Correspondence

Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm

Sassan Ahmadi and Andreas S. Spanias

***Abstract*— An improved cepstrum-based voicing detection and pitch determination algorithm is presented. Voicing decisions are made using a multifeature voiced/unvoiced classification algorithm based on statistical analysis of cepstral peak, zero-crossing rate, and energy of short-time segments of the speech signal. Pitch frequency information is extracted by a modified cepstrum-based method and then carefully refined using pitch tracking, correction, and smoothing algorithms. Performance analysis on a large database indicates considerable improvement relative to the conventional cepstrum method. The proposed algorithm is also shown to be robust to additive noise.**

***Index Terms*—Feature classification, pitch determination, speech processing, threshold adaptation, voicing detection.**

I. Introduction

Pitch detection is an essential task in a variety of speech processing applications. Although many pitch detection algorithms (PDA's), both in the time and frequency domains, have been proposed in the literature [2], accurate and robust voicing detection and pitch frequency determination remain an open problem. The difficulty involved in pitch detection stems from the nonstationarity and quasiperiodicity of the speech signal as well as the interaction between the glottal excitation and the vocal tract. Threshold-based classifiers are typically used for voicing decisions (e.g., conventional cepstrum and autocorrelation methods [7]). The voicing decision is often made by examining if the value of a certain feature exceeds a predetermined threshold. Inappropriate selection of the threshold, regardless of input signal characteristics, results in performance degradation.

The PDA presented in this work overcomes some of the aforementioned problems by exploiting an improved method for voiced/unvoiced (V/UV) classification based on statistical analysis of cepstral peak, zero-crossing rate, and energy of short-time speech segments. Although the proposed algorithm was originally inspired by the work reported in [4], there are some significant differences relative to the conventional cepstrum method. Unlike the conventional cepstrum method, the proposed algorithm uses a multifeature classification scheme as well as signal-dependent initial-thresholds, and a different cepstral weighting function, which improves the detectability of low-frequency pitch peaks.

The proposed multifeature V/UV classification algorithm, as depicted in Figs. 1 and 2, consists of two passes. In the first pass, certain features of the input speech are extracted and statistical analysis is performed to obtain the initial-thresholds required for the second stage. Preliminary voicing decisions and pitch frequency estimates are obtained in the second pass. Pitch frequency tracking and correlation between adjacent frames are then exploited to achieve an accurate and consistent estimation for the pitch frequency and voicing. A median filter is used to smooth the pitch contour and correct isolated errors in the data.

Performance analysis on a large speech database reveals relatively accurate and reliable pitch detection. Furthermore, the performance is maintained at low segmental signal to noise ratios (SSNR). It is also shown that the algorithm yields considerable performance improvement when compared to the conventional cepstrum method [4].

The rest of the correspondence is organized as follows. In Section II, a detailed description of the V/UV classification algorithm is given. In Section III, the pitch frequency determination algorithm is discussed. In Section IV, some meaningful objective error measures are defined and the results of the performance analysis are presented. Concluding remarks are given in Section V.

II. V/UV Classification Algorithm

The classification of the short-time speech segments into voiced, unvoiced, and transient states is critical in many speech analysis-synthesis systems. The essence of classification is to determine whether the speech production involves vibration of the vocal cords [5], [11]. The V/UV classification can be performed using a single feature, whose behavior could be significantly affected by the presence or absence of voicing activity. The accuracy of such an approach would not go beyond a certain limit, because the range of values of any single parameter generally overlaps between different categories. The confusion caused by overlapping between different regions is further intensified if speech has not been recorded in a high-fidelity environment. Although V/UV classification has been traditionally tied to the problem of pitch frequency determination, the vibration of the vocal cords does not necessarily result in periodicity in the speech signal [5]. Therefore, a failure in the detection of periodicity in some voiced regions would result in V/UV classification errors.

In this algorithm, a binary V/UV classification is performed based on three features, which can be divided into two categories: 1) features which provide a preliminary V/UV discrimination and 2) a feature which directly corresponds to the periodicity in the input speech. The analysis for extracting the aforementioned features is performed during the first pass, as illustrated in Fig. 1. The speech signal, sampled at 8 kHz, is analyzed at 10 ms intervals using a 40 ms Hamming window. An optional bandpass noise-suppression filter (i.e., a ninth-order Butterworth filter with lower cutoff frequency of 200 Hz and upper cutoff frequency of 3400 Hz) is applied to deemphasize the out-of-band noise when the input speech is contaminated with additive noise as well as providing an appropriate high-frequency spectral roll-off. After this preprocessing stage, the following features are extracted and analyzed.

1) Cepstral Peaks: The cepstrum, defined as the real part of the inverse Fourier transform of the log-power spectrum, has a strong peak corresponding to the pitch period of the voiced speech segment being analyzed [4]. A 512-point fast Fourier transform (FFT) was found sufficient for accurate computation of the cepstrum. The cepstral peaks corresponding to the voiced segments are clearly resolved and quite sharp. Hence, the peak picking scheme is to determine the cepstral peak in the interval [2.5–15 ms], corresponding to pitch frequencies between 60–400 Hz, which exceeds some

Manuscript received July 27, 1996; revised August 21, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

S. Ahmadi is with Nokia Mobile Phones, Inc., San Diego, CA 92121 USA (e-mail: sassan.ahmadi@nmp.nokia.com)

A. S. Spanias is with the Department of Electrical Engineering, Arizona State Univesity, Tempe, AZ 85287 USA (e-mail: spanias@asu.edu).

Publisher Item Identifier S 1063-6676(99)02735-2.

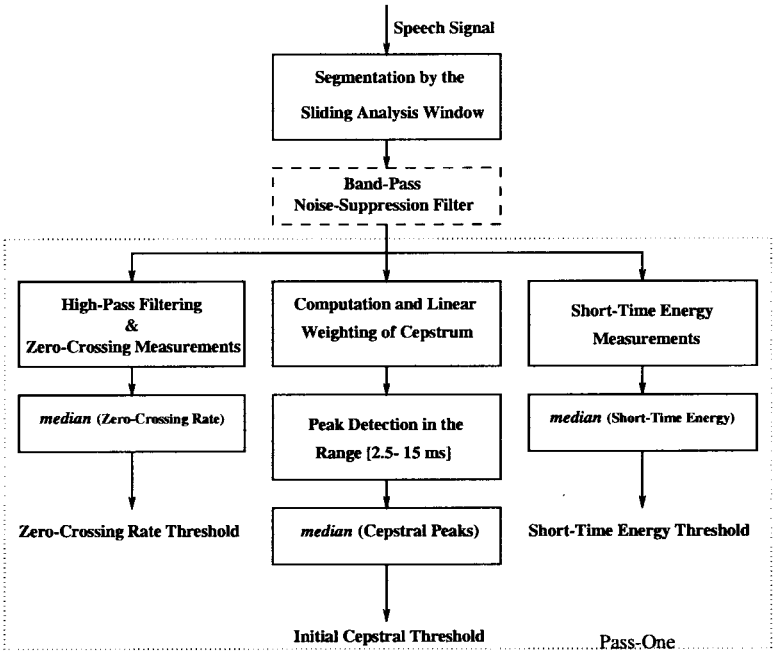

Fig. 1. Flowchart of the first pass of the proposed algorithm.

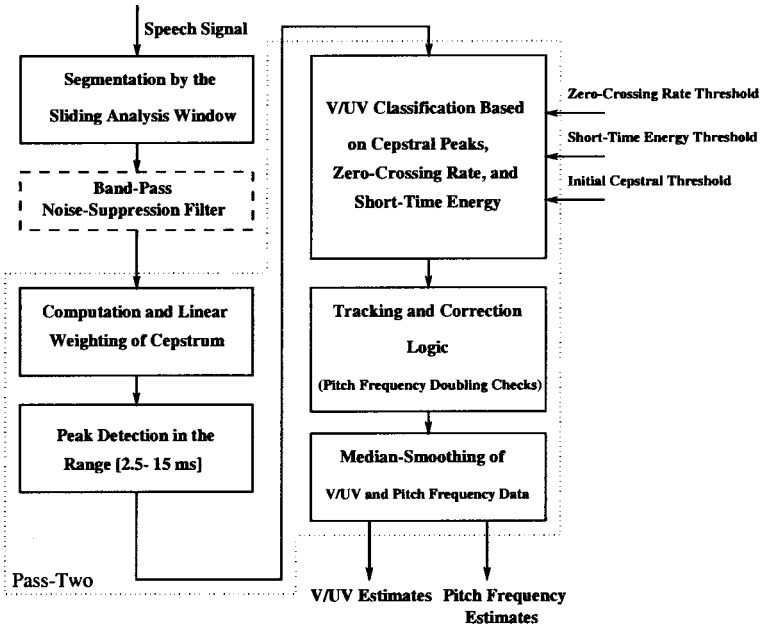

Fig. 2. Flowchart of the second pass of the proposed algorithm.

specified threshold. Since the cepstral peaks decrease in amplitude with increasing quefrency, a linear cepstral weight is applied over the 2.5 to 15 ms range. The linear cepstral weighting with range of one to eight was found empirically by using periodic pulse trains with varying periods as the input to the pitch determination program. The strength and existence of a cepstral peak for voiced speech is dependent on a variety of factors, including the length of the analysis window applied to the signal and the formant structure of the input signal. The window length and the relative positions of the window and the speech signal will have considerable effect on the height of the cepstral peaks [8]. If the window length is less than two pitch period long, a strong indication of periodicity cannot be expected. The longer the window, the greater the variation of the speech signal from the beginning to the end. Therefore, considering the tapering effect of the analysis window, the window length was set to 40 ms to capture at least two clearly defined periods in the windowed speech segment.

The extraction of the cepstral peaks is a deterministic problem. However, to decide if a cepstral peak represents a voiced segment requires a decision level (i.e., the threshold) that is not deterministic and strongly depends on the characteristics of the input speech. A plot of the histograms of the cepstral peaks corresponding to four different male and female utterances is shown in Fig. 3. In order to determine the optimum threshold, statistical distributions of the cepstral peaks corresponding to the voiced and unvoiced segments of speech must be known in advance. This *a priori* information is not generally provided. If such information were available, a *maximum*

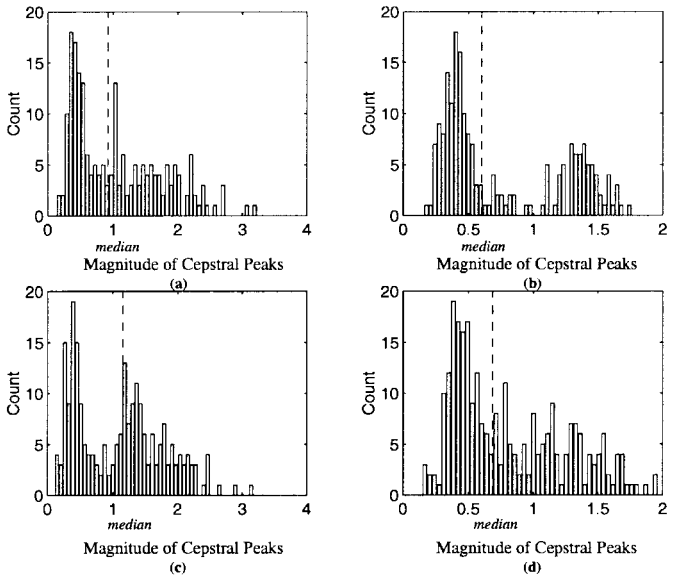

Fig. 3. Histograms of cepstral peaks. (a), (c) Distributions for two different male speakers. (b), (d) Distributions for two different female speakers.

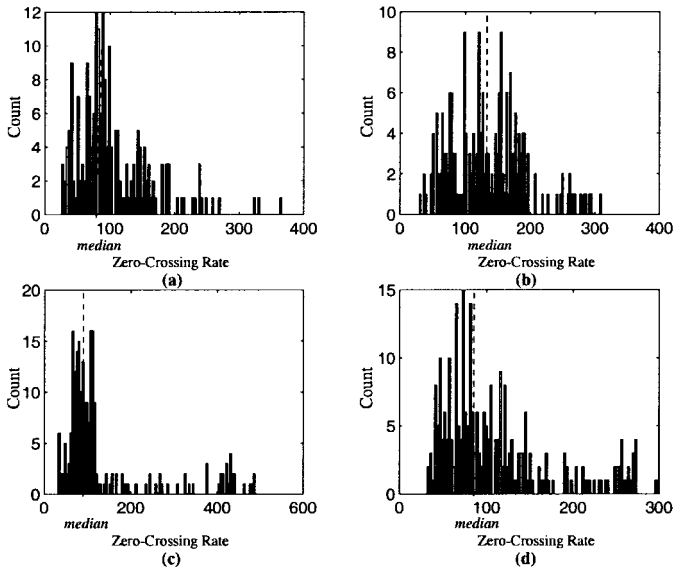

Fig. 4. Histograms of zero-crossing rate. (a), (c) Distributions for two different male speakers. (b), (d) Distributions for two different female speakers.

a posteriori probability (MAP) estimate of the initial-threshold could be obtained by finding the value of θ for which the following cost function was minimized:

$$\eta(\theta) = P_{\rm v} \int_{-\infty}^{\theta} f_{\rm v}(x)\,dx + P_{\rm uv} \int_{\theta}^{\infty} f_{\rm uv}(x)\,dx \tag{1}$$

where $P_{\rm v}$ and $P_{\rm uv}$ denote the probabilities that speech is voiced or unvoiced, respectively. The functions $f_{\rm v}(x)$ and $f_{\rm uv}(x)$ represent the statistical distributions of the cepstral peaks associated with voiced and unvoiced segments of the speech signal, respectively. Similar expressions can be used to determine the optimum thresholds corresponding to the other features.

It is a well-known fact that the cepstral peaks corresponding to the unvoiced segments have smaller magnitudes than those associated with the voiced segments. However, the regions that contain voiced and unvoiced cepstral peaks overlap and an absolute discrimination is not possible. It must be noted that, even if the actual statistical distributions were known, the initial-threshold obtained in (1) could not strictly discriminate between voiced and unvoiced cases because of the unavoidable overlapping between the regions.

A practical approach is to seek a value that minimizes some meaningful error criteria. Based on statistical analysis of the observations and the properties mentioned above, it was found that the *median* of the cepstral peaks is relatively a good criterion to be used as the initial-threshold. This choice of the threshold divides the set of observations into two subsets of equal number of entries. These regions can be defined as follows:

$$\begin{aligned} R_{\mathcal{C}}^{L} &= \{C_i |\, \min(\mathcal{C}) \le C_i < \text{median}(\mathcal{C})\}; \\ R_{\mathcal{C}}^{H} &= \{C_i | \text{median}(\mathcal{C}) \le C_i < \max(\mathcal{C})\} \end{aligned} \tag{2}$$

where $\mathcal{C} = \{C_m\}_{m=1}^{M}$ represents the set of all cepstral peaks, M is the total number of speech segments used in the experiment, and C_i denotes the ith cepstral peak. In practice, the parameter M is equal to the number of segments in the speech file being analyzed. It must be noted that the choice of median of a feature as the initial-threshold for preliminary classification of that feature does not constrain the number of voiced and unvoiced frames in an utterance. At the end of the first pass, the median of the cepstral peaks is computed and used as the initial-threshold for the second pass. Other values for the threshold such as mean and a percentage of the maximum value of the corresponding feature, as well as a constant-threshold were also investigated. These values were either signal-independent or strongly affected by extreme values measured for the corresponding feature. The choice of median will be further justified in Section IV.

2) Short-Time Zero-Crossing Rate: In the context of discrete-time signals, a zero-crossing occurs if successive samples have different algebraic signs. Although the basic algorithm needs only a comparison of signs of two successive samples, the speech signal has to be preprocessed to ensure a correct measurement. Noise, DC offset, and 60-Hz hum have deleterious effects on zero-crossing measurements. In this algorithm, the speech signal is filtered by a ninth order highpass Chebyshev filter with lower cutoff frequency of 100 Hz to avoid the aforementioned difficulties. The sampling frequency of the speech signal also determines the time resolution of the zero-crossing measurements. The zero-crossing rate corresponding to the ith segment of the filtered speech is computed as follows:

$$ZCR_i = \sum_{n=1}^{N-1} |\text{sgn}[x_i(n)] - \text{sgn}[x_i(n-1)]| \tag{3}$$

where N = 320 (i.e., corresponding to 40 ms analysis window) denotes the length of the windowed speech segment, $x_i(n)$. A reasonable criterion is that, if the zero-crossing rate exceeds a given threshold, the corresponding segment is likely to be unvoiced; otherwise, the speech segment is likely to be voiced. This, however, could be an imprecise statement, because the distributions of the zero-crossing rates of voiced and unvoiced segments inevitably overlap. Fig. 4 shows the distribution of zero-crossing rates of various male and female utterances. It will be shown that the median of the zero-crossing rates is usually the most appropriate value to be used as the threshold. The validity of this choice is further justified by considering the above properties and the fact that this value is not affected by extreme values in the data. This signal-dependent threshold divides the region between the minimum and the maximum value of the zero-crossing rate into two regions with equal number of elements, where the decision regions can be defined as follows:

$$\begin{aligned} R_{\mathcal{Z}}^{L} &= \{ZCR_i |\, \min(\mathcal{Z}) \le ZCR_i < \text{median}(\mathcal{Z})\}; \\ R_{\mathcal{Z}}^{H} &= \{ZCR_i | \text{median}(\mathcal{Z}) \le ZCR_i < \max(\mathcal{Z})\} \end{aligned} \tag{4}$$

where $\mathcal{Z} = \{ZCR_m\}_{m=1}^{M}$ denotes the set of all zero-crossing rates. Therefore, the median of the zero-crossing rate is computed in the

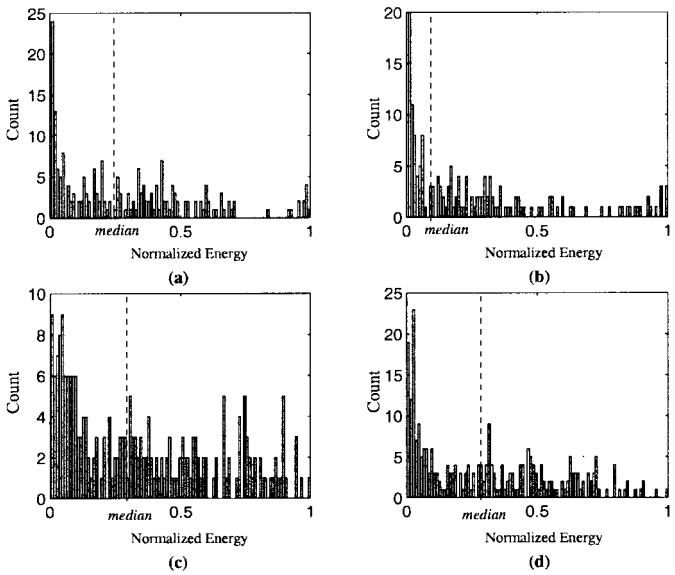

Fig. 5. Histograms of normalized short-time energy. (a), (c) Distributions for two different male speakers. (b), (d) Distributions for two different female speakers.

first pass and used as the threshold in the second pass. Since the preliminary decisions are further refined in the second pass, this choice of the threshold will not restrict the number of voiced and unvoiced frames in the input speech signal.

3) Short-Time Energy: The energy of the ith speech segment, defined as $E_i = \sum_{n=0}^{N-1} |x_i(n)|^2$, provides a convenient representation that reflects the variations of the amplitude of the speech signal [8]. The energy of unvoiced segments is generally much lower than that of voiced segments. The histograms of normalized short-time energies of various male and female utterances are depicted in Fig. 5. The differences in level between voiced and unvoiced regions are well pronounced. However, transient and low-level voiced segments cannot be easily discriminated; therefore, regions that contain the energies of voiced and unvoiced segments usually overlap. The results of our studies show that the median of the short-time energies usually provides a good criterion to roughly distinguish between voiced and unvoiced regions, where the regions are defined as follows:

$$R_{\mathcal{E}}^{L} = \{E_i | \min(\mathcal{E}) \le E_i < \text{median}(\mathcal{E})\};$$
$$R_{\mathcal{E}}^{H} = \{E_i | \text{median}(\mathcal{E}) \le E_i < \max(\mathcal{E})\} \quad (5)$$

where $\mathcal{E} = \{E_m\}_{m=1}^{M}$ is the set of all short-time energies.

Based on the above discussion, the ith segment is roughly declared unvoiced if the following logical expression is satisfied:

$$[(C_i \in R_{\mathcal{C}}^{L}) \wedge (ZCR_i \in R_{\mathcal{Z}}^{H}) \wedge (E_i \in R_{\mathcal{E}}^{L})] \Rightarrow (i \in \mathcal{UV}) \quad (6)$$

where "$\wedge$" denotes the logical *and* operation, and $\mathcal{UV}$ is the set of unvoiced indices.

Although the presence of features in the complementary regions could be a strong indication that the corresponding segment is voiced, due to overlapping between the decision regions, this may not be true in general. The cepstral peaks at the end of a voiced interval usually decrease in amplitude and would fall below the initial-threshold. There is also the possibility that an isolated cepstral peak exceeds the threshold [4]. In fact, some isolated flaps of the vocal cords may result in such isolated cepstral peaks. Low-level voiced segments and rapid fluctuations of the amplitude of the voiced segments contaminated with additive noise may also lead to erroneous decisions. Some of the above problems may not be detected, which would result in single or multiple errors in final decisions. A median smoothing of order five is applied to remove single and double errors (i.e., two consecutive errors) and to smooth the output pitch frequency contours. Isolated cepstral peaks are not considered as voiced, and this is done by ignoring any cepstral peak exceeding the threshold if the immediately preceding and succeeding cepstra indicate unvoiced speech. Therefore, the immediately following cepstrum must be searched for a peak prior to making a decision about the present segment. Cepstral information of the adjacent segments are also required to detect pitch frequency doubling. It was mentioned that the cepstral peaks at the end of a voiced interval may fall below the initial-threshold. The solution is to reduce the threshold to one-half of its initial value over a quefrency range of ± 1 ms of the immediately preceding cepstral peak when tracking the cepstral peaks in a sequence of voiced speech segments [2], [4]. The threshold is reset to its initial value at the end of a series of voiced segments. Finally, the ith segment is declared voiced if either of the following conditions is satisfied.

1) $[(C_{i+1} \ge \lambda_{i+1}) \wedge (C_i \ge \lambda_i)] \Rightarrow i \in \mathcal{V}$ *(start or continue pitch tracking)*;
2) $[(C_{i+1} \ge \lambda_{i+1}) \wedge (C_{i-1} \ge \lambda_{i-1})] \Rightarrow i \in \mathcal{V}$ *(isolated absence of pitch peak)*;
3) $[(C_{i+1} \ge \lambda_{i+1}) \wedge (ZCR_i \in R_{\mathcal{Z}}^{L}) \wedge (E_i \in R_{\mathcal{E}}^{H})] \Rightarrow i \in \mathcal{V}$ *(beginning of a voiced interval)*;
4) $[(C_i \ge \lambda_i) \wedge (C_{i-1} \ge \lambda_{i-1}) \wedge (C_{i+1} < \lambda_{i+1})] \Rightarrow i \in \mathcal{V}$ *(stop pitch tracking)*;
5) $[(C_i \ge \lambda_i) \wedge (ZCR_i \in R_{\mathcal{Z}}^{L}) \wedge (E_i \in R_{\mathcal{E}}^{H})] \Rightarrow i \in \mathcal{V}$ *(a potential voiced segment)*;

where $\lambda_i \le \text{median}(\mathcal{C})$ denotes the value of the cepstral threshold at the ith segment, and $\mathcal{V}$ is the set of voiced indices.

III. Pitch Frequency Determination

If the ith speech segment is declared voiced, the pitch period is the location of the cepstral peak provided that the value of this peak exceeds the instantaneous threshold; otherwise, an estimate of the pitch frequency based on the values of the pitch frequencies of the adjacent segments is given. Erroneous pitch frequency doubling is an important issue that must be detected and eliminated. There are two types of pitch frequency doubling, which usually occur at the end of a voiced interval. The algorithm given in [4] capitalizes on this observation by looking for a cepstral peak exceeding the instantaneous threshold in an interval of ± 0.5 ms of one-half the quefrency of the double-pitch peak.

The voicing and pitch frequency data are each smoothed by a median filter of order five. Median smoothing is capable of preserving sharp discontinuities of reasonable duration in the data and still able to filter out noise (e.g., single and double errors) superimposed on the data [6]. The size of the median smoother is strictly dependent on the minimum duration of discontinuity that one wishes to preserve. It was found that a median smoother of order five would eliminate sharp discontinuities of short duration, but would preserve longer duration discontinuities. The results of informal listening tests carried out by other researchers indicate that the smoothed pitch contours are not detrimental in any way to the quality of the synthetic speech [6], [9].

IV. Experimental Results

The performance of the proposed algorithm was evaluated on speech data taken from TIMIT database. The speech material used in our experiments contained 186 speech files, corresponding to approximately 50 000 speech frames at 10 ms frame update rate, with lengths ranging from 2 to 15 s and covered a variety of speakers and a full range of pitch frequencies. An equal number of male and female

speakers from various dialect regions were utilized. The following objective error measures are used to compare the pitch frequency and voicing estimates obtained from the proposed algorithm with reference pitch frequency contours that have been constructed for the database [10], [12]. Voiced-to-unvoiced (V-UV) and unvoiced-to-voiced (UV-V) error rates denote the accuracy in correctly classifying voiced and unvoiced intervals, respectively. A UV-V error occurs when an unvoiced frame is classified erroneously as voiced. On the other hand, a V-UV error occurs if a voiced frame is detected as unvoiced by the algorithm. These errors are computed by averaging the per-frame UV-V and V-UV errors over all frames in the database. The weighted gross pitch error (GPE) [10], [12] represents a correctly classified voiced frame where the reference and the estimated pitch frequency tracks differ in fundamental frequency. This is defined as follows:

$$\text{GPE} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{E_k}{E_{\max}}\right)^{1/2}\left|\frac{f_k - \hat{f}_k}{\hat{f}_k}\right| \tag{7}$$

where K denotes the number of elements in the set of all correctly classified voiced indices in the database, $E_{\max}$ represents the maximum short-time energy, and f_k and $\hat{f}_k$ are the reference and estimated pitch frequencies for the kth frame, respectively.

It is obvious that a standard and perfectly labeled database does not exist. A labeled reference database was generated using 186 speech files taken from the TIMIT database. The preliminary reference pitch and voicing estimates were obtained using a dynamic pitch tracking algorithm. The preliminary estimates were further refined using an algorithm based on maximizing the reconstruction energy and spectral matching during harmonic analysis [3]. Then for about 2027 frames the original waveform, the synthesized waveform, the spectrograms, and the pitch frequency contour were displayed on a graphic terminal. By visual inspection and listening to the original and synthesized speech, a decision was made interactively and compared to the initial estimates of the reference pitch frequency and voicing. Correction factors were calculated and applied to the entire set of reference pitch frequency and voicing. The nature of the refinement was as follows: The frequency interval [60–400 Hz], corresponding to the range of valid pitch frequency values, was divided into small frequency bins, then average pitch errors, if any, were computed in each frequency bin. The average reference pitch errors were normalized by the value of the central frequency of the corresponding bin and then smoothed over consecutive frequency bins. The correction factors obtained in this manner were used to correct other pitch frequency estimates in the entire reference database. Further experiments such as partial comparison of the results with those obtained from other algorithms and the use of the reference pitch frequency and voicing estimates in a variety of speech coders have verified the accuracy and reliability of the reference data. After the reference database was created and refined, the performance of the proposed algorithm was evaluated.

As already mentioned, the median of the features, on the average, provides more appropriate values for the thresholds to roughly distinguish between voiced and unvoiced regions in preliminary classification. Nevertheless, this choice does not restrict the final classification of voiced and unvoiced speech segments in an utterance. In fact, the output results for many known pitch tracks were carefully examined, and the final results did not show any restriction on the number of voiced and unvoiced frames. The proposed algorithm was applied to several cases where the percentage of voiced and unvoiced frames were different from 50%, and good results were obtained. It must be noted that the initial-thresholds are set per file and they are dependent on the characteristics of the input speech file. Moreover, the initial value obtained for the cepstral threshold is adapted in consecutive voiced segments. To further justify the choice of the initial-threshold, the performance of the algorithm was evaluated based on different values for the initial-threshold and it is tabulated in Table I. Clean speech was used in all experiments. As an example, the percentage-threshold was taken 65% of the maximum value of the corresponding feature. It should be clear that the performance of the algorithm significantly changes with different percentage values. The UV-V or V-UV errors are also affected by the choice of the percentage value. On the other hand, the constant-threshold, whose value is chosen empirically, is also independent of the input speech characteristics which does not generally result in the best performance.

TABLE I
PERFORMANCE OF THE PROPOSED ALGORITHM WITH DIFFERENT VALUES FOR THE INITIAL-THRESHOLDS

Initial-Threshold	GPE (%)	V-UV (%)	UV-V (%)
Median	0.79	1.06	0.62
Mean	0.89	1.47	0.65
Constant	0.96	2.65	0.49
65% of the Maximum Value	1.09	7.22	0.15

TABLE II
PERFORMANCE OF THE PROPOSED ALGORITHM COMPARED TO THE CONVENTIONAL CEPSTRUM METHOD

Method	GPE (%)	V-UV (%)	UV-V (%)
Proposed Method	0.79	1.06	0.62
Conventional Cepstrum	1.25	7.63	1.91

TABLE III
PERFORMANCE OF THE PROPOSED ALGORITHM AT DIFFERENT SEGMENTAL SNR'S FOR MALE SPEAKERS, WHERE THE ADDITIVE NOISE IS A ZERO-MEAN WHITE GAUSSIAN NOISE. GPE, V-UV, AND UV-V DENOTE GROSS PITCH ERROR, VOICED-TO-UNVOICED ERROR RATE, AND UNVOICED-TO-VOICED ERROR RATE, RESPECTIVELY

SSNR (dB)	GPE (%)	V-UV (%)	UV-V (%)
∞	0.53	1.08	0.65
10	0.60	1.23	0.84
5	1.28	1.62	1.38
0	2.10	2.19	2.14

TABLE IV
PERFORMANCE OF THE PROPOSED ALGORITHM AT DIFFERENT SEGMENTAL SNR'S FOR FEMALE SPEAKERS, WHERE THE ADDITIVE NOISE IS A ZERO-MEAN WHITE GAUSSIAN NOISE. GPE, V-UV, AND UV-V DENOTE GROSS PITCH ERROR, VOICED-TO-UNVOICED ERROR RATE, AND UNVOICED-TO-VOICED ERROR RATE, RESPECTIVELY

SSNR (dB)	GPE (%)	V-UV (%)	UV-V (%)
∞	1.06	1.05	0.58
10	1.63	1.13	0.73
5	3.95	1.74	1.42
0	8.79	3.82	4.18

The performance of the proposed algorithm was also compared against the conventional cepstrum method [4] and the results are shown in Table II. The same reference database was used to evaluate the performance of both algorithms. The use of extra features as well as the choice of the signal-dependent initial-threshold and other modifications have caused the proposed algorithm to outperform the conventional cepstrum method.

Finally, the performance of the proposed PDA was evaluated under noisy conditions. The results of the analysis for male and female speakers at different SSNR's are shown in Tables III and IV. Pitch frequency contours of a typical male utterance at different SSNR's

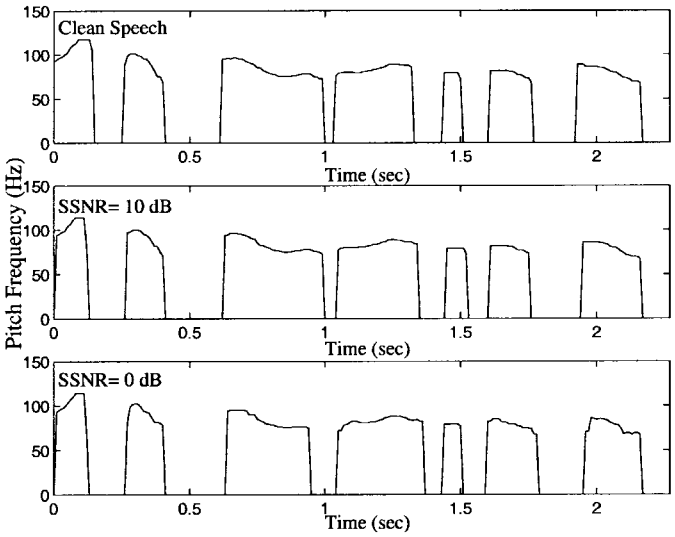

Fig. 6. Performance of the proposed algorithm under noisy conditions for a typical male utterance.

are demonstrated in Fig. 6. A white Gaussian noise was added to the clean speech, and the performance was evaluated at SSNR's of 10 and 0 dB. It is evident that the algorithm performs satisfactorily even in such noisy environment. Still, no multiple and half-pitch frequency values could be found. The intuitive reasoning for maintaining the performance under noisy condition can be summarized as follows:

1) noise samples are uncorrelated from one segment to the next segment;
2) cepstral weighting at high quefrencies, which improves the detectability of low-frequency pitch peaks;
3) the use of a multifeature classification algorithm and statistical analysis of data;
4) the use of tracking and correction algorithm;
5) the use of median smoothing to remove single and double errors in voicing and pitch frequency data.

As can be seen from Tables III and IV, the algorithm performs satisfactorily at SSNR's down to 5 dB. The proposed PDA has been utilized in various sinusoidal speech coders at rates from 9.6 to 2.4 Kb/s, where reconstructed speech of very good quality was obtained [1].

V. Conclusions

An improved multifeature voicing detection and pitch frequency determination algorithm was presented. Reliable estimations for the voicing parameters are obtained by extracting certain features of the input speech, statistical analysis of the data, and postprocessing based on signal-adaptive thresholds obtained in the first stage of the algorithm. The performance of the proposed algorithm was evaluated on a large speech database and compared to the conventional cepstrum method. It was also shown that the performance is maintained under noisy conditions.

Acknowledgment

Dynamic-pitch-tracker, pitch extractor, and pitch period marking program were provided by C. Tuerk, Engineering Department, Cambridge University, Cambridge, U.K.

References

[1] S. Ahmadi, "Low bit rate speech coding based on the sinusoidal model," Ph.D dissertation, Arizona State Univ., Tempe, AZ, June 1997.
[2] W. Hess, *Pitch Determination of Speech Signals.* Berlin, Germany: Springer-Verlag, 1983.
[3] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE ICASSP'90*, pp. 249–252.
[4] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.
[5] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and network classifier," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 250–255, Apr. 1993.
[6] 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.
[7] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comprehensive performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–417, Oct. 1976.
[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
[9] A. E. Rosenberg, "Effect of pitch averaging on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1592–1595, Aug. 1968.
[10] B. G. Secrest and G. R. Doddington, "Postprocessing techniques for voice pitch trackers," in *Proc. IEEE ICASSP'82*, pp. 172–175.
[11] L. J. Siegel and A. C. Bessey, "Voiced/unvoiced/mixed excitation classification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 451–460, June 1982.
[12] V. R. Viswanathan and W. H. Russell, "New objective measures for the evaluation of pitch extractors," in *Proc. IEEE ICASSP'85*, pp. 11.10.1–11.10.4.