited to about 3 kHz, as were the telephone recordings. The recordings made on the high-quality microphone were wideband recordings which were filtered at 4 kHz prior to digitization.

IV. Measurement of the Standard Pitch Contour

The method used to measure the standard pitch contour for each of the utterances in the data base was the semiautomatic pitch detector of McGonegal *et al.* [13] which was developed for this study. This method is a highly sophisticated, user-interactive, pitch detector which estimated pitch on a 10-ms frame-by-frame basis. Extensive analysis of the results obtained from this semiautomatic pitch detector across several users on the same utterances showed this method to be highly reliable [13].

Using the semiautomatic method the analysis time for an experienced user was about 30 min to process 1 s of speech (i.e., 100 frames). For the data base used in this study, a total of 60 h of computer processing was required to estimate the standard pitch contours for the entire data base.

V. Error Analysis Results

The way in which objective comparisons of the performance of each of the individual pitch detectors were made was as follows. For each of the utterances in the data base, a standard pitch contour was obtained using the semiautomatic method of Section IV. We denote the standard pitch contour as $p_s(m)$ were m goes from 1 to M, and M is the number of 10-ms frames in the utterance. The contour $p_s(m)$ has the value 0 if the mth frame is unvoiced; otherwise it has the value of the pitch period for the mth frame.

Next, each of the utterances was used as input to each of the seven pitch detectors and a set of pitch contours was obtained as output. We denote the pitch contour from the jth pitch detector ($j = 1, 2, \cdots, 7$) as $\{p_j(m), m = 1, 2, \cdots, M\}$. Of course, special attention had to be given in the Fortran code to compensate the processing delay of each pitch detector to ensure that the pitch contours from each of the seven pitch detectors registered properly with the standard pitch contour.

To quantitatively measure the performance of each of the pitch detectors relative to the semiautomatic analysis, a series of error measurements was defined for each utterance. In Sec-

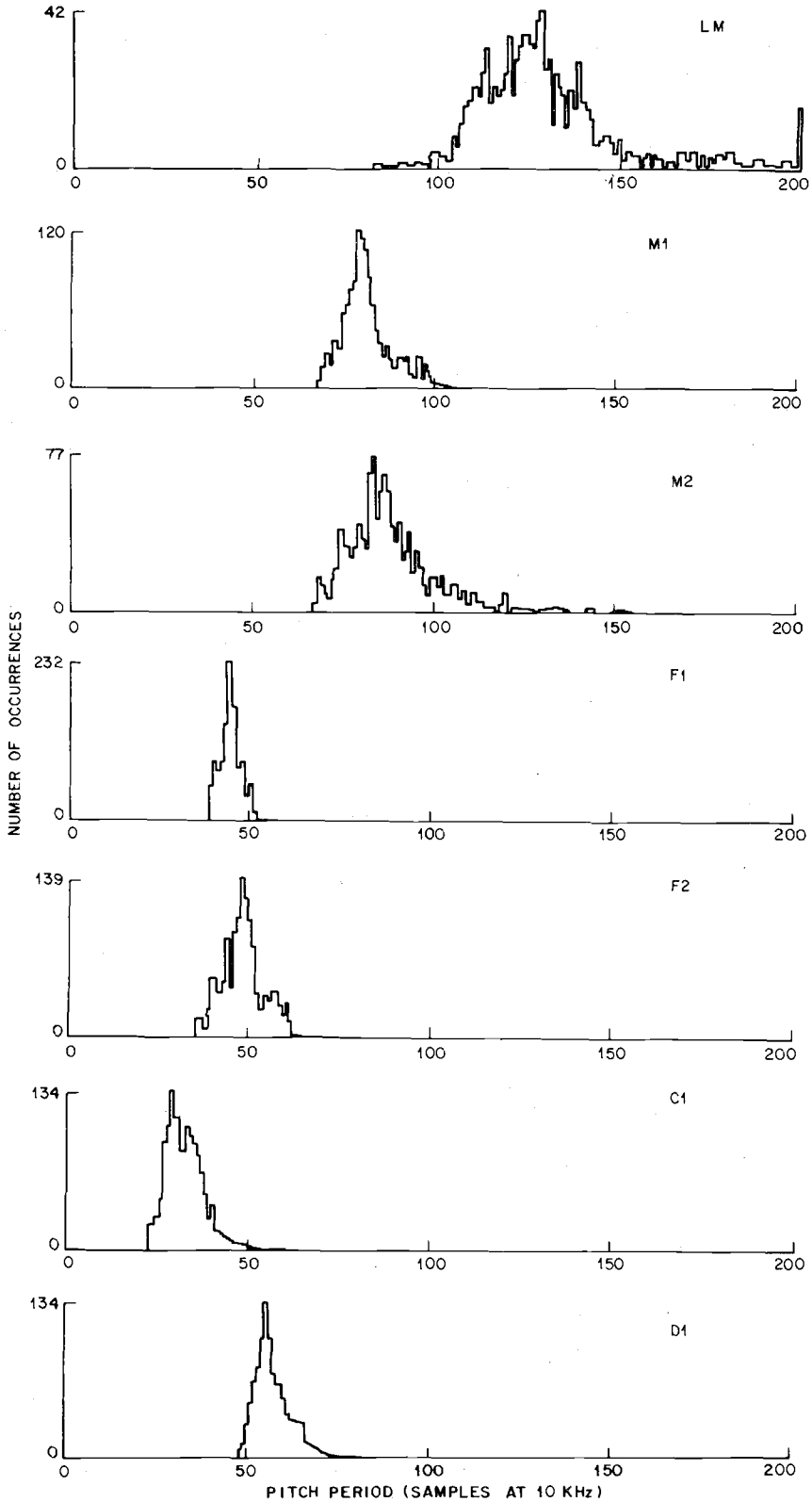


Fig. 11. Individual pitch period histograms for each of the speakers used in this study.

tion V-A we discuss the problems associated with defining these error measurements and attaching physical significance to their values.

The result of the error measurements was a set of scores of the performance of each pitch detector for each utterance, for each recording condition, and for each speaker. Due to the excessive amount of data, the individual results were averaged over the utterances of a single speaker, for each recording condition. A complete set of these results for all the performance classes is given in Section V-B. Finally, where appropriate, the results were averaged over recording conditions and an absolute ranking of each of the pitch detectors for each speaker was

given. Such rankings provided a good picture of the performance strengths and weaknesses of each of the seven pitch detectors.

Before proceeding to the discussion of the error measurements, an additional dimension to the error analysis should be mentioned. This added dimension was the application of nonlinear smoothing (error correcting) methods to detect and correct several types of errors which occur in pitch detection [17]. Such a nonlinear smoother was incorporated into this investigation to see what the effects would be on this data base. Extensive examples of the applications of nonlinear smoothers to speech processing are given in [17].

A. Definition of Error Parameters

As mentioned above, for every utterance in the data base there is a standard pitch contour, $p_s(m)$, and a pitch contour for each pitch detector, $p_j(m)$, where j denotes the pitch detector used in the comparison, i.e., $j=1$ is the AUTOC method, $j=2$ is the CEP method, etc. By comparing $p_s(m)$ to $p_j(m)$, (for each value of j) it can be seen that four possibilities can occur for each value of m. These four possibilities are the following.

1) $p_s(m)=0$, $p_j(m)=0$ in which case both the standard analysis and the pitch detector classified the mth interval as unvoiced. No error results here.

2) $p_s(m)=0$, $p_j(m)\neq 0$ in which case the standard analysis classified the mth interval as unvoiced, but the pitch detector classified the mth interval as voiced. An unvoiced-to-voiced error results here.

3) $p_s(m)\neq 0$, $p_j(m)=0$ in which case the standard analysis classified the mth interval as voiced, but the pitch detector classified the mth interval as unvoiced. A voiced-to-unvoiced error results here.

4) $p_s(m)=P_1\neq 0$, $p_j(m)=P_2\neq 0$ in which case both the standard analysis and the pitch detector classified the mth interval as voiced. For this case two types of errors can exist, depending on the values of P_1 and P_2, the pitch periods from the standard analysis and from the pitch detector. If we define the voiced error $e(m)$ as

$$e(m)=P_1-P_2, \tag{1}$$

then, if $|e(m)|\geq 10$ samples (i.e., more than 1-ms error in estimating the pitch period), the error was classified as a *gross* pitch period error. For such cases, the pitch detector has failed dramatically in estimating the pitch period. Possible causes of such gross pitch errors are pitch period doubling or tripling, inadequate suppression of formants so as to effect pitch measurements, etc. The second type of pitch error was the *fine* pitch period error in which case $|e(m)|<10$ samples. For such cases the pitch detector has estimated the pitch period sufficiently accurately to attribute the errors (primarily) to the measurement techniques.

Based on the above four possibilities for comparing each frame of the reference pitch contour to each frame of each pitch detector contour, five distinct measurements of the performance of each pitch detector were derived. These five error measurements are the following.

1) Gross Error Count: For this measurement the number of gross pitch period errors (as defined above) per utterance was tabulated.

2) Mean of the Fine Pitch Errors: The mean $\bar{e}$ is defined as

$$\bar{e}=\frac{1}{N_i}\sum_{j=1}^{N_i} e(m_j) \tag{2}$$

where m_j is the jth interval in the utterance for which $|e(m)|<10$ (fine pitch error), and N_i is the number of such intervals in the utterance. Thus $\bar{e}$ is a measure of the bias in the pitch measurement during voiced intervals.

3) Standard Deviation of the Fine Pitch Errors: The standard deviation, σ_e, is defined as

$$\sigma_e=\sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i} e^2(m_j)-\bar{e}^2}. \tag{3}$$

The standard deviation of the fine pitch errors is a measure of the accuracy of the pitch detector in measuring pitch period during voiced intervals.

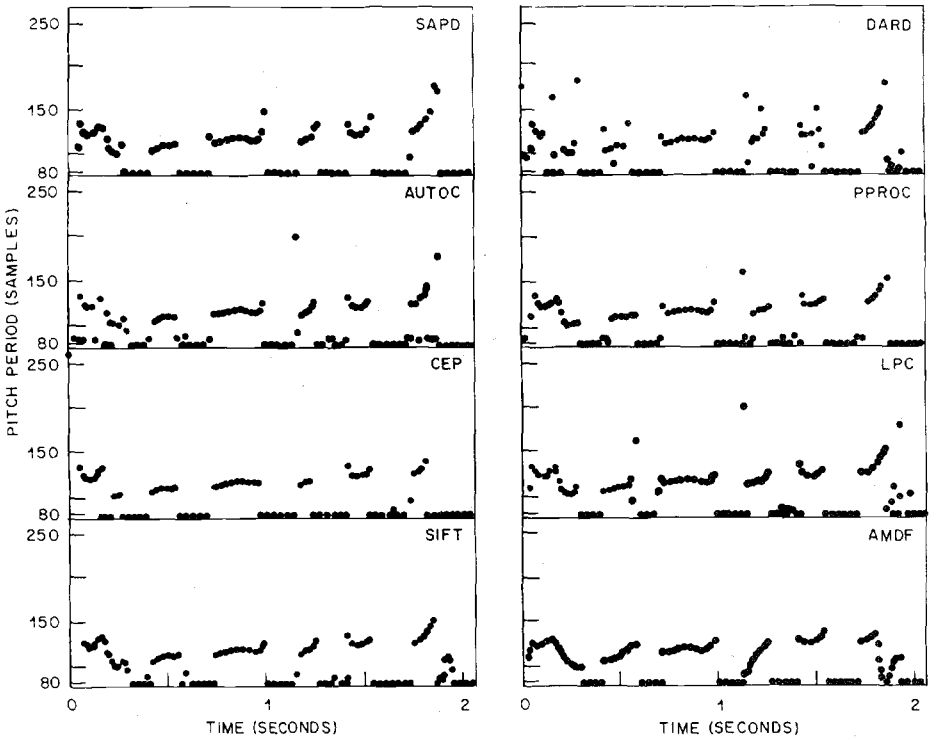
4) Voiced-to-Unvoiced Error Rate: This measurement shows the accuracy in correctly classifying voiced intervals.

5) Unvoiced-to-Voiced Error Rate: This measurement shows the accuracy in correctly classifying unvoiced intervals.

Although other error analyses are possible, it was felt that these five error measurements provided a good description of the performance strengths and weaknesses of each of the seven pitch detectors. The results of these error analyses are given in Section V-B. Before presenting these results, we first show some examples of the individual pitch contours for three of the utterances used in this study. Figs. 12-14 show typical sets of pitch contours for the seventh utterance of Section III-B, for the wideband condition, for speakers LM, M2, and C1. The curve of the upper left in each figure is the result of the semiautomatic analysis (i.e., the standard pitch contour). It can be seen from these figures that each of the types of errors discussed above occurs in these examples. Finally, Fig. 15 shows the result of processing each of the pitch contours of Fig. 13 by a nonlinear smoother (a combination of running medians of length 7 and some simple logic). The overall similarity among the smoothed pitch contours is startlingly evident in Fig. 15. As will be seen later, a good nonlinear smoother (error correcter) is able to correct a large number of the errors in pitch detection and considerably improve the performance of a pitch detector. However, if the error rate is too high, no amount of nonlinear smoothing will suffice.

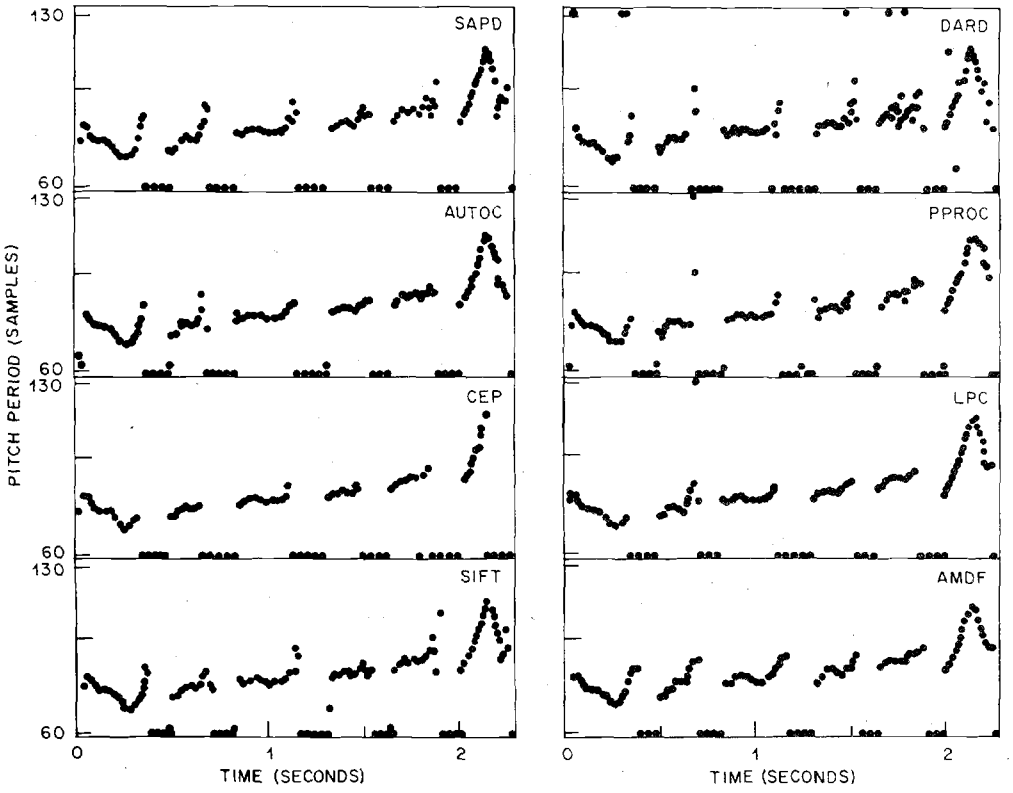
B. Results of Error Analysis

The complete set of error analyses discussed in Section V-A was performed on the entire data base of Section III, and the major results are presented in Tables I–XX, along with corresponding performance scores for each error category. Several points about the analysis should first be noted before discussing the individual tables and the resulting performance scores. First it must be pointed out that for the microphone and telephone recording conditions all eight sentences were processed, whereas for the wideband case only the four sentences were processed (i.e., the four nonsense words were not used).



UTTERANCE: LM07W, UNSMOOTHED

Fig. 12. Representative set of raw pitch contours for utterance 7, speaker LM, recording condition W.



UTTERANCE: M207W, UNSMOOTHED

Fig. 13. Representative set of raw pitch contours for utterance 7, speaker M2, condition W.

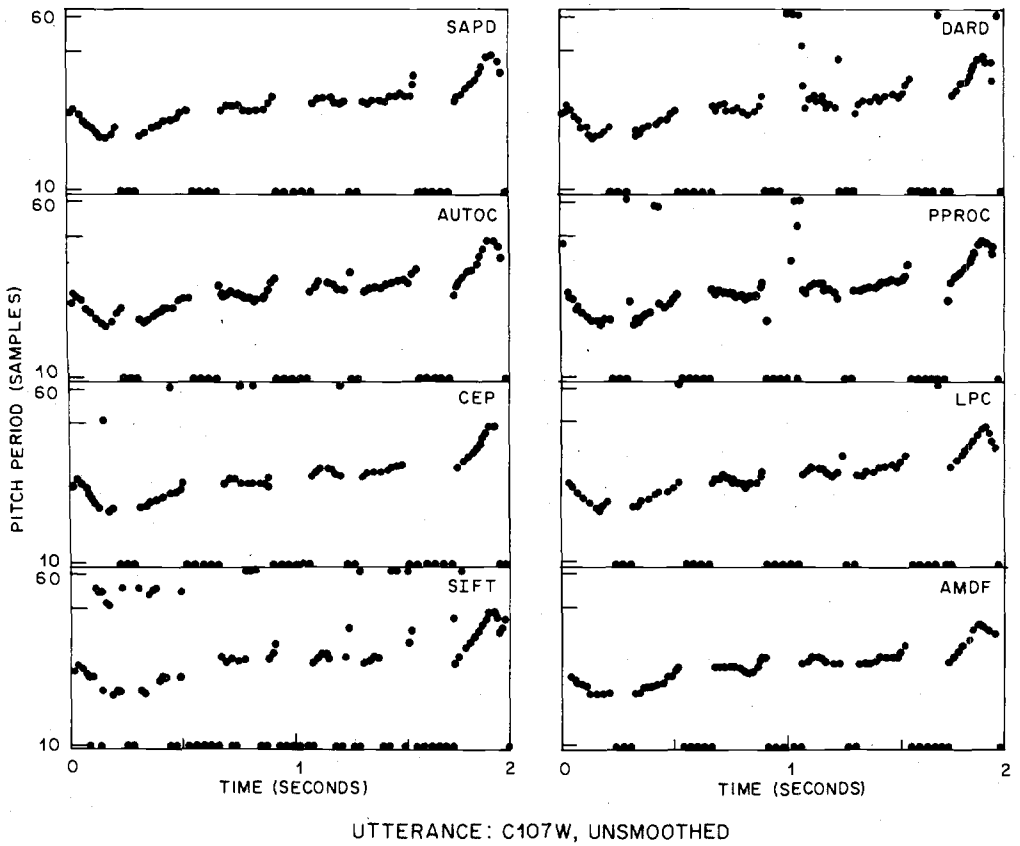


Fig. 14. Representative set of raw pitch contours for utterance 7, speaker C1, condition W.

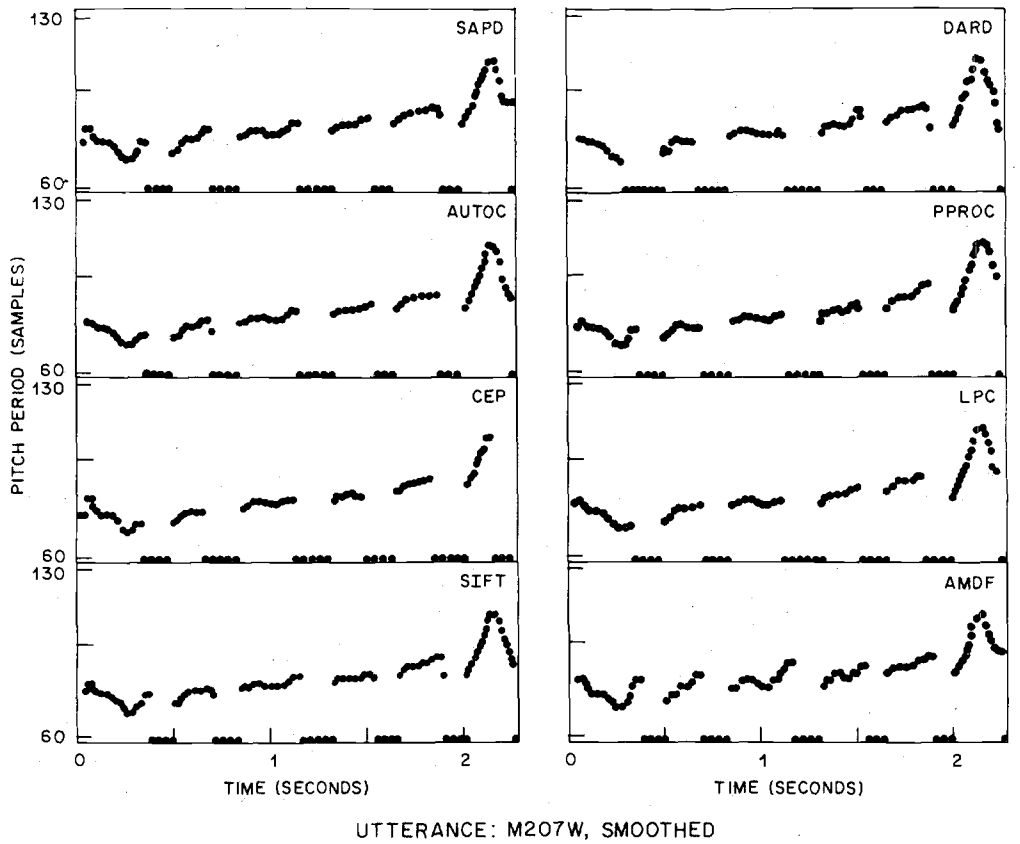


Fig. 15. Representative set of nonlinearly smoothed pitch contours for utterance 7, speaker M2, condition W.

Second, the results obtained for the diplophonic speaker (D1) were omitted entirely from the error analysis because of the universal difficulties of all the pitch detectors (including the semiautomatic method) in estimating the correct pitch period for this speaker. Some of the raw analysis results for speaker D1 are presented in the M.S. thesis of Cheng [18]. Finally, because of the large number of factors involved in the analysis, each of the error measurements was averaged over the utterances of each speaker. This is justified in that it is not anticipated that the sentence material is a factor in the performance evaluation of any pitch detector.

The format for the results presented in Tables I-XX is as follows. First we present the average (over utterances) error scores for each category for each speaker, recording condition, and pitch detector. Also included in the tables is a sum of the raw averages across the three recording conditions. Each table of raw data is followed by one (or sometimes two) table(s) of performance rankings based on an empirical (but hopefully physically justifiable) evaluation of the average scores for each pitch detector and for each speaker. From these performance rankings for each error category the performance strengths and weaknesses of each pitch detector can readily be seen and evaluated. Following the tables of unsmoothed raw averages, the results obtained after nonlinear smoothing are presented. Comparisons between the unsmoothed and smoothed performance rankings show cases where the error rate is too high to be properly corrected with simple nonlinear smoothing techniques. We now proceed to discuss the results for each of the five error categories of Section V-A.

1) Gross Pitch Errors: Tables I-IV present the results obtained for the gross pitch error measurements. From Table I it can be seen that, for the most part, a great deal of homogeneity existed between the scores for the three recording conditions, although in some cases there were fairly substantial differences in the average gross error scores. Table II shows the performance rankings based on the sum of the average gross error scores across the three recording conditions. The best rankings are the lowest scores in Table II, i.e., 1 is the best score, 5 is the worst score. Rank 1 was given to a score of from 0 to 6; rank 2 for a score from 6 to 18; rank 3 for a score from 18 to 42; rank 4 for a score from 42 to 90; and rank 5 for scores over 90. The scale in this case is logarithmic because the difficulty in detecting and correcting such gross errors inherently appears to be logarithmically related to the number of such errors per utterance. Based on these assumptions, the rankings of Table II show that each pitch detector performed better for some speakers (i.e., range of pitch variation) than for others. For example, the AUTOC pitch detector performed much worse on the two low-pitch speakers (LM and M2) than on the three higher pitch speakers (F1, F2, C1); whereas the CEP pitch detector performed much better on the lower pitch speakers than on the higher pitch speakers. An overall ranking score for each pitch detector (i.e., the sum of the rankings over the speakers) is given at the bottom of Table II, and the ranking scores in the rightmost column of Table II (the sum of the rankings over the pitch detector) is a measure of the difficulty of detecting pitch for a given speaker. Table II shows that the overall ranking scores for five of the seven pitch detectors were comparable, and that the two others were somewhat inferior for this error category. Additionally, it is seen that the speakers with the most extreme pitch (LM, C1) presented the most difficulty in terms of this error category.

TABLE I
NUMBER OF GROSS PITCH ERRORS—UNSMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	15.3	0.5	0.6	5.8	10.0	4.4	12.8
	T	26.1	1.1	4.5	5.8	11.0	5.6	15.6
	W	19.5	1.3	4.5	13.8	23.8	13.0	23.8
	Sum	60.9	2.9	9.6	25.4	44.8	23.0	52.2
M1	M	0.6	0.1	0.0	5.9	2.0	0.1	0.3
	T	3.4	0.1	0.8	6.3	3.0	0.8	0.8
	W	2.8	0.5	3.0	23.5	6.0	0.8	2.8
	Sum	6.8	0.7	3.8	35.7	11.0	1.7	3.9
M2	M	6.1	0.4	1.3	15.9	4.9	2.9	7.3
	T	9.9	0.6	3.4	4.0	5.8	4.0	9.8
	W	7.3	1.3	5.3	26.8	12.3	5.5	8.5
	Sum	23.3	2.3	10.0	46.7	23.0	12.4	25.6
F1	M	1.9	9.1	4.4	7.3	4.0	2.4	0.5
	T	1.6	8.5	1.8	6.3	2.8	1.4	0.0
	W	0.0	29.0	8.0	0.8	4.0	2.0	0.0
	Sum	3.5	46.0	14.2	14.4	10.8	5.8	0.5
F2	M	0.4	1.4	2.1	7.1	2.4	1.6	0.6
	T	0.6	2.0	1.5	5.6	1.5	1.5	1.0
	W	2.0	2.5	3.8	8.5	5.0	2.0	1.8
	Sum	3.0	5.9	7.4	21.2	8.9	5.1	3.4
C1	M	1.0	13.6	65.3	6.1	7.8	8.3	10.6
	T	1.9	14.8	62.6	12.9	9.0	12.3	9.1
	W	0.0	12.5	40.8	3.0	7.3	5.5	6.5
	Sum	2.9	40.9	168.7	22.0	24.1	26.1	26.2

TABLE II
PERFORMANCE SCORES BASED ON SUM OF GROSS PITCH ERRORS—UNSMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	4	1	2	3	4	3	4	21
M1	2	1	1	3	2	1	1	11
M2	3	1	2	4	3	2	3	18
F1	1	4	2	2	2	1	1	13
F2	1	1	2	3	2	1	1	11
C1	1	3	5	3	3	3	3	21
Sum	12	11	14	18	16	11	13	

Code: (0-6) = 1, (6-12) = 2, (12-42) = 3, (42-90) = 4, (90-) = 5.

Tables III and IV present the results for the gross pitch error category after processing by a nonlinear smoother. This type of error is most easily detected and corrected by the nonlinear smoother used in this study as verified by Tables III and IV. It can be seen from Table IV that only 12 out of the 42 pairs in Table IV were not given the best ranking of 1. These 12 represent cases where the gross error rate was too high to be corrected entirely by a nonlinear smoother. The overall ranking scores for the smoothed results showed all seven pitch detectors (with the exception of speaker C1 for pitch detector SIFT) to be essentially identical in their overall performance in this error category.

2) Fine Pitch Error—Average Value: The results of the analysis of the average value of the fine pitch error indicated that all seven pitch detectors yielded average values of $\bar{e}$ on the order of ± 0.5 samples across all utterances, speakers, and re-

TABLE III
NUMBER OF GROSS PITCH ERRORS—SMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	5.3	2.3	2.0	3.5	5.8	3.6	9.4
	T	6.1	1.6	2.0	2.0	6.0	3.1	10.8
	W	3.8	5.0	6.8	5.0	13.3	9.5	17.3
	Sum	15.2	8.9	10.8	10.5	25.1	16.2	37.5
M1	M	0.3	0.4	0.5	0.8	1.3	0.4	0.6
	T	0.0	0.1	0.0	0.5	0.1	0.3	0.5
	W	0.3	0.5	0.5	2.8	1.5	0.8	2.8
	Sum	0.6	1.0	1.0	4.1	2.9	1.5	3.9
M2	M	0.8	1.5	1.5	6.0	2.9	3.5	6.4
	T	1.3	1.8	1.4	2.3	3.0	3.6	7.4
	W	2.3	2.3	3.0	8.3	7.0	5.0	5.3
	Sum	4.4	5.6	5.9	16.6	12.9	12.1	19.1
F1	M	0.0	0.8	0.0	0.1	0.0	0.0	0.0
	T	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Sum	0.0	0.8	0.0	0.2	0.0	0.0	0.0
F2	M	0.0	0.1	0.0	0.4	0.3	0.0	0.1
	T	0.1	0.0	0.1	0.6	0.0	0.3	0.3
	W	0.8	0.0	1.3	0.3	1.3	0.0	0.0
	Sum	0.9	0.1	1.4	1.3	1.6	0.3	0.4
C1	M	0.0	0.0	57.4	0.1	0.0	0.0	0.5
	T	0.0	0.0	55.8	0.3	0.0	0.0	1.8
	W	0.0	0.0	15.8	0.0	0.0	0.0	0.0
	Sum	0.0	0.0	129.0	0.4	0.0	0.0	2.3

TABLE IV
PERFORMANCE SCORES BASED ON SUM OF GROSS PITCH ERRORS—SMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	2	2	2	2	3	2	3	16
M1	1	1	1	1	1	1	1	7
M2	1	1	1	2	2	2	2	11
F1	1	1	1	1	1	1	1	7
F2	1	1	1	1	1	1	1	7
C1	1	1	5	1	1	1	1	11
Sum	7	7	11	8	9	8	9	

Code same as Table II.

cording conditions. No consistent bias (either positive or negative) in the value of $\bar{e}$ was noted in the data. Thus for all practical purposes the average value of the fine pitch error was essentially 0 in all cases and, therefore, no results are tabulated here.

3) Fine Pitch Error–Standard Deviation: Tables V-VIII present the results of the analysis of the standard deviation of the fine pitch error. The units of the standard deviation are samples. The results here were quite homogeneous across recording conditions and thus the sum of the standard deviations over recording conditions was used as the performance measure in Tables VI (raw averages) and VIII (smoothed averages). Based on the analysis results, a standard deviation of less than 0.5 samples per condition, or 1.5 samples for the sum was given a score of 1. A linear scale was used for this measurement –thus a standard deviation sum from 1.5 to 3 samples was given the next best score (2), etc.

As seen in Table VI, four of the pitch detectors (AUTOC, CEP, SIFT, and LPC) performed almost uniformly across all speakers, and had comparably high overall performance scores.

TABLE V
STANDARD DEVIATION OF FINE PITCH ERRORS—UNSMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	1.0	0.9	0.9	1.1	1.8	1.1	2.3
	T	1.0	1.0	0.9	1.2	1.9	1.2	2.1
	W	0.8	1.0	0.9	1.1	1.6	1.1	2.1
	Sum	2.8	2.9	2.7	3.4	5.3	3.4	6.5
M1	M	0.7	0.5	0.7	0.9	0.9	0.6	1.1
	T	0.6	0.5	0.7	0.8	0.6	0.6	1.0
	W	1.0	0.8	0.8	1.3	1.1	0.7	1.3
	Sum	2.3	1.8	2.2	3.0	2.6	1.9	3.4
M2	M	0.7	0.7	0.7	1.0	1.1	0.8	1.5
	T	0.6	0.7	0.8	1.0	1.0	0.9	1.3
	W	1.0	0.9	1.0	1.2	1.4	0.9	1.6
	Sum	2.3	2.3	2.5	3.2	3.5	2.6	4.4
F1	M	0.6	0.5	0.7	1.1	0.8	0.6	1.0
	T	0.6	0.6	0.7	1.0	0.9	0.6	1.1
	W	0.5	0.5	0.7	0.9	0.8	0.5	0.9
	Sum	1.7	1.6	2.1	3.0	2.5	1.7	3.0
F2	M	0.6	0.6	0.6	0.8	0.8	0.4	1.1
	T	0.6	0.5	0.7	0.9	0.8	0.5	1.3
	W	0.6	0.5	0.7	0.8	0.7	0.6	1.1
	Sum	1.8	1.6	2.0	2.5	2.3	1.5	3.5
C1	M	0.4	0.5	0.8	0.7	0.6	0.5	0.9
	T	0.4	0.5	0.8	0.8	0.6	0.5	1.0
	W	0.5	0.4	0.9	0.8	0.7	0.5	1.0
	Sum	1.3	1.4	2.5	2.3	1.9	1.5	2.9

TABLE VI
PERFORMANCE SCORES BASED ON SUM OF STANDARD DEVIATIONS OF FINE PITCH ERRORS—UNSMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	2	2	2	3	4	3	5	21
M1	2	2	2	2	2	2	3	15
M2	2	2	2	3	3	2	3	17
F1	2	2	2	3	2	2	2	15
F2	2	2	2	2	2	1	3	14
C1	1	1	2	2	2	1	2	11
Sum	11	11	12	15	15	11	18	

Code: (0–1.5) = 1, (1.5–3) = 2, (3–4.5) = 3, (4.5–6) = 4, (6–) = 5.

The two simple time-domain pitch detectors (DARD and PPROC) had somewhat higher scores (poorer performance) due to the lower resolution which is obtained in estimating a pitch period directly on the waveform due to effects discussed earlier. Finally, the worst performance in this category was for the AMDF pitch detector. This result is due to the lack of resolution in the AMDF measurement which is made only every third or fourth sample—thus the pitch period is only estimated to within a couple of samples.

Tables VII and VIII for the smoothed standard deviations show that the nonlinear smoother does not strongly affect the raw results presented in Tables V and VI. Slight differences in the overall performance scores do exist both because of the gross pitch period errors which are detected and corrected to fine pitch period errors, and because of the smoothing of the fine pitch errors themselves.

4) Voiced-to-Unvoiced Errors: Tables IX-XIV present the results of the voiced-to-unvoiced errors for each pitch detector. Table IX gives the raw average scores for each recording condi-

TABLE VII
STANDARD DEVIATION OF FINE PITCH ERRORS—SMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	1.1	1.1	1.0	1.2	1.5	1.1	1.9
	T	1.0	1.2	1.0	1.3	1.6	1.3	1.9
	W	1.2	1.3	0.9	1.2	1.5	1.0	2.0
	Sum	3.3	3.6	2.9	3.7	4.6	3.4	5.8
M1	M	0.6	0.6	0.6	0.8	0.7	0.5	1.2
	T	0.6	0.6	0.6	0.7	0.6	0.5	1.0
	W	0.8	0.9	0.9	1.1	0.9	0.7	1.3
	Sum	2.0	2.1	2.1	2.6	2.2	1.7	3.5
M2	M	0.8	0.8	0.8	1.0	1.0	0.8	1.5
	T	0.8	0.9	0.9	0.9	0.8	0.9	1.6
	W	0.9	1.1	1.1	1.2	1.3	0.8	1.5
	Sum	2.5	2.8	2.8	3.1	3.1	2.5	4.6
F1	M	0.5	0.5	0.6	0.8	0.6	0.5	0.9
	T	0.5	0.6	0.6	0.9	0.7	0.5	1.0
	W	0.5	0.5	0.8	0.7	0.6	0.4	0.8
	Sum	1.5	1.6	2.0	2.4	1.9	1.4	2.7
F2	M	0.6	0.6	0.6	0.9	0.7	0.4	1.0
	T	0.5	0.5	0.6	0.8	0.7	0.5	1.2
	W	0.6	0.5	0.6	0.7	0.6	0.5	1.0
	Sum	1.7	1.6	1.8	2.4	2.0	1.4	3.2
C1	M	0.4	0.6	1.1	0.6	0.6	0.5	0.9
	T	0.5	0.5	0.7	0.7	0.5	0.6	1.1
	W	0.5	0.5	0.9	0.6	0.7	0.4	0.8
	Sum	1.4	1.6	2.7	1.9	1.8	1.5	2.8

TABLE VIII
PERFORMANCE SCORES BASED ON SUM OF STANDARD DEVIATIONS OF FINE PITCH ERRORS—SMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	3	3	2	3	4	3	4	22
M1	2	2	2	2	2	2	3	15
M2	2	2	2	3	3	2	4	18
F1	1	2	2	2	2	1	2	12
F2	2	2	2	2	2	1	3	14
C1	1	2	2	2	2	1	2	12
Sum	11	13	12	14	15	10	18	

Code same as Table VI.

tion as well as the sum of the scores across recording conditions. Each of the scores is given as a ratio of the number of voiced-to-unvoiced errors to the number of voiced intervals for each condition. As might be anticipated, there is a great lack of homogeneity of the results across recording conditions, especially for the LPC pitch detector.

Table X gives a performance evaluation of the pitch detectors for the raw data of the voiced-to-unvoiced error rate averaged over the three recording conditions. The scores at the top of this table are the percentage of voiced-to-unvoiced errors for each pitch detector. A ranking of 1 was given to a pitch detector with an average error rate less than 5 (percent). A linear scale was used for these performance scores as shown in Table X.

Based on the overall rankings, it can be seen that five of the pitch detectors (AUTOC, DARD, PPROC, LPC, and AMDF) had essentially equivalent performance scores and all tended to be homogeneous across speakers. The SIFT pitch detector had a somewhat poorer performance than the top five, and the CEP pitch detector had a poor performance for this error category. We defer a discussion of these results to Section VI.

Because of the lack of homogeneity across recording conditions, a second set of performance ratings was made for this error category based solely on the wideband recordings. These results are presented in Table XI. From this table it can be seen that four of the pitch detectors (AUTOC, PPROC, LPC, and AMDF) performed extremely well on this condition. The SIFT and DARD methods had somewhat poorer performance scores, while the CEP method had the worst score.

Tables XII–XIV show the error scores and performance rankings for voiced-to-unvoiced errors for the smoothed pitch contours. The effect of the smoother is to change slightly the number of voiced-to-unvoiced errors. The performance rankings for the data averaged over recording conditions (Table XIII) shows slightly different results than for the raw data; however, the rankings for the wideband condition (Table XIV) are quite similar to the raw data rankings of Table XI.

The results of Tables IX–XIV also show that the most difficult speakers were the two low-pitched speakers (LM, M2) and the high-pitched speaker (C1).

5) Unvoiced-to-Voiced Errors: The last set of tables (Tables XV–XX) show the results of the unvoiced-to-voiced error analysis. The form of the data in these tables is identical to that used in the voiced-to-unvoiced error category. A performance ranking of 1 was given to an unvoiced-to-voiced error rate of less than 10 percent. The remaining ranking scores were assigned linearly as shown in Table XVI. The overall performance scores for the raw data averaged across recording conditions showed the CEP pitch detector to have a very low score (high performance), in contrast to the very high scores it obtained in the previous error category. The AUTOC, SIFT, DARD, PPROC, and AMDF pitch detectors all had similar performance rankings and the LPC pitch detector had a very poor score. (Again we defer discussion of these results to Section VI.)

Table XVII (for the raw wideband data only) shows the performance of the LPC pitch detector to be substantially improved and comparable to all but the CEP pitch detector.

As seen in Tables XVIII–XX, the nonlinear smoother substantially helps almost all the pitch detectors for the unvoiced-to-voiced error category. The performance rankings for all but the LPC pitch detector are almost comparable for the smoothed data averaged over recording conditions (Table XIX); for the wideband smoothed data (Table XX) all seven pitch detectors had comparable performance scores.

VI. DISCUSSION OF ERROR ANALYSIS RESULTS

The error analysis and performance evaluation presented in Section V points up the strengths and weaknesses of each of the pitch detectors used in the study. No single pitch detector was uniformly top ranked across all speakers, recording conditions, and error measurements. In this section we discuss the results presented in Section V with a view towards explaining the general trends in the performance scores and how they relate back to the specific methods of pitch detection used in this study.

TABLE IX
VOICED-TO-UNVOICED ERRORS—UNSMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	32/631	168/631	58/631	66/631	16/631	1/631	27/631
	T	36/631	235/631	105/631	66/631	37/631	78/631	40/631
	W	33/533	130/533	46/533	77/533	18/533	4/533	15/533
	Sum	101/1795	533/1795	209/1795	209/1795	71/1795	83/1795	82/1795
M1	M	19/703	54/703	11/703	30/703	25/703	3/703	28/703
	T	45/703	75/703	37/703	75/703	51/703	36/703	57/703
	W	6/654	88/654	7/654	39/654	14/654	14/654	5/654
	Sum	70/2060	217/2060	55/2060	144/2060	90/2060	53/2060	90/2060
M2	M	48/772	89/772	38/772	65/772	28/772	1/772	40/772
	T	60/772	123/772	60/772	104/772	67/772	194/772	67/772
	W	27/660	123/660	12/660	37/660	15/660	26/660	16/660
	Sum	135/2204	335/2204	110/2204	196/2204	110/2204	221/2204	113/2204
F1	M	10/762	99/762	45/762	15/762	18/762	6/762	21/762
	T	38/762	97/762	42/762	40/762	45/762	148/762	26/762
	W	7/603	70/603	28/603	18/603	14/603	1/603	17/603
	Sum	55/2127	266/2127	115/2127	73/2127	77/2127	155/2127	64/2127
F2	M	18/810	62/810	36/810	14/810	17/810	3/810	23/810
	T	46/810	67/810	37/810	32/810	41/810	68/810	36/810
	W	16/670	68/670	30/670	49/670	33/670	12/670	30/670
	Sum	80/2290	197/2290	103/2290	95/2290	91/2290	83/2290	89/2290
C1	M	38/935	93/935	130/935	27/935	20/935	5/935	21/935
	T	68/935	100/935	137/935	58/935	52/935	43/935	52/935
	W	9/568	66/568	139/568	18/568	12/568	5/568	13/568
	Sum	115/2438	259/2438	406/2438	103/2438	84/2438	53/2438	86/2438

TABLE X
PERFORMANCE SCORES BASED ON SUM OF VOICED-TO-UNVOICED ERRORS—UNSMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	5.6	29.7	11.6	11.6	4.0	4.6	4.6
M1	3.4	10.5	2.7	7.0	4.4	2.6	4.4
M2	6.1	15.2	5.0	8.9	5.0	10.0	5.1
F1	2.6	12.5	5.4	3.4	3.6	7.3	3.0
F2	3.5	8.6	4.5	4.1	4.0	3.6	3.9
C1	4.7	10.6	16.7	4.2	3.4	2.2	3.5

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	2	5	3	3	1	1	1	16
M1	1	3	1	2	1	1	1	10
M2	2	4	2	2	2	3	2	17
F1	1	3	2	1	1	2	1	11
F2	1	2	1	1	1	1	1	8
C1	1	3	4	1	1	1	1	11
Sum	8	20	13	10	7	9	7	

(b) Performance Scores

Code: (0–5) = 1, (5–10) = 2, (10–15) = 3, (15–20) = 4, (20–) = 5.

TABLE XI
PERFORMANCE SCORES BASED ON VOICED-TO-UNVOICED ERRORS—WIDEBAND DATA—UNSMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	6.2	24.4	8.6	14.4	3.4	0.8	2.8
M1	0.9	13.5	1.1	6.0	2.1	2.1	0.8
M2	4.1	18.6	1.8	5.6	2.3	3.9	2.4
F1	1.2	11.6	4.6	3.0	2.3	0.2	2.8
F2	2.4	10.1	4.5	7.3	4.9	1.8	4.5
C1	1.6	11.6	24.5	3.2	2.1	0.9	2.3

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	2	5	2	3	1	1	1	15
M1	1	3	1	2	1	1	1	10
M2	1	4	1	2	1	1	1	11
F1	1	3	1	1	1	1	1	9
F2	1	3	1	2	1	1	1	10
C1	1	3	5	1	1	1	1	13
Sum	7	21	11	11	6	6	6	

(b) Performance Scores

Code same as Table X.

The results on the gross pitch period errors (Tables I–IV) showed that the time-domain and hybrid pitch detectors had greatest difficulty with the low-pitched speakers (LM, M2) whereas the spectral pitch detector (CEP) had the greatest difficulty with the high-pitched speakers (C1, F1). The difficulties of time-domain methods for low-pitched speakers are due to the fixed 30–40-ms analysis frame which is generally inadequate for low-pitched speakers. The difficulties of spectral methods for high-pitched speakers are due to the small number of harmonics which are present in their spectra, leading to analysis difficulties in choosing the correct pitch. The poor performance of the SIFT pitch detector on speaker C1 is related to the problem of reliably spectrally flattening (by inverse filtering) a signal in which generally only one harmonic occurs.

TABLE XII
VOICED-TO-UNVOICED ERRORS—SMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	112/626	137/626	48/626	56/626	31/626	6/626	34/626
	T	213/626	226/626	135/626	60/626	60/626	79/626	85/626
	W	69/512	97/512	28/512	61/512	18/512	6/512	11/512
	Sum	404/1764	460/1764	211/1764	117/1764	109/1764	91/1764	130/1764
M1	M	20/706	43/706	8/706	33/706	26/706	0/706	24/706
	T	64/706	65/706	42/706	91/706	67/706	25/706	60/706
	W	9/657	74/657	7/657	44/657	17/657	14/657	5/657
	Sum	93/2069	182/2069	57/2069	168/2069	110/2069	39/2069	89/2069
M2	M	90/782	80/782	47/782	86/782	45/782	0/782	46/782
	T	134/782	104/782	86/782	114/782	89/782	213/782	85/782
	W	32/660	116/660	9/660	43/660	25/660	15/660	16/660
	Sum	256/2224	300/2224	142/2224	243/2224	159/2224	228/2224	147/2224
F1	M	7/769	119/769	50/769	32/769	16/769	2/769	19/769
	T	40/769	92/769	34/769	53/769	45/769	157/769	22/769
	W	7/607	81/607	22/607	24/607	18/607	2/607	21/607
	Sum	54/2145	292/2145	106/2145	109/2145	79/2145	161/2145	62/2145
F2	M	13/815	61/815	29/815	17/815	8/815	4/815	32/815
	T	44/815	70/815	34/815	43/815	46/815	47/815	48/815
	W	15/676	72/676	28/676	83/676	38/676	12/676	32/676
	Sum	72/2306	203/2306	91/2306	143/2306	92/2306	63/2306	112/2306
C1	M	40/941	99/941	198/941	44/941	31/941	31/941	86/941
	T	70/941	107/941	175/941	107/941	65/941	85/941	97/941
	W	8/600	70/600	230/600	26/600	15/600	5/600	33/600
	Sum	118/2482	276/2482	603/2482	177/2482	111/2482	121/2482	216/2482

TABLE XIII
PERFORMANCE SCORES BASED ON SUM OF VOICED-TO-UNVOICED ERRORS—SMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	22.9	26.1	12.0	10.0	6.2	5.2	7.4
M1	4.5	8.8	2.8	8.1	5.3	1.9	4.3
M2	11.5	13.5	6.4	10.9	7.1	10.3	6.6
F1	2.5	13.6	4.9	5.1	3.7	7.5	2.9
F2	3.1	8.8	3.9	6.2	4.0	2.7	4.9
C1	4.8	11.1	24.3	7.1	4.5	4.9	8.7

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	5	5	3	3	2	2	2	22
M1	1	2	1	2	2	1	1	10
M2	3	3	2	3	2	3	2	18
F1	1	3	1	2	1	2	1	11
F2	1	2	1	2	1	1	1	9
C1	1	3	5	2	1	1	2	15
Sum	12	18	13	14	9	10	9	

(b) Performance Scores

Code same as Table X.

TABLE XIV
PERFORMANCE SCORES BASED ON VOICED-TO-UNVOICED ERRORS—WIDEBAND DATA—SMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	13.5	18.9	5.5	11.9	3.5	1.2	2.1
M1	1.4	11.3	1.1	6.7	2.6	2.1	0.8
M2	4.8	17.6	1.4	6.5	3.8	2.3	2.4
F1	1.2	13.3	3.6	4.0	3.0	0.3	3.5
F2	2.2	10.7	4.1	12.3	5.6	1.8	4.7
C1	1.3	11.7	38.3	4.3	2.5	0.8	5.5

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	3	4	2	3	1	1	1	15
M1	1	3	1	2	1	1	1	10
M2	1	4	1	2	1	1	1	11
F1	1	3	1	1	1	1	1	9
F2	1	3	1	3	2	1	1	12
C1	1	3	5	1	1	1	2	14
Sum	8	20	11	12	7	6	7	

(b) Performance Scores

Code same as Table X.

The results on the fine pitch period errors (Tables V–VIII) showed that (aside from the AMDF method which inherently lacked pitch resolution) the time-domain waveform pitch detectors (DARD, PPROC) had somewhat lower resolution than the other methods. This is due to the sensitivity of waveform peaks, valleys, and zero crossings to formant changes, noise, distortion, etc.

The error measurements of voiced-to-unvoiced and unvoiced-to-voiced errors provided several interesting results. These categories *cannot* be examined separately because they are often intimately related. For example, a voiced–unvoiced detector which is biased towards the category voiced will generally have a low voiced-to-unvoiced error rate, but in compensation will have a high unvoiced-to-voiced error rate. There are three types of voiced–unvoiced decision methods used in the seven pitch detectors. One method is the use of a

TABLE XV
UNVOICED-TO-VOICED ERRORS—UNSMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	44/277	16/277	45/277	24/277	52/277	165/277	48/277
	T	35/277	5/277	46/277	25/277	71/277	133/277	69/277
	W	32/180	3/180	33/180	25/180	29/180	46/180	31/180
	Sum	111/734	24/734	124/734	74/734	152/734	344/734	148/734
M1	M	32/292	5/292	52/292	27/292	25/292	73/292	35/292
	T	26/292	5/292	52/292	36/292	74/292	96/292	68/292
	W	14/132	9/132	22/132	9/132	16/132	11/132	18/132
	Sum	72/716	19/716	126/716	72/716	115/716	180/716	121/716
M2	M	42/324	19/324	55/324	88/324	65/324	226/324	43/324
	T	39/324	13/324	59/324	11/324	52/324	141/324	38/324
	W	20/128	4/128	30/128	19/128	26/128	22/128	24/128
	Sum	101/772	36/772	144/772	118/772	143/772	389/772	105/772
F1	M	40/219	9/219	50/219	56/219	48/219	68/219	29/219
	T	30/219	8/219	47/219	25/219	43/219	51/219	29/219
	W	21/125	5/125	31/125	8/125	13/125	15/125	6/125
	Sum	91/563	22/563	128/563	89/563	104/563	134/563	64/563
F2	M	86/400	20/400	91/400	147/400	128/400	254/400	126/400
	T	51/400	23/400	84/400	71/400	107/400	188/400	89/400
	W	16/160	11/160	27/160	3/160	9/160	15/160	13/160
	Sum	153/960	54/960	202/960	221/960	244/960	457/960	228/960
C1	M	43/312	7/312	73/312	29/312	51/312	89/312	37/312
	T	35/312	10/312	57/312	43/312	74/312	102/312	43/312
	W	15/132	7/132	21/132	17/132	15/132	11/132	9/132
	Sum	93/756	24/756	151/756	89/756	140/756	202/756	89/756

TABLE XVI
PERFORMANCE SCORES BASED ON SUM OF UNVOICED-TO-VOICED ERRORS—UNSMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	15.1	3.3	16.9	10.0	20.7	46.9	20.2
M1	10.1	2.7	17.6	10.1	16.1	25.1	16.9
M2	13.1	4.7	18.7	15.3	18.5	50.4	13.6
F1	16.2	3.9	22.7	15.8	18.5	23.8	11.4
F2	15.9	5.6	21.0	23.0	25.4	47.6	23.8
C1	12.3	3.2	20.0	11.8	18.5	26.7	11.8

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DART	PPROC	LPC	AMDF	Sum
LM	2	1	2	2	3	5	3	18
M1	2	1	2	2	2	3	2	14
M2	2	1	2	2	2	5	2	16
F1	2	1	3	2	2	3	2	15
F2	2	1	3	3	3	5	3	20
C1	2	1	3	2	2	3	2	15
Sum	12	6	15	13	14	24	14	

(b) Performance Scores

Code: (0-10) = 1, (10-20) = 2, (20-30) = 3, (30-40) = 4, (40-) = 5.

TABLE XVII
PERFORMANCE SCORES BASED ON UNVOICED-TO-VOICED ERRORS—WIDEBAND DATA—UNSMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	17.8	1.7	18.3	13.9	16.1	25.6	17.2
M1	10.6	6.8	16.7	6.8	12.1	8.3	13.6
M2	15.6	3.1	23.4	14.8	20.3	17.2	18.8
F1	16.8	4.0	24.8	6.4	10.4	12.0	4.8
F2	10.0	6.9	16.9	1.9	5.6	9.4	8.1
C1	11.4	5.3	15.9	12.9	11.4	8.3	6.8

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	2	1	2	2	2	3	2	14
M1	2	1	2	1	2	1	2	11
M2	2	1	3	2	3	2	2	15
F1	2	1	3	1	2	2	1	12
F2	2	1	2	1	1	1	1	9
C1	2	1	2	2	2	1	1	11
Sum	12	6	14	9	12	10	9	

(b) Performance Scores

Code same as Table XVI.

simple threshold on one or more measurements to classify an interval as voiced or unvoiced. For example, the preliminary voiced-unvoiced detector used in the AUTOC, CEP, SIFT, and PPROC methods used a waveform threshold to remove intervals of silence. The second type of voiced-unvoiced detector is the periodicity measurement. For example, the AUTOC, AMDF, and SIFT methods used a threshold on the autocorrelation peak to decide if the interval was periodic whereas the CEP method used a threshold on the cepstral peak for this purpose. The third type of voiced-unvoiced detector is the pattern recognition statistical approach used in the LPC pitch detector. Each of these methods has some advantages and disadvantages. For example, the periodicity measurement tends to be extremely robust with regard to noise, distortion, and spurious transients in the signal. Thus methods like the AUTOC and AMDF pitch detectors tended to work uni-

TABLE XVIII
UNVOICED-TO-VOICED ERRORS—SMOOTHED

Speaker		Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	M	5/282	3/282	6/282	18/282	24/282	95/282	51/282
	T	3/282	0/282	9/282	11/282	25/282	59/282	64/282
	W	3/201	2/201	17/201	11/201	20/201	43/201	35/201
	Sum	11/765	5/765	32/765	40/765	69/765	197/765	150/765
M1	M	19/289	3/289	45/289	17/289	9/289	55/289	35/289
	T	3/289	5/289	33/289	5/289	29/289	65/289	54/289
	W	8/129	7/129	20/129	5/129	3/129	9/129	15/129
	Sum	30/707	15/707	98/707	27/707	41/707	129/707	104/707
M2	M	5/314	13/314	18/314	45/314	21/314	225/314	31/314
	T	4/314	8/314	5/314	2/314	13/314	131/314	37/314
	W	17/128	4/128	20/128	9/128	16/128	20/128	24/128
	Sum	26/756	25/756	43/756	56/756	50/756	376/756	92/756
F1	M	31/212	5/212	34/212	3/212	21/212	33/212	15/212
	T	19/212	1/212	44/212	2/212	19/212	19/212	18/212
	W	17/121	5/121	20/121	0/121	6/121	11/121	6/121
	Sum	67/545	11/545	98/545	5/545	46/545	63/545	39/545
F2	M	34/395	5/395	52/395	42/395	48/395	82/395	44/395
	T	26/395	11/395	64/395	35/395	55/395	132/395	44/395
	W	8/154	10/154	20/154	0/154	7/154	6/154	9/154
	Sum	68/944	26/944	136/944	77/944	110/944	220/944	97/944
C1	M	16/306	4/306	31/306	4/306	16/306	10/306	13/306
	T	10/306	6/306	18/306	3/306	15/306	3/306	9/306
	W	13/130	5/130	10/130	0/130	7/130	10/130	7/130
	Sum	39/742	15/742	59/742	7/742	38/742	23/742	29/742

TABLE XIX
PERFORMANCE SCORES BASED ON SUM OF UNVOICED-TO-VOICED ERRORS—SMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	1.4	0.7	4.2	5.2	9.0	25.8	19.6
M1	4.2	2.1	13.9	3.8	5.8	18.2	14.7
M2	3.4	3.3	5.7	7.4	6.6	49.7	12.2
F1	12.3	2.0	18.0	0.9	8.4	11.6	7.2
F2	7.2	2.8	14.4	8.2	11.7	23.3	10.3
C1	5.3	2.0	8.0	0.9	5.1	3.1	3.9

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	1	1	1	1	1	3	2	10
M1	1	1	2	1	1	2	2	10
M2	1	1	1	1	1	5	2	12
F1	2	1	2	1	1	2	1	10
F2	1	1	2	1	2	3	2	12
C1	1	1	1	1	1	1	1	7
Sum	7	6	9	6	7	16	10	

(b) Performance Scores

Code same as Table XVI.

TABLE XX
PERFORMANCE SCORES BASED ON UNVOICED-TO-VOICED ERRORS—WIDEBAND DATA—SMOOTHED

Speaker	Pitch Detector AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF
LM	1.5	1.0	8.5	5.5	10.0	21.4	17.4
M1	6.2	5.4	15.5	3.9	2.3	7.0	11.6
M2	13.3	3.1	15.6	7.0	12.5	15.6	18.8
F1	14.0	4.1	16.5	0.0	5.0	9.1	5.0
F2	5.2	6.5	13.0	0.0	4.5	3.9	5.8
C1	10.0	3.8	7.7	0.0	5.4	7.7	5.4

(a) Percentage Error Rate

Speaker	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Sum
LM	1	1	1	1	2	3	2	11
M1	1	1	2	1	1	1	2	9
M2	2	1	2	1	2	2	2	12
F1	2	1	2	1	1	1	1	9
F2	1	1	2	1	1	1	1	8
C1	2	1	1	1	1	1	1	8
Sum	9	6	10	6	8	9	9	

(b) Performance Scores

Code same as Table XVI.

formly well across recording conditions, whereas a method like the LPC pitch detector, which used a pattern recognition voiced-unvoiced detector, worked much better for wideband recordings than for microphone or telephone recordings. The distortions (especially high-level transients) and band-limiting in both the microphone and telephone recordings made reliable voiced-unvoiced decisions almost impossible for the pattern recognition approach (of the LPC method) using the five parameters discussed in [15]. However, for the wideband recordings, this method worked quite well.

The only method which had no formal voiced-unvoiced detector was the DARD method. This method just identified pitch period markers directly on the speech waveform. The method used to classify an interval as voiced was to measure

the spacing between adjacent markers centered around the interval and to call the interval as unvoiced if the marker spacing exceeded 200 samples (20-ms period). This method provided surprisingly good results yielding a reasonable voiced-unvoiced error rate.

Finally, it can be seen that the CEP pitch detector had a strong tendency to classify voiced intervals as unvoiced. In compensation the unvoiced-to-voiced error rate for the CEP method was very low. Readjustment of the cepstral peak threshold and the following zero-crossing threshold would yield a tradeoff in these scores.

VII. Computational Considerations

Since none of the pitch detectors used in this study are commercially available, another factor in comparing these pitch detectors is their speed of execution on the computer (a Data General NOVA 800 minicomputer[2]) on which all the simulations were run. Table XXI shows such a comparison along with other computational considerations for implementing the various algorithms. The execution times given in the table are the time required to process 1 s of speech. It can be seen that the two waveform time-domain pitch detectors (DARD and PPROC) ran the fastest, whereas all the others were on the order of 1 to 2 orders of magnitude slower. The AMDF pitch detector would take about four times longer if the resolution in the measurement were increased to 1 sample at a 6.67-kHz rate. The AUTOC pitch detector is a factor of 2 or more faster than the SIFT, LPC, and CEP pitch detectors because of the simplified autocorrelation function which is computed using a counter rather than a multiplier and an adder.

Table XXI also includes some of the details of how the various pitch detection algorithms were implemented on the NOVA 800 computer. The numerical method of realization (i.e., fixed or floating point) is indicated in the column labeled "arithmetic type." Three of the algorithms were realized in integer arithmetic (DARD, PPROC, and AUTOC); three were realized in floating-point arithmetic (AMDF,[3] SIFT, and LPC): the CEP method used both integer arithmetic [for windowing and fast Fourier transforms (FFT's)], and floating arithmetic (for the log magnitude operation). The next column indicates whether or not downsampling (i.e., reduction of the sampling rate of the signal to a lower rate) was used in the realization to reduce the computation. Although not used for the AMDF and AUTOC methods, it could easily be incorporated into these methods to speed up the realization. Finally, the last column shows the dependence of the computation on the sampling rate of the input. As seen in this table, all the methods are approximately linearly or quadratically dependent on the sampling rate, assuming all the parameters of the analysis (i.e., analysis section length, pitch range, etc.) remain the same.

TABLE XXI
Computational Considerations for the Seven Pitch Detectors on the Nova 800 Minicomputer

Pitch Detector	Speed/s of Speech	Arithmetic Type	Down-sampling Used	Dependence on Sampling Rate
DARD	5 s	Integer	No	Linear
PPROC	7.5 s	Integer	No	Linear
AMDF	50 s	Floating point	No[a]	Quadratic
AUTOC	120 s	Integer	No[a]	Quadratic
SIFT	250 s	Floating point	Yes	Quadratic
LPC	300 s	Floating point	Yes	Quadratic
CEP	400 s	Mixed	No	Linear

[a]These algorithms could easily incorporate downsampling.

VIII. Summary

This paper has reported on the results of a rather extensive performance evaluation of seven pitch detection algorithms. Using a variety of error measurements, the performance strengths and weaknesses of each of the pitch detectors for different speakers and different recording conditions were highlighted.

A major issue which arises when trying to understand the results of this study is how to interpret the various error scores. This is one problem for which we have no simple answer other than it all depends on the intended application of the pitch analysis. For example, classifying a low-level voiced speech interval as unvoiced may be perfectly acceptable for a vocoder, but may cause great problems for a recognition system. Similarly, the level at which various types of errors become significant also depends strongly on the application. We have presented performance scores based on a criterion related to the applications with which the authors are most familiar, i.e., speaker verification systems [2] and digit recognition systems [17].

Finally, an important consideration in interpreting the results presented here is the perceptual effect of each of the types of errors discussed in Section V. A parallel series of investigations is required to provide perceptual comparisons among the seven pitch detectors. Such an investigation is currently being made by the authors.

[2]Cycle time 800 ns, add time of 1.6 μs, multiply time of 3.6 μs. The machine also had floating-point hardware.

[3]The AMDF algorithm as provided to us was implemented in integer arithmetic. However the 16-bit integer representation of the NOVA 800 is inadequate for this implementation. Consequently, the computations were converted to floating point.

References

[1] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687-1697, Dec. 1972.

[2] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 169-176, Apr. 1975.

[3] H. Levitt, "Speech processing aids for the deaf: An overview," *IEEE Trans. Audio Electroacoust. (Special Issue on 1972 Conference on Speech Communication and Processing)*, vol. AU-21, pp. 269-273, June 1973.

[4] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. New York: Springer-Verlag, 1972.

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

[6] 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.
[7] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust. (Special Issue on Speech Communication and Processing—Part II)* vol. AU-16, pp. 262-266, June 1968.
[8] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.
[9] N. J. Miller, "Pitch detection by data reduction," *IEEE Trans. Acoust., Speech, Signal Processing (Special Issue on IEEE Symposium on Speech Recognition)*, vol. ASSP-23, pp. 72-79, Feb. 1975.
[10] 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.
[11] 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.
[12] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634-648, Feb. 1970.
[13] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A semiautomatic pitch detector (SAPD)," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 570-574, Dec. 1975.
[14] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer, 1976.
[15] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201-212, June 1976.
[16] R. E. Crochiere and L. R. Rabiner, "Optimum FIR digital filter implementations for decimation, interpolation, and narrow-band filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 444-456, Oct. 1975.
[17] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 552-557, Dec. 1975.
[18] M. J. Cheng, "A comparative performance study of several pitch detection algorithms," M. S. thesis, Mass. Inst. Technol., Cambridge, June 1975.
[19] L. R. Rabiner and M. R. Sambur, "Some preliminary experiments in the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 170-182, Apr. 1976.

Maximum Likelihood Pitch Estimation

JAMES D. WISE, STUDENT MEMBER, IEEE, JAMES R. CAPRIO, MEMBER, IEEE, AND THOMAS W. PARKS, MEMBER, IEEE

Abstract—A method for estimating the pitch period of voiced speech sounds is developed based on a maximum likelihood (ML) formulation. It is capable of resolution finer than one sampling period and is shown to perform better in the presence of noise than the cepstrum method.

I. Introduction

MANY current speech encoding techniques attempt to achieve low-rate digitized speech transmission by modeling the speech source as a linear filter, representing the vocal tract resonances, excited by either random noise or a quasiperiodic pulse train, representing the source signal for unvoiced and voiced speech, respectively. Achieving good-quality resynthesized speech with this model requires that both the filter parameters and the excitation signal be accurately estimated. The mechanical character of the early vocoders was due largely to their inability to extract the excitation signal and motivated the development of the voice-excited vocoder [1]. It is this same difficulty in accurately modeling the excitation signal which is responsible for the popularity of systems which transmit a quantized version of the estimated excitation signal even though they have a higher data rate than those which parameterize the excitation.

Manuscript received October 23, 1975; revised March 30, 1976 and April 6, 1976. This work was supported in part by the National Science Foundation under Grant ENG 70-01349 A03.

J. D. Wise and T. W. Parks are with the Department of Electrical Engineering, Rice University, Houston, TX 77001.
J. R. Caprio is with Comptek Research, Inc., Buffalo, NY.

The difficulty in estimating the excitation is compounded by departures from the idealizations used in developing the method which are encountered in a realistic situation: noisy environment, absence of the fundamental due to band-limiting, simultaneous presence of periodic and random excitation, phase distortion, or rapid changes in pitch period. In particular, the sensitivity of the pitch detector to ambient noise is a serious limitation in many potential applications of analysis-synthesis telephony systems [2].

The pitch detection scheme to be discussed is designed to be resistant to white, Gaussian noise, may be extended to colored noise, and shows promising performance in the presence of realistic environmental noise. In addition, it is capable of determining the pitch period with a resolution finer than one sample period, resulting in improved performance for high-pitched speech.

In revising this paper, the authors discovered that Noll [3] proposed a maximum likelihood (ML) pitch estimation method similar to that described in Section II of this paper. The modifications developed in Section III reduce his problem of multiple peaks with increasing amplitude caused by noise, and enable this method to provide a parameter which is useful in making the voiced-unvoiced decision. An interpretation in terms of the signal autocorrelation function is presented which is useful in suggesting possible efficient implementations and in providing insight into the frequency-domain behavior of the estimator. This method has been evaluated on several utterances with various types of noise and signal-to-